Entity-Aspect Linking: Providing Fine-Grained Semantics of Entities in Context

Federico Nanni, Simone Paolo Ponzetto
Data and Web Science Group
University of Mannheim
Germany
federico,simone@informatik.uni-mannheim.de

Laura Dietz
Department of Computer Science
University of New Hampshire
U.S.A.
dietz@cs.unh.edu

ABSTRACT

The availability of entity linking technologies provides a novel way to organize, categorize, and analyze large textual collections in digital libraries. However, in many situations a link to an entity offers only relatively coarse-grained semantic information. This is problematic especially when the entity is related to several different events, topics, roles, and – more generally – when it has different aspects. In this work, we introduce and address the task of entity-aspect linking: given a mention of an entity in a contextual passage, we refine the entity link with respect to the aspect of the entity it refers to. We show that a combination of different features and aspect representations in a learning-to-rank setting correctly predicts the entity-aspect in 70% of the cases. Additionally, we demonstrate significant and consistent improvements using entity-aspect linking on three entity prediction and categorization tasks relevant for the digital library community.

KEYWORDS

entities, entity-aspects, wikification, information retrieval, knowledge bases

ACM Reference format:
Federico Nanni, Simone Paolo Ponzetto and Laura Dietz. 2018. Entity-Aspect Linking:
Providing Fine-Grained Semantics of Entities in Context. In *Proceedings of The 18th ACM/IEEE Joint Conference on Digital Libraries, Fort Worth, TX, USA, June 3–7, 2018 (JCDL '18),* 10 pages.
DOI: 10.1145/3197026.3197047

1 INTRODUCTION

The ability of enriching a collection of documents with entity-link (EL) annotations is a major advancement achieved in recent years by the natural language processing (NLP) community and a great improvement to the accessibility of large-scale digital library (DL) collections, from historical documents,[1] over newspaper corpora [43, 44] to web archives [11, 18].

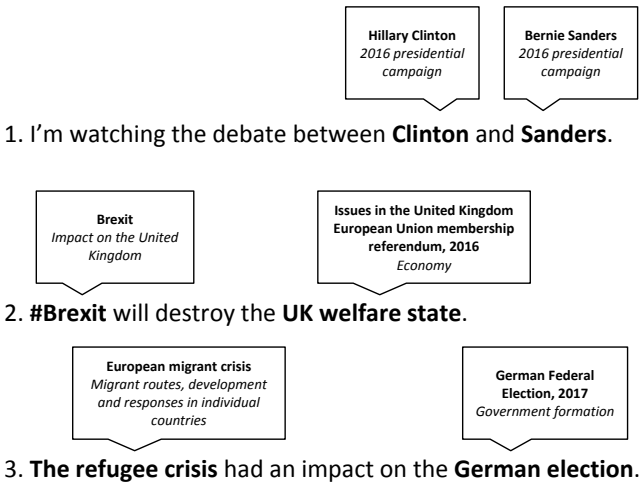

Figure 1: Example of entity-aspect linking: Boldface indicates entity links, italics presents the related aspects.

Knowing that the mentions "Clinton" and "Sanders" presented in the first example in Figure 1 refer to the DBpedia [2] entities Hillary_Clinton and Bernie_Sanders, and not to Bill_Clinton or Sarah_Huckabee_Sanders, can support several downstream applications, such as entity-centric information retrieval [9], question answering [6] as well as corpus exploration and collection-building [19, 27], which are staple tasks of the fields of Digital Humanities and Computational Social Science [28, 39, 44].

Current limitations. Although entity-linking systems such as TagMe [12] or DBpedia spotlight [22] have recently brought several benefits to the research community, the coarseness of entity-link annotations limits their adoption in many digital library tasks. For instance, if we consider the first example in Figure 1, we can notice that detecting whether a sentence mentions the entity Hillary_Clinton might not be enough for a digital librarian who wants to organize documents in different event-collections (e.g., the different U.S. Presidential campaigns – as the Internet Archive aims to do[2]). As a matter of fact, a mention of Hillary_Clinton might refer to her role as First Lady of the United Stated during the Bill Clinton presidency, as the Senator of the state of New York (2001-2009), as the opponent of Barack Obama in the 2008 Democratic Primary race, as the Secretary of State during the first Obama administration or as the Democratic presidential candidate for the

[1] See the Europeana Entity API: https://pro.europeana.eu/resources/apis/entity

JCDL '18, Fort Worth, TX, USA

978-1-4503-5178-2/18/06…$15.00
DOI: 10.1145/3197026.3197047

[2] https://archive-it.org/collections/8118

2016 Presidential Elections. While state-of-the-art entity linking technologies [12, 22] can identify that the mention "Clinton" refers to Hillary_Clinton, they do not provide further information on the aspect of the entity that is most related to the context.

The same issue emerges in social-science and humanities applications. Imagine a researcher examining the different angles from which the entity Brexit has been mentioned on social media messages (as the tweet presented in Figure 1) [21] or a scholar studying in which discourses the entity European_Migrant_Crisis emerges in relation to the vote for the 2017 German federal elections (Figure 1, example 3) [29]: in both cases the simple linking of a mention to an entry in a knowledge base will not capture the specificity of the context. For all these reasons and in order to support digital librarians in offering collections that are even more semantically enriched, we present the task of entity-aspect linking (EAL).

The task. Given an entity-mention in a specific context (e.g., in a tweet, a sentence or a paragraph), our goal is to link it to one from a set of predefined aspects that captures the addressed topic. In our setting, we assume that each of these *entity-aspects* is accompanied by a textual description and a heading. While our approach is general and applies to any source of predefined aspects, from biomedical catalogs[3] to historical knowledge resources [39, 43], in this work we exclusively focus on predefined aspects extracted from Wikipedia. Each top-level section of an entity's Wikipedia page defines a single aspect (i.e., a specific entity sub-topic), which comes with heading, textual content and entity links. The mention "Clinton" presented above would therefore be linked to the aspect 2016_Presidential_Campaign, from Hillary Clinton's Wikipedia page. The idea of using sections and the definition of "aspects" comes from previous work on automatic Wikipedia enrichment and entity recommendation [1, 13, 36, 37].

Advantages. The outcome of this task is a refined entity link, which provides richer information about the entity and enables the user to choose the granularity appropriate for the task at hand. In Figure 1, we present a series of examples where, together with the linked entities, we offer the heading of the most relevant section, given the context of the mention.

