# What's New? Analysing Language-specific Wikipedia Entity Contexts to Support Entity-Centric News Retrieval

Yiwei Zhou[1], Elena Demidova[2], and Alexandra I. Cristea[1]

[1] Department of Computer Science, University of Warwick, Coventry, UK
{Yiwei.Zhou, A.I.Cristea}@warwick.ac.uk
[2] University of Southampton, UK and L3S Research Center, Germany
demidova@L3S.de

**Abstract.** Representation of influential entities, such as celebrities and multinational corporations on the web can vary across languages, reflecting language-specific entity aspects, as well as divergent views on these entities in different communities. An important source of multilingual background knowledge about influential entities is Wikipedia — an online community-created encyclopaedia — containing more than 280 language editions. Such language-specific information could be applied in entity-centric information retrieval applications, in which users utilise very simple queries, mostly just the entity names, for the relevant documents. In this article we focus on the problem of creating language-specific entity contexts to support entity-centric, language-specific information retrieval applications. First, we discuss alternative ways such contexts can be built, including *Graph-based* and *Article-based* approaches. Second, we analyse the similarities and the differences in these contexts in a case study including 220 entities and five Wikipedia language editions. Third, we propose a context-based entity-centric information retrieval model that maps documents to aspect space, and apply language-specific entity contexts to perform query expansion. Last, we perform a case study to demonstrate the impact of this model in a news retrieval application. Our study illustrates that the proposed model can effectively improve the recall of entity-centric information retrieval while keeping high precision, and provide language-specific results.

## 1 Introduction

Entities with world-wide influence, such as celebrities and multinational corporations, can be represented differently on webpages or in other documents originating from various cultures and written in different languages. These various representations can reflect language-specific entity aspects as well as views on the entity in different language-speaking communities. In order to enable a better representation of the language-specific entity aspects for information retrieval applications, methods to systematically identify an entity's context — i.e. the aspects in the entity's descriptions typical in a specific language — need to be developed.

For example, in the English news articles, the entity "Angela Merkel", the Chancellor of Germany, is often associated with US and UK politicians such as "Barack Obama" and "David Cameron". Also, recent discussions of European importance, such as the Greek financial situation, are included. On the contrary, although the news articles from German media also include European topics, they frequently focus on the domestic political topics, featuring discussions of political parties in Germany, scandals arising around German politicians, local elections, finances and other country-specific topics. Taking another example, in the case of multinational companies, such as GlaxoSmithKline (a British healthcare company), aspects related to the local activities are prevalent in the news articles in specific languages. These aspects range from the effectiveness of the various vaccines developed by the company, to the sports events sponsored by this company in a specific country.

In this article we focus on the problem of creating language-specific entity contexts to support entity-centric, language-specific information retrieval applications. Whilst this work is based on our prior research [29], where we introduced language-specific entity contexts, this article additionally introduces a *context-based entity-centric information retrieval model* that applies these contexts to improve the overall recall and provides language-specific results, which differs our research from all former relevant studies.

In order to obtain a comprehensive overview over the language-specific entity aspects and their representations , multilingual background knowledge about this entity is required. We choose Wikipedia as a knowledge base to obtain such background knowledge. Wikipedia — a multilingual encyclopaedia available in more than 280 language editions — contains language-specific representations of millions of entities and can provide a rich source for cross-cultural analytics. For example, recent studies focused on the manual analysis of the controversial entities in Wikipedia and identified significant cross-lingual differences (e.g. [22]). As entity representations in different Wikipedia language editions can evolve independently, they often include overlapping as well as language-specific entity aspects. We discuss different ways of creating these contexts using Wikipedia, including *Article-based* and *Graph-based* approaches and propose a measure to compute the context similarity. We use this measure to analyse the similarities and the differences of the language-specific entity contexts in a case study using 220 entities of four different entity types and the representations of these entities in five languages. Demonstrating the benefits of finding entity contexts for five languages has been considered enough to illustrate the principles and methods, which can then be transferred to other languages and other entities. Our experiments show that the proposed *Graph-based* entity context can effectively provide a comprehensive overview over the language-specific entity aspects.