Aspects vs. types vs. properties. On the one hand, this definition of aspects has similarities to entity types, which however refer to groups of semantically equivalent entities and are therefore coarser than the definition of an entity [8, 30]. For instance, Hillary_Clinton in the first example in Figure 1 will be assigned to the types "debater", "politician", or "person", depending on the vocabulary of types. These are some of the types that are currently associated to Hillary_Clinton on DBpedia: "First Lady", "Lawyer", "Senator", "Human Being", and "Grammy Award Winner". In contrast, our definition of an aspect refers to a list of events, topics, or roles that are specific to the given entity. It is therefore more fine-grained than the definition of an entity: In the example above, the relevant aspect of Hillary_Clinton is the one referring to her participation in the 2016 presidential race.

On the other hand, our definition of aspects has similarities with properties, which describe specific attributes of an entity, such as height or date-of-birth. However, as properties are generally not accompanied with textual descriptions, we do not consider them in our work. Nevertheless, we are aware that the boundaries between types, properties, and aspects are often fluid.

Explicit vs. implicit aspect representations. In this paper, we present and address the entity-aspect linking task in particular to meet the demands of the fields of Digital Humanities and Computational Social Science [15, 34]. These communities often employ entity-linking as a semantically explicit alternative to Latent Dirichlet Allocations (LDA) [3, 4] for corpus exploration and topic-based document selection [19, 31], as latent topics detected by LDA are often difficult to interpret and evaluate [7, 24, 45]. To satisfy the need of fine-grained interpretable topics, we have developed a system that adopts and provides explicit representations of pre-defined entity-aspects harvested from Wikipedia. As a future step for supporting the exploration of digital library collections, we envision the possibility to associate each entity in each document with its aspect, which will offer rich information on the topics addressed.

Our contribution. We present a method that employs learning-to-rank on a tailored set on lexical and semantic features in order to perform the newly introduced entity-aspects linking task, with both minimal and noisy contextual information (i.e., from a sentence to a section). The system outperforms several established baselines on a set of new datasets.[4] In addition, we show the usefulness of EAL for three different tasks that are relevant for the digital library community: Query- and event-based entity-ranking and collection organization.

Outline. The rest of the paper is organized as follows. First of all, we offer an overview of related work on the topic. Next, we present the proposed method in detail. Then we describe how the entity-aspect dataset has been created. A quantitative evaluation of the different approaches follows, together with an error analysis of the different systems. Finally, we present three different applications of entity-aspects in digital library tasks.

2 RELATED WORK

In this section we offer an overview of the task of entity linking, which serves as a starting point for the entity-aspect linking task. Next, we describe previous works that have focused on identifying entity-aspects.

Entity linking. Also known as entity resolution, the task of linking textual mentions to an entity in a knowledge base is called entity linking [35]. As an information extraction task, it involves the ability to recognize named entities in text (such as people, locations, organizations, as well as products and events), to resolve coreference between a set of mentions that could refer to the same entity (e.g. "Hillary Clinton" and "Mrs. Clinton") and to disambiguate the entity by linking it to a specific entry in a knowledge base such as DBpedia [2], Yago [41] or Freebase [5]. The disambiguation process is the most challenging step of any entity linking pipeline, as mentions of an entity in text can be ambiguous (as in the "Clinton" and "Sanders" examples presented in the introduction). For this reason, entity linking systems such as TagMe! [12], DBpedia Spotlight [22] or Babelfy [23] examine the mention's context to precisely disambiguate it. For instance, in the expression "the debate between Clinton and Sanders", "Clinton" is more likely to refer to

[3] For instance MeSH: https://www.nlm.nih.gov/mesh/

[4] All the resources used in this paper are available at: https://federiconanni.com/entity-aspect-linking/

the DBpedia entity Hillary_Clinton than to Bill_Clinton. However, in the expression "Clinton vs. Bush debate", the mention "Clinton" is more likely to refer to Bill_Clinton. In recent years, EL has been tackled with approaches based on artificial neural networks, which have been employed both for producing embedding representations of entities from text and knowledge graphs [17, 38, 46], and for entity-disambiguation and linking [47].

In this paper, we address what we envision as the further fine-grained step in entity-linking, namely to identify which specific section (i.e., aspect) in the Wikipedia page of the entity is relevant, given its mention in context. In building our pipeline, we test both established approaches from the entity-linking task and we also examine the usefulness of both word and entity embeddings.

Sections as entity-aspect. The formalization of entity-aspect that we adopt in our work is derived by Fetahu et al. [13]. In their paper, the authors enrich Wikipedia sections with news-article references. They do so in two steps: *a)* article-entity placement and *b)* article-section placement. While the authors do not explicitly use the word "aspect" in their paper, they consider – as we do – each section as a different sub-topic. In a similar way, Banerjee and Mitra [1] enrich Wikipedia stubs by assigning content retrieved from the web to the specific section. Following these works, Reinanda et al. [37] recently adopted the definition of "aspect" of an entity referring to its section; in their paper the authors use aspect-features to identify relevant documents for long tail entities. Unfortunately, the datasets used in these works are not available to the research community (we will expand on this in the Dataset section).

Our work differs from these studies in a few ways: *1)* we develop a system able to deal with both short and long snippets of text, and not only with long documents, as in previous work [1]. This is because we believe that the research community will benefit from such approach in order to analyze different types of digital collections, from historical archives to datasets of social media messages; *2)* we build a system that is applicable to any entity mention in a context, and not only to the most salient entities in text (as opposed to Fetahu et al. [13]); *3)* we do not adopt any type of feature that is specific to a particular Wikipedia category (as in Banerjee and Mitra [1]), in order to present a pipeline that is completely domain-independent and can be used to link any type of entity; *4)* finally, we present a fine-grained entity-linking tool to be used in digital library applications for supporting research in the humanities and social sciences. While our approach could be adopted to enrich Wikipedia sections with new snippets (i.e., the focus of previous papers on entity aspects), the long-term goal of our work is a system that could directly be used by digital librarians for creating topic-specific collections from large archives, for instance by retrieving all mentions of Hillary_Clinton related to the 2016 Election Campaign from the Internet Archive.

Other ways of obtaining "aspects". While in our paper we use sections to describe different aspects of an entity, there are other ways to detect them: for instance, by extracting information from large-scale query logs [32, 36, 48], or by conducting entity profiling in order to generate salient content about an entity from a given corpus [20, 42]. Even though these two types of approaches could be useful in many contexts, they strongly depend on the corpus under study and cannot be extended to, for example, long-tail entities that do not frequently appear in the collection, as we would like to do in our work.

3 APPROACH

In order to assign a mention m of the entity e in context cx to the most relevant aspect a of its Wikipedia page w (see Figure 2), we test different aspect representations and ranking strategies and combine them together in a learning-to-rank (L2R) system.

As opposed to previous work [1, 13], this method can be applied to any entity mention, regardless of its entity-type or prominence in a context. Moreover, in the evaluation section we provide evidence of the usefulness of this approach with text of different length.

While the methodology here described relies on Wikipedia, it can be adapted to any knowledge resource which provides textual descriptions of entity "sub-topics".

3.1 Aspect Representations

We consider three different ways of representing the entity-aspect in order to detect its similarity to the mention in context:

Header. We compare the similarity of the mention (m) in context (cx) and the header (h) of each section in the Wikipedia page (w) and rank aspects based on that. We do so, because often headers are very short summaries of the content of their section (as 2016_Presidential_Campaign, regarding the example in Figure 1).

Content. We measure the similarity between the mention in context (cx) and the content (co) of each section of the Wikipedia page of the entity and rank aspects based on that, as already done in previous work [1, 13].

Entity. We compare the entities (el) mentioned in cx and co. This follows the intuition that referring to similar entities should be an additional signal on the relation between the two texts. For detecting entities we used TagMe [12] with standard parameters.

3.2 Features

We employ two types of feature-vectors for ranking aspects.

3.2.1 Word Vector Models. These features consider symbolic representations of each word as a single token. We employ them to rank aspects using header, content and entity representations:

tf-idf (cs). We compute the cosine similarity (cs) between the tf-idf (logarithmic, L2-normalized variant) vector of contextual mention and aspect. We tested both lemmatization and stemming as pre-processing steps during the experimental phase; as lemmatization has always offered better performance we report these results in the paper. We excluded numbers and stopwords[5].

BM25. We rank aspect representations given the contextual mention as a query using Okapi BM25 with k_1=2 and b=0.75. As in the previous approach, we adopt lemmatization and we remove numbers and stopwords.

3.2.2 Distributional Semantic Models. The following features consider the representations of each word/entity as a vector in an embedding space. We employ word embeddings for ranking aspects with header and content representations, while we use entity-embeddings with the entity representation of each aspect:

[5]Based on the NLTK English stopword list: https://www.nltk.org/

Table 1: Type of features generated from each aspect-representation.

Representation	Features
Header	tf-idf (cs), BM25, w-emb (cs)
Content	tf-idf (cs), BM25, w-emb (cs)
Entity	tf-idf (cs), BM25, ent-emb (cs)

w-emb (cs). We compute the cosine similarity between the mention in context and the aspect using the pre-trained word embeddings GloVe [33] of 300 dimensions. As in previous work [16, 25], every word w in cx and co is represented as a K-dimensional vector $\vec{w}$. A vector representation $\vec{v}$ for the whole text d is then obtained by a weighted element-wise average of word vectors $\vec{w}$ in the text. To give more attention to infrequent word, we additionally use the tf-idf of each word w to weight:

$$\vec{v} = \frac{1}{|d|} \sum_{w \in d} \text{tf-idf}(w) \cdot \vec{w}$$

ent-emb (cs). As in previous work [26, 27], we obtain latent vector representations $\vec{el}$ of each linked entity el appearing in cx and co using pre-computed RDF2Vec 500 dimensions entity embeddings [38]. A vector representation $\vec{v}$ for the whole text d is then computed as a weighted element-wise average of entity vectors $\vec{el}$. By casting mentions in context and aspects as bag-of-entities we adapt tf-idf to entity links (link statistics from DBpedia 2015-04 [2]):

$$\vec{v} = \frac{1}{|\{el \in d\}|} \sum_{el \in d} \text{tf-idf}(el) \cdot \vec{el}$$

3.3 Machine Learning

For measuring the similarity between every entity in context and aspect, the features presented above generates a vector of length 9, as summarized in Table 1.

Aspects are then ranked with a list-wise learning-to-rank (L2R) approach implemented in RankLib.[6] The weight parameter is learned by optimizing for the precision at 1 (P@1) using coordinate ascent with linear normalization. We then conduct a feature-ablation study by applying L2R on held-out test data using 5-fold cross-validation. In the experimental section we provide evidence on the usefulness of the selected features.

4 ENTITY-ASPECT DATASET

Given the fact that the resources employed in previous works [1, 13, 37] were too general for our task (they consider only certain type of entities and context sizes) and are not available anymore,[7] we decided to build a new collection of entity-aspect links to evaluate the presented method, based upon human-curated Wikipedia section links.

[6]https://sourceforge.net/p/lemur/wiki/RankLib/

[7]The dataset from Fetahu et al.[13] is the only resource still available online; however, this collection reflects only a sub-part of the dataset used in their paper. This issue impeded us from employing it in our work. More information here: https://federiconanni.com/aspect-linking-previous-dataset/

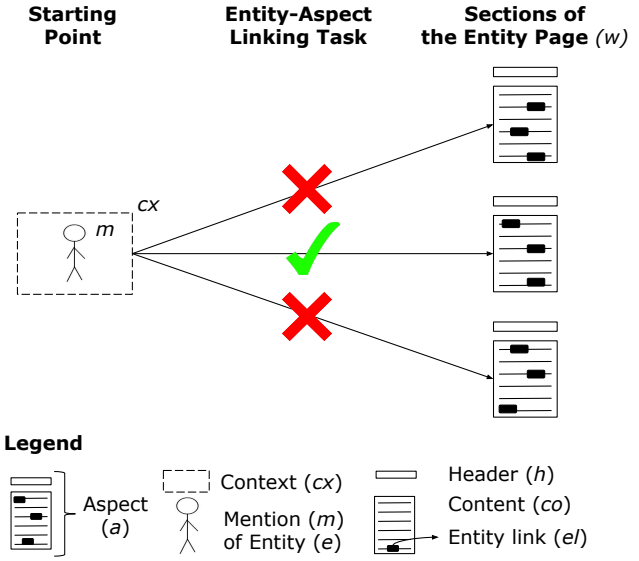

Figure 2: Graphical representation of the task.

Wikipedia section links. As described in Wikipedia's Manual of Style and Linking,[8] "if an existing article has a section specifically about the topic" Wikipedia contributors are encouraged to point hyper-links directly to it. This should be done by suffixing the article name with a # and the name of the section (e.g., "Hillary_Clinton #2016_presidential_campaign).