Moreover, we propose a *context-based information retrieval model*, which applies language-specific contexts to support entity-centric information retrieval applications. This model enables retrieval of the documents that describe information relevant to the entity, even if the entity is not mentioned by name. This information can include relevant events, which are likely to impact the en-

tity, or are otherwise related to it. At the same time, while using this model, the relevance of the documents mentioning the entity name is only marginally reduced. We implement the proposed model on an information retrieval application, which includes: (i) targeted retrieval of entity-centric information using language-specific contexts; (ii) an overview of the language-specific entity aspects in the each retrieved document. We perform a case study to demonstrate the impact of this model in the context of news articles retrieval through the application. Our results illustrate that language-specific — and in particular *Graph-based* — entity contexts can enhance the recall of the information retrieval application, while keeping high precision, providing positive results are news articles that describe current events relevant to the entity, even if the entity is not explicitly mentioned. We further propose *Result Specificity* to measure the level of language specificity of the proposed information retrieval model when serving users with different language backgrounds. Our results show that our context-based information retrieval model is able to provide highly language-specific results through exploiting language-specific contexts.

## 2 Creation of the Language-Specific Entity Context

In this section we define the language-specific entity context, present a measure of the context similarity and discuss alternative ways to extract such contexts from the multilingual Wikipedia.

### 2.1 Language-specific Entity Context Definition

We define the language-specific entity context as follows:

**Definition 1** *The context $C(e, l_i)$ of the entity $e$ in the language $l_i$ is represented through the weights of aspects $\{a_1, \ldots, a_M\}$ relevant to $e$, $C(e, l_i) = (w(a_1, e, l_i), \ldots, w(a_M, e, l_i))$.*

In this article, we consider entity aspects being *noun phrases* that co-occur with the entity in a given language. In addition, we can also consider the titles of the *in-linked* Wikipedia articles as an additional source of the entity aspects. The weights of the aspects are based on two factors: (1) the language-specific *aspect frequency*, i.e. the frequency of the co-occurrence of the aspect and the entity in a language, and (2) the *language frequency* — the number of languages in which the entity contexts contain the aspect. The first weighting factor prioritises the aspects that frequently co-occur with the entity in a particular language. The second factor assigns higher weights to the language-specific aspects of the entity not mentioned in many other languages.

Inspired by the term frequency–inverse document frequency ($tf - idf$), given a multilingual data collection containing the languages $\mathbb{L} = \{l_1, \ldots, l_N\}$, the weight $w(a_k, e, l_i)$ of the entity aspect $a_k$ in the language-specific context $C(e, l_i)$ is calculated as follows:

$$w(a_k, e, l_i) = af(a_k, e, l_i) \cdot log \frac{N}{lf(a_k, e, \mathbb{L})}, \quad (1)$$

where $af(a_k, e, l_i)$ is the language-specific *aspect frequency*, which represents the frequency of the co-occurrence of the aspect $a_k$ and the entity $e$ in $C(e, l_i)$; $N$ is the number of languages in the multilingual collection $\mathbb{L}$; $lf(a_k, e, \mathbb{L})$ is the *language frequency*, which represents the number of languages from $\mathbb{L}$ in which the contexts of $e$ contain the aspect $a_k$.

As for the titles of the manually-defined *in-linked* articles, which are also used to represent aspects, we assign them with an average language-specific *aspect frequency* computed for the noun phrases.