We derived a collection of these section-links from the recently released TREC Complex Answer Retrieval dataset,[9] where organizers have processed a large amount of articles from Wikipedia, organizing them in sections, subsections, paragraphs and disambiguating entity-links [10].[10] From this dataset, we have retrieved a set of 201 entity-section links and for each of them we have manually reassessed the existence of a relation between the mention of the entity in context and the linked section.

Aspect structure. For each of these links, we consider the top level section heading as the associated aspect. Each aspect has therefore a name, which is the heading of the section, and a content, comprising all text included in the section (both textual descriptions and sub-headings).

Gold standard. We additionally provide a set of negative section-candidates for each correct section-link, which are extracted from the same Wikipedia page. The number of candidates, depending on the Wikipedia page under study, varies between 2 and 29, with an average of 6 potentially relevant aspects for each section-link. Generating this type of natural gold standard leads to a total of 1274 entity-section pairs.

If we reconsider the first example in Figure 1, together with the section 2016_presidential_campaign, we also present the sections 2008_presidential_campaign, U.S._Secretary_of_State, Email_ controversy, and so on.

[8]https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking

[9]http://trec-car.cs.unh.edu/

[10]We used the unprocessed_train file, version 1.5.

Table 2: Statistics of the dataset: number of tokens, depending on the context.

Context	Min.	Avg.	Max.
sentence	5	27	102
paragraph	5	94	318
section	48	19478	1259176

4.1 Different Types of Context

For each entity-aspect link, we examine the performance of our system considering different types of context (statistics in Table 2).

Sentence. The first type of context we test is the sentence in which the entity is mentioned. While this context should be highly relevant for the mentions, its brevity allows to simulate entity-aspect linking in situations where more contextual information is lacking (e.g., on social media platforms).

Paragraph. The second type of context is the paragraph in which the entity is mentioned. Paragraphs are taken from the Trec-CAR paragraph corpus [10] and are generally composed of 3–4 sentences, offering richer context to the mention.

Section. We additionally consider the section in which the entity is mentioned. This will allow us to test EAL, given a large context (which could simulate, for example, the length of a newspaper article). As opposed to previous work [13], we do not filter entities based on whether they are salient for the context or not; for this reason the task is very challenging, as already pointed out [27].

5 EXPERIMENTAL EVALUATION

We present a quantitative evaluation of our system on the new entity-aspect dataset described in the previous section. These experiments aim to answer the following research questions:

- **RQ1**. To which extent is it possible to link an entity-mention in context to the most appropriate aspect?
- **RQ2**. Does the combination of different representations of aspects (header, entity) improve the performance, over the use of content features?
- **RQ3**. Do different context sizes have an impact on the performance of the proposed method?
- **RQ4**. If yes, what is the best context setting for performing entity aspect-linking?

We begin by introducing the set of baselines we tested in our work, then we offer an overview of our system setting. We evaluate both baselines and our approach in terms of Precision at 1 (P@1) and mean average precision (MAP) using trec-eval.[11] We conclude this section with a discussion of the results.

5.1 Baselines

Inspired by previous work [1, 13, 37] and by the literature on entity linking [23, 35], we test a variety of standard approaches and heuristics for performing entity-aspect linking.

[11] Version 8.1: http://trec.nist.gov/trec_eval/

5.1.1 Simple Heuristics. The following baselines test a few simple heuristics:

- **size.** We consider the length of each section (in number of tokens) and we link the entity-mention to the longest.
- **content overlap.** We measure the overlap between tokens in the context of the mention and the section. Words have been tokenized, lemmatized and we excluded numbers and stopwords.
- **entity overlap.** We measure the overlap between entities in the context of the mention and in the section.

5.1.2 Text Similarity Baselines. The second group of baselines considers each aspect representation and ranking strategy presented in the Approach section as an independent system, for example: **tf-idf (cs)** with header representation, **BM25** with content representation, etc.

5.1.3 Learning To Rank Baselines. This third group of baselines uses all ranking features from each single aspect representation:

- **L2R (Header).** All the features generated by the comparison between the mention in context and the header of the section.
- **L2R (Content).** All the features generated by the comparison between the mention in context and the content of the section. This approach is supposed to be a very hard-to-beat baseline.
- **L2R (Entity).** All the features generated by the comparison between the entities mentioned in the context and the entities mentioned in the section.

This final group of baselines will tell us whether considering only a single type of aspect representation (e.g., only the content) is sufficient for performing the task.

5.2 Our Models

For what concerns the learning-to-rank approach proposed in this paper for entity-aspect linking (EAL), we present the results adopting two different settings:

- **EAL (All).** This setting adopts all features presented in the Approach section (denoted with † in the results).
- **EAL (Selected).** This setting employs a tailored set of features (denoted with ⋄ in the results), namely: tf-idf (header, content and entity), BM25 (content) and w-emb (content). We have identified them by conducting a series of ablation studies in the different test settings, using an additional validation set. We refer to this as EAL (Selected) in the rest of the paper .

5.3 Aspect-Linking with Sentence Context

The first experiment we present is on linking the correct aspect (a) of an entity (e), when only a sentence is given as a context cx. An example could be the following sentence (entity mention in boldface):

*I'm watching the debate between **Clinton** and Sanders.*

As it can be seen by the results presented in Table 3,[12] the use of a simple heuristics that checks the content overlap between the

[12] In tables, * marks improvements statistically significant (p-value lower than 0.10; paired Student's t-test) over all other baselines.

Table 3: Precision at 1 (P@1) and mean average precision (MAP) on aspect-linking using sentence context.

Model	P@1	MAP
random baseline	0.16	0.41
Simple Heuristics		
size	0.39	0.60
content overlap	0.57	0.72
entity overlap	0.50	0.69
Header		
tf-idf (cs) †⋄	0.44	0.63
BM25 †	0.42	0.62
w-emb (cs) †	0.32	0.54
L2R (Header)	0.44	0.63
Content		
tf-idf (cs) †⋄	0.62	0.76
BM25 †⋄	0.60	0.74
w-emb (cs) †⋄	0.54	0.70
L2R (Content)	0.65	0.77
Entity		
tf-idf (cs) †⋄	0.50	0.69
BM25 †	0.54	0.71
RDF2Vec (cs) †	0.49	0.68
L2R (Entity)	0.57	0.73
Our Approach		
EAL (All) †	0.66	0.79
EAL (Selected) ⋄	***0.70**	***0.81**

Table 4: Precision at 1 (P@1) and mean average precision (MAP) on aspect-linking using paragraph context.

Model	P@1	MAP
random baseline	0.16	0.41
Simple Heuristics		
size	0.39	0.60
content overlap	0.55	0.72
entity overlap	0.53	0.71
Header		
tf-idf (cs) †⋄	0.42	0.62
BM25 †	0.41	0.62
w-emb (cs) †	0.31	0.53
L2R (Header)	0.42	0.62
Content		
tf-idf (cs) †⋄	0.62	0.76
BM25 †⋄	0.58	0.74
w-emb (cs) †⋄	0.53	0.69
L2R (Content)	0.58	0.75
Entity		
tf-idf (cs) †⋄	0.53	0.70
BM25 †	0.55	0.70
RDF2Vec (cs) †	0.49	0.66
L2R (Entity)	0.53	0.70
Our Approach		
EAL (All) †	0.63	0.77
EAL (Selected) ⋄	**0.65**	**0.78**

sentence and the section is already a good indicator for the entity-aspect linking task. On the other hand, the use of header features by themselves rarely provide good results. Adopting entity features alone does not lead to good performance either; this is due to the fact that a section often mentions a very large number of entities and only a small fraction of them is relevant to the topic. The use of content features in L2R leads to good performance, strongly improving over the simple use of content overlap.

For what concerns the apporaches introduced in this paper, our tailored selection of features and aspect representations, named EAL (Selected), achieves a statistically significant improvement over all baselines, both for P@1 and MAP. This is due to the fruitful combination of features derived from different aspect representations. Our EAL (All) approach also shows very good performance, significantly outperforming the majority of the baselines.

Take-Away. Linking an entity to its most relevant aspect given a sentence of context is often possible, as our EAL-Selected system shows by reaching a P@1 of 0.70 (**RQ1**). In particular, our experiments reveal that for facing the scarce amount of contextual information available in a sentence, it is often necessary to combine features derived from the content with header and entity representation of the aspect (**RQ2**).

5.4 Aspect-Linking with Paragraph Context

Next, we tackle the task of entity-aspect linking at paragraph-level. In the following example, the mentions (in boldface) of the entity Bernie_Sanders point to the section 2016_Presidential_Campaign on its Wikipedia page. This is evident, thanks to the words "Trump", "general election", "Democratic National Committee", etc., which appear in the surrounding text:

> *Although **Sanders** had not formally dropped out of the race, he announced on June 16, 2016, that his main goal in the coming months would be to work with Clinton to defeat Trump in the general election. On July 8, appointees from the Clinton campaign, the Sanders campaign, and the Democratic National Committee negotiated a draft of the party's platform. On July 12, **Sanders** formally endorsed Clinton at a rally in New Hampshire, appearing with her.*

Our starting hypothesis was that having more context surrounding the mention would support the different models, as it provides additional information. However, we discovered that the overall topic of the paragraph is not always relevant for disambiguating the referred aspect, as the entity is often mentioned on a side.

As it can be noticed by looking at Table 4, there is in particular a drop in performance for the supervised approaches, in comparison to the results with sentence context. This is mainly due to two reasons: *a)* often different unsupervised models produce very similar rankings and therefore provide no new information to the L2R systems; *b)* moreover semantic features, which capture the meaning of the paragraph, tend to add noise to the model.

Take-Away. Having more context surrounding the mention often hurts the performance of the tested approaches, as the topic of the

Table 5: Precision at 1 (P@1) and mean average precision (MAP) on aspect-linking using section context.

Model	P@1	MAP
random baseline	0.16	0.41
Simple Heuristics		
size	0.39	0.60
content overlap	0.46	0.66
entity overlap	0.52	0.70
Header		
tf-idf (cs) †⋄	0.41	0.62
BM25 †	0.37	0.59
w-emb (cs) †	0.27	0.49
L2R (Header)	0.41	0.62
Content		
tf-idf (cs) †⋄	0.53	0.70
BM25 †⋄	0.50	0.69
w-emb (cs) †⋄	0.45	0.64
L2R (Content)	0.49	0.68
Entity		
tf-idf (cs) †⋄	0.51	0.67
BM25 †	0.52	0.69
RDF2Vec (cs) †	0.48	0.65
L2R (Entity)	0.48	0.66
Our Approach		
EAL (All) †	0.55	0.72
EAL (Selected) ⋄	**0.57**	**0.73**

paragraph is not always relevant for the linking task (**RQ1**). While both our methods outperform all other L2R approaches, they do not significantly improve the performance of their strongest feature, namely tf-idf (cs) with content representation of the aspect (**RQ2**).

5.5 Aspect-Linking with Section Context

The final set of experiments we conduct on our new dataset adopts an entire section as a context for the EAL task. As already remarked at paragraph level, during our experiments we noticed that most of the time the entity is not central to the context of the section. For this reason, as we do not perform any entity-filtering (as opposed to previous work [13]), the surrounding text adds a large amount of noise to the linking task. To better understand this issue, consider the following example, where Barack Obama is mentioned in the description of the attempt of the city of Chicago in organizing the 2016 Summer Olympics.

> ***President*** *and First Lady traveled to Denmark to support Chicago's bid for the 2016 Summer Olympics.*

While the sentence highlights the aspect of the entity Barack_Obama that is most related to the mention (in boldface), i.e., its presidency, the rest of the surrounding context is not relevant if not misleading for the task. For instance, the city "Chicago" – which is often mentioned in the section – is relevant for another aspect of the entity which is only slightly related, namely the fact that Obama studied and worked in the city, and was senator of Illinois.

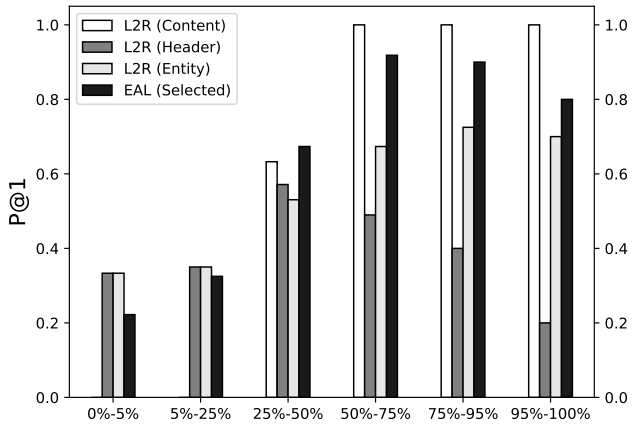

Figure 3: Difficulty-test for P@1, comparing L2R (Content) to other learning to rank systems with sentence context.

The impact of this can be noticed by looking at the results in Table 5, which show a drastic drop in performance for the majority of the approaches (**RQ1,3**). In an in-depth study we noticed in particular how semantic features derived from the content become drastically less informative in cases of large contexts.

Take-Away. Nevertheless, the two settings of our EAL system both outperform all other L2R baselines. This shows again the advantages of fruitfully combining the different types of features and aspect representations in a learning-to-rank setting (**RQ2**). However, it is also important to remark that, in this very noisy setting, a simple heuristics such as measuring the overlap between the entities provides already good results.

5.6 Discussion

In this experimental section we have shown how it is possible to link entity mentions in context to the correct aspect, reaching in certain settings high performance (both in terms of MAP and P@1 - **RQ1**). The use of content features by themselves for performing entity-aspect linking is a very strong baseline; however to deal with complex scenarios (from sentence to section levels), the combination of these features with the ones derived from other types of aspect representations often enables to our EAL systems to significantly improve the performance (**RQ2**). To better understand the benefit of this, we present in Figure 3 an additional test where we have divided all queries into different levels of difficulty according to the performance (P@1) achieved by the L2R (Content) at sentence level: the 5% most difficult queries for this method are to the left, the 5% easiest queries to the right. By comparing the performance of the L2R (Content) to the other L2R methods in the different bins it emerges that whenever it is difficult to perform the task with content features, other features derived from the heading and the entity support our learning-to-rank system.

It is also important to remark on the fact that different context-sizes have a serious impact on the performance of all methods (**RQ3**); due to this fact our error-analysis suggests that a good strategy is to employ as context only the most relevant one to three sentences surrounding each mention (**RQ4**). This take-away is not

Table 6: MAP and P@5 regarding entity ranking for web queries on Robust04 and ClueWeb12.

	Robust04		ClueWeb12	
Model	**MAP**	**P@5**	**MAP**	**P@5**
REWQ [40]	-	0.79	-	**0.84**
Fusion [14]	0.57	0.85	0.39	0.82
EAL	**0.73**	**0.93**	**0.57**	0.74

only important for the entity-linking task presented in this paper, but could be relevant also to the researchers interested in automatic Wikipedia enrichment [1, 13]: the relation between documents and entity sections may be better captured by considering only a few relevant sentences surrounding the mention, instead of using the entire document as contextual information.

Given the promising results achieved at sentence-level, in the next section we additionally test the performance of our EAL system in a few downstream applications relevant for the digital library community, from query-expansion to tweet classification.

6 EAL: DOWNSTREAM APPLICATIONS

After presenting the results of our experiments, we now examine a series of applications of the obtained entity-aspects in digital libraries, derived with our **EAL (Selected)** approach.

6.1 Predicting Latent Entities for a User Query

As a first type of application, we explore the usefulness of entity-aspects for the task of ranking latent relevant entities, given a user query. This task supports both the suggestion of named entities related to a search and the expansion of the query with semantic information from a knowledge base such as DBpedia.

Approach. In this setting, we rank each candidate entity, based on its most related section to the user query.

Dataset. We conduct experiments for measuring the usefulness of this approach on the two datasets for the task presented by Schuhmacher et al. [40] and extended by Foley et al. [14] (Robust04 and ClueWeb12). In these resources, each user-query is associated with a list of entities. For example the query "oic balkans 1990s" is associated with the relevant entities Bosnia-Herzegovina and Organisation_of _Islamic_Cooperation.

Comparison. In Table 6 we report the results of the currently two best systems for the task in terms of MAP and precision at 5 (P@5): REWQ [40] employs a learning-to-rank approach based upon features of entities linked in pseudo-relevant documents; The Fusion approach presented by Foley et al. [14] also employs learning-to-rank but, instead of relying on entity linking annotations, uses lexical features in a minimal linguistic resource setting. Since we adopt the extended version of the dataset created by Foley et al., for REWQ we are able to report only the precision at 5.

Results. The results of our system (EAL), presented in Table 6, show the usefulness of employing aspects for ranking entities, given a user query. In fact, we obtain a drastic improvement on the Robust04 dataset compared to the state-of-the-art approaches, and our method also shows good performance on ClueWeb12, outperforming Fusion [14] in term of MAP.

Table 7: MAP and P@5 regarding entity ranking for named events using dataset from Nanni et al. [27].

Model	**MAP**	**P@5**
RDF2Vec (cs)	0.65	0.80
EAL (EvAsp)	0.74	0.73
EAL (EntAsp)	**0.82**	**0.90**

6.2 Predicting Relevant Entities for an Event

Another relevant task that could be addressed through the use of entity-aspects is to suggest relevant entities, given a specific named event. Knowing which entities are the most relevant supports the creation of comprehensive event collections by capturing even the documents that do not explicitly mention the event name [26].

Approach. The system we present for this task follows the intuition that only a few aspects (i.e., sections) of an entity will be related to an event and vice-versa.

We rank entities in two different ways. First, by the similarity between the entity-name and all sections on the Wikipedia page of the event. We call this approach **EvAsp**, as it employs event-aspects. Second, by the event-name and all sections on the Wikipedia page of the entity, which we call **EntAsp**. While it is important to remark that our EAL approach has not been designed for dealing with such small context (i.e., only the name of the event/entity is given), our intention is to assess whether it could benefit from its combination of lexical and semantic features.

Dataset. We adopt the event-entity dataset introduced in our previous work [27], where global events such as the Orange Revolution are associated with a set of potentially relevant entities, which have to be ranked.

Comparison. In our previous work [27], we evaluated different approaches and found that ranking by the cosine similarity of pre-computed RDF graph embedding representations [38] of entities and events is an efficient solution for the task. However, we suspect that using pre-computed vector representations is not always an ideal approach, as new and long-tail entities are not present in the entity embeddings vocabulary. We evaluate the performance of the systems in terms of MAP and p@5.

Results. As can be seen in Table 7, the use of aspects reveals to be an efficient solution also for ranking entities, given a named event. As a matter of fact, in the EntAsp setting, this approach significantly outperforms RDF2Vec both for what concerns MAP and P@5. This boost in performance is given in particular by the use of semantic features (word embeddings) for measuring the relatedness between the event name and the content of the entity sections.

6.3 Predicting Event-Aspects for Tweets

As the last downstream application tested in this paper, we examine the use of entity aspects to improve the organization and navigation of topic-based collections. In our previous work [27], we focused on how to help digital librarians in the creation of event-collections, by automatizing the building process; we addressed in particular how to obtain a comprehensive coverage of each event analyzed by also capturing their premises and consequences. We envision the integration of these two technologies as the next step of our work.

We adopt for this experiment the large-scale collection that Zhang et al. [49] have recently created and made available to the research community regarding the Brexit referendum. It comprises around 12 million tweets and 240,000 news articles related to the event, crawled starting from 30 days before the vote. The researchers have entity-linked each tweet and assigned it to a DBpedia category that should capture the most relevant sub-topic of the event, such as Category:Immigration_to_Europe or Category:European_Union_Law.

However, while examining the dataset, we noticed that the associated categories often identify very general themes and are not useful for structuring the collection in sub-topics (for instance, many tweets are associated with Category:Euroscepticism_in_the_United_Kindgom). For this reason, in this final application we explore whether the use of entity aspects could be an alternative approach for describing the topics of the tweets and for facilitating the exploration of the collection.

Approach. To represent the topic Brexit, we consider the eight aspects of the entity Issues_in_the_United_Kingdom_European_Union_Membership_Referendum,_2016 as they describe the event and offer a clear overview of the major stories discussed during the campaign (see Table 8). Our EAL approach is then used to assign each tweet to one of the different aspects.

Dataset. We took a sample of 750 tweets in English, randomly selected from the Brexit collection [49]. For each tweet, we asked three human experts to assign it, when possible, to the most related section of the page. We obtained a good agreement between the annotators with an inter-annotator agreement measured in Cohen's kappa of 0.66. Annotators also flagged tweets that were too general (270) for the task or not related to the event (106), for example:

*The end is near #**brexit***

*Congrats!!! ICELAND 2-1 **Brexit** in Euro 2016*

This analysis already indicates the complexity of organizing tweets in event sub-topics: in fact, almost half of the messages are too general to be classified or not related to the event. We decided to remove these tweets from the dataset when conducting the final evaluation, which left us with 372 tweets, each of them assigned to a specific section depending on the aspect of the event Brexit that was most relevant.

Comparison. We test the quality of our approach in comparison with two types of baselines:

- **Ranking Approaches.** We report the results of two of the best performing baselines from the entity-aspect linking task, namely: BM25 and word-embedding (cosine similarity) between the tweet and the content of the section;
- **Classification Approaches.** As additional hard-to-beat baselines, we treat this problem as a classification task, where each tweet is associated with a label (i.e., the section-heading). We compare two types of classifiers (Naive-Bayes and Support Vector Machine) with different features (tf-idf and word embeddings) in a 10-fold cross validation setting.

Results. As can be seen in the results reported in Table 9, the performance of all approaches reveal the high complexity of the task, where it is necessary to identify the topic of a political message, which is often conveyed in just a few words. Nevertheless, our EAL approach outperforms all other methods and in particular the classifiers trained with in-domain data. Once again, it is important to remark that this is due to the fact that our learning-to-rank system combine both lexical and semantic features and different representation of aspects. These initial results motivate us in further exploring the use of aspects for structuring event collections in explicit sub-topics and in improving our system for addressing these specific challenges.

Table 8: Statitics on the annotation of Brexit aspects.

Topic	# Tweets
Economy	155
Immigration	52
Sovereignty and influence	50
Security, law enforcement and defense	3
Risk to the Unity of the United Kingdom	30
Transatlantic Trade and Investment Partnership	5
Enlargement of the European Union	12
Proposed consequences of a vote to leave	65
Total	372
Excluded	
General	270
Out-of-topic	108

Table 9: Precision at 1 on Brexit-aspect dataset.

Model	P@1
random baseline	0.12
Ranking Approaches	
Content - BM25	0.37
Content - w-emb (cs)	0.36
EAL	**0.42**
Classification Approaches	
Naive Bayes (tf-idf)	0.27
SVM (tf-idf)	0.27
Naive Bayes (w-emb)	0.38
SVM (w-emb)	0.37

7 CONCLUSION

In this paper we presented and addressed the task of entity-aspect linking. Given a mention of an entity in a context, we developed a method to link it to its most related Wikipedia section. We showed how our method significantly outperforms several established baselines and delivers good results in different contexts (a sentence, a paragraph, or a section). When applying entity-aspect links to a series of relevant tasks for the digital library community – query- and event-based entity-ranking and sub-topics collection organization – we achieve state-of-the-art improvements. This demonstrates that our method, which can be applied to large-scale corpora, provides an effective way to gain detailed topical insights into a text collection. Refining entity-links in text with aspect information opens now the way to the next step of our research, which will be focused

on combining the linked aspects of multiple entities in the same context for generating an explicit description of the topic addressed in text. This type of output will be useful for digital librarians who intend to offer semantically richer collections and will also support the adoption of advanced computational approaches in the humanities and social sciences.

Acknowledgements

This work was funded in part by a scholarship of the Eliteprogramm for Postdocs of the Baden-Württemberg Stiftung (project "Knowledge Consolidation and Organization for Query-specific Wikipedia Construction"). Furthermore, this work was partially funded by the Junior-professor funding programme of the Ministry of Science, Research and the Arts of the state of Baden-Württemberg (project "Deep semantic models for high-end NLP application").

REFERENCES

[1] Siddhartha Banerjee and Prasenjit Mitra. 2015. WikiKreator: Improving Wikipedia Stubs Automatically.. In *Proc. of ACL*. 867–877.
[2] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia – A Crystallization Point for the Web of Data. *Journal of Web Semantics* 7, 3 (2009), 154–165.
[3] David M Blei. 2012. Topic modeling and digital humanities. *Journal of Digital Humanities* 2, 1 (2012), 8–11.
[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.
[5] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. of SIGMOD*. 1247–1250.
[6] Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. *Proc. of EMNLP* (2014).
[7] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*. 288–296.
[8] Luciano del Corro, Abdalghani Abujabal, Rainer Gemulla, and Gerhard Weikum. 2015. Finet: Context-aware fine-grained named entity typing. In *Proc. of ACL*.
[9] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity query feature expansion using knowledge base links. In *Proc. of SIGIR*.
[10] Laura Dietz and Ben Gamari. 2017. TREC CAR: A Data Set for Complex Answer Retrieval. Version 1.5. (2017). http://trec-car.cs.unh.edu
[11] Pavlos Fafalios, Helge Holzmann, Vaibhav Kasturia, and Wolfgang Nejdl. 2017. Building and Querying Semantic Layers for Web Archives. In *Proc. of JCDL*, Vol. 2017.
[12] Paolo Ferragina and Ugo Scaiella. 2010. TagMe: on-the-fly annotation of short text fragments (by Wikipedia entities). In *CIKM*. 1625–1628.
[13] Besnik Fetahu, Katja Markert, and Avishek Anand. 2015. Automated news suggestions for populating wikipedia entity pages. In *Proc. of CIKM*. 323–332.
[14] John Foley, Brendan O'Connor, and James Allan. 2016. Improving Entity Ranking for Keyword Queries. In *Proc. of CIKM*. 2061–2064.
[15] Francesca Frontini, Carmen Brando, and Jean-Gabriel Ganascia. 2015. Semantic web based named entity linking for digital humanities and heritage texts. In *First International Workshop on the Semantic Web for Scientific Heritage at ESWC-15*.
[16] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Unsupervised cross-lingual scaling of political texts. In *Proc. of EACL*.
[17] Hongzhao Huang, Larry Heck, and Heng Ji. 2015. Leveraging deep neural networks and knowledge graphs for entity disambiguation. *arXiv preprint arXiv:1504.07678* (2015).
[18] Nattiya Kanhabua, Philipp Kemkes, Wolfgang Nejdl, Tu Ngoc Nguyen, Felipe Reis, and Nam Khanh Tran. 2016. How to search the internet archive without indexing it. In *Proc. of TPDL*. Springer, 147–160.
[19] Anne Lauscher, Federico Nanni, Pablo Ruiz Fabo, and Simone Paolo Ponzetto. 2016. Entities as topic labels: combining entity linking and labeled LDA to improve topic interpretability and evaluability. *IJCol-Italian journal of computational linguistics* 2, 2 (2016), 67–88.
[20] Peng Li, Yinglin Wang, Wei Gao, and Jing Jiang. 2011. Generating aspect-oriented multi-document summarization with event-aspect model. In *Proc. of EMNLP*. 1137–1146.
[21] Julio Cesar Amador Diaz Lopez, Sofia Collignon-Delmar, Kenneth Benoit, and Akitaka Matsuo. 2017. Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data. *Statistics, Politics and Policy* 8, 1 (2017), 85–104.
[22] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proc. of Semantic Systems*. ACM, 1–8.
[23] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics* 2 (2014), 231–244.
[24] Federico Nanni, Hiram Kümper, and Simone Paolo Ponzetto. 2016. Semi-supervised Textual Analysis and Historical Research Helping Each Other: Some Thoughts and Observations. *International Journal of Humanities and Arts Computing* 10, 1 (2016), 63–77.
[25] Federico Nanni, Bhaskar Mitra, Matt Magnusson, and Laura Dietz. 2017. Benchmark for Complex Answer Retrieval. *Proc. of ICTIR* (2017).
[26] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. 2016. Entity relatedness for retrospective analyses of global events. In *NLP+CSS Workshop at WebSci*.
[27] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. 2017. Building entity-centric event collections. In *Proc. of JCDL*. IEEE, 1–10.
[28] Federico Nanni, Yang Zhao, Simone Paolo Ponzetto, and Laura Dietz. 2017. Enhancing Domain-Specific Entity Linking in DH. *Digital Humanities* (2017).
[29] Fabian G Neuner and Christopher Wratil. 2017. The myth of the boring election: populism and the 2017 German election. *LSE European Politics and Policy (EUROPP) Blog* (2017), 1–5.
[30] Andrea Giovanni Nuzzolese, Anna Lisa Gentile, Valentina Presutti, Aldo Gangemi, Darío Garigliotti, and Roberto Navigli. 2015. Open knowledge extraction challenge. In *Semantic Web Evaluation Challenge*. Springer, 3–15.
[31] Alex Olieman, Kaspar Beelen, Milan van Lange, Jaap Kamps, and Maarten Marx. 2017. Good Applications for Crummy Entity Linkers? The Case of Corpus Selection in Digital Humanities. *Proc. of Semantic Systems* (2017).
[32] Patrick Pantel, Thomas Lin, and Michael Gamon. 2012. Mining entity types from query logs via user intent modeling. In *Proc. of ACL*. 563–571.
[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *Proc. of EMNLP*, Vol. 14.
[34] Thierry Poibeau and Pablo Ruiz. 2015. Generating navigable semantic maps from social sciences corpora. *Digital Humanities* (2015).
[35] Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*. Springer, 93–115.
[36] Ridho Reinanda, Edgar Meij, and Maarten de Rijke. 2015. Mining, ranking and recommending entity aspects. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 263–272.
[37] Ridho Reinanda, Edgar Meij, and Maarten de Rijke. 2016. Document filtering for long-tail entities. In *Proc. of CIKM*. ACM, 771–780.
[38] Petar Ristoski and Heiko Paulheim. 2016. Rdf2vec: Rdf graph embeddings for data mining. In *Proc. of ISWC*. Springer.
[39] Marco Rovera, Federico Nanni, Simone Paolo Ponzetto, and Anna Goy. 2017. Domain-specific named entity disambiguation in historical memoirs. In *Proc. of Clic-IT*.
[40] Michael Schuhmacher, Laura Dietz, and Simone Paolo Ponzetto. 2015. Ranking Entities for Web Queries Through Text and Knowledge. In *Proc. of CIKM*.
[41] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proc. of WWW*. 697–706.
[42] Bilyana Taneva and Gerhard Weikum. 2013. Gem-based entity-knowledge maintenance. In *Proc. of CIKM*. ACM, 149–158.
[43] Theo van Veen, Juliette Lonij, and Willem Jan Faber. 2016. Linking Named Entities in Dutch Historical Newspapers. In *Proc. of MTSR*. Springer, 205–210.
[44] Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, German Rigau, and others. 2016. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems* 110 (2016), 60–85.
[45] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proc. of ICML*. ACM, 1105–1112.
[46] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. *Proc. of CoNLL* (2016).
[47] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2017. Learning Distributed Representations of Texts and Entities from Knowledge Base. *Transactions of the Association for Computational Linguistics* (2017).
[48] Xiaoxin Yin and Sarthak Shah. 2010. Building taxonomy of web search intents for name entity queries. In *Proc. of WWW*. ACM, 1001–1010.
[49] Lei Zhang, Maribel Acosta, Michael Färber, Steffen Thoma, and Achim Rettinger. 2017. BreXearch: Exploring Brexit Data Using Cross-Lingual and Cross-Media Semantic Search. In *Proc. of ISWC*.