### 2.2 Context Similarity Measure

In order to compute the similarity between language-specific entity contexts, we use a vector space model. Each axis in the vector space represents an aspect $a_k$. We represent the context $C(e, l_i)$ of the entity $e$ in the language $l_i$ as a vector in this space. An entry for $a_k$ in the vector represents the weight of the aspect $a_k$ in $C(e, l_i)$. Then the $e$'s context similarity between languages $l_i$ and $l_j$ is computed as the cosine similarity of the context vectors:

$$Sim(C(e, l_i), C(e, l_j)) = \frac{C(e, l_i) \cdot C(e, l_j)}{|C(e, l_i)| \times |C(e, l_j)|}. \quad (2)$$

In order to allow for cross-lingual similarity computations, we represent the aspects in a common language using machine translation. To simplify the description in this article, we always refer to the original language of the entity context, keeping in mind that the similarity is computed in a common language representation.

### 2.3 Article-based Context Creation

Wikipedia articles describing an entity in different language editions (i.e. the articles that use the named entity as titles) can be a useful source for the creation of the language-specific context vectors. Thus, we first propose the *Article-based* context creation approach, which simply uses the articles describing the entity in different language editions of Wikipedia. We use all sentences from an article describing the entity in a language edition as the only source of the *Article-based* language-specific entity context.

One drawback of this approach is the possible limitation of the aspects coverage due to the incompleteness of the Wikipedia articles. Such incompleteness can be more prominent in some language editions, making it difficult to create fair cross-lingual comparisons. For example, when reading the English Wikipedia article about the entity "Angela Merkel", a lot of basic aspects about this politician, such as her background and early life, her domestic policy and her foreign affairs, are provided. However, not all aspects about Angela Merkel occur in this

Wikipedia article. We can observe, that other articles in the same Wikipedia language edition mention other important aspects. For example, the Wikipedia article about "Economic Council Germany" mentions Angela Merkel's economic policy: "Although the organisation is both financially and ideologically independent it has traditionally had close ties to the free-market liberal wing of the conservative Christian Democratic Union (CDU) of Chancellor Angela Merkel.". Even the English Wikipedia article about an oil painting, "The Nightmare", which does not seem connected to "Angela Merkel" at the first glance, also mentions "Angela Merkel" as: "On 7 November 2011 Steve Bell produced a cartoon with Angela Merkel as the sleeper and Silvio Berlusconi as the monster." The aspects contained in the examples above do not occur in the English Wikipedia article entitled "Angela Merkel". As this example illustrates, just employing the Wikipedia article describing the entity can not entirely satisfy the need to obtain a comprehensive overview over the language-specific aspects.

### 2.4 Graph-based Context Creation

To alleviate the shortcomings of the *Article-based* approach presented above and obtain a more comprehensive overview of the entity aspects in the entire Wikipedia language edition (rather than in a single article), we propose the *Graph-based* context creation approach. The idea behind this approach is to use the link structure of Wikipedia to obtain a comprehensive set of sentences mentioning the target entity and to use this set to create the context. To this extent, we use the *in-links* to the Wikipedia article describing the entity and the *language-links* of this article to efficiently collect the articles that are probable to mention the target entity in different language editions. We extract the sentences mentioning the target entity using state-of-the-art named entity disambiguation methods and use these sentences to create language-specific contexts.

To illustrate our approach, we use the creation of the context in the English edition of Wikipedia for the entity "Angela Merkel" as an example. For the Wikipedia article in English entitled "Angela Merkel", there are several *in-links* from other articles in English that mention the entity. Besides that, there are also *language-links* from the articles describing "Angela Merkel" in other Wikipedia language editions to this entity's English Wikipedia article.

In Fig. 1, we use the arrows to represent the *in-links*, and dashed lines to represent the *language-links*. The black nodes represent articles in English, while the white nodes represent the articles in other languages.

Overall, the creation of the *Graph-based* context for "Angela Merkel" using these links includes the following steps:

1. *Graph Creation.* We create a subgraph for "Angela Merkel" from Wikipedia's link structure in the following way: we first expand the node set from the article in English describing the entity (the central node) to all language editions of this Wikipedia article; we further expand the node set with all the articles having *in-links* to the nodes in the node set; we finally expand the node set with all the articles having *language-links* to the existing nodes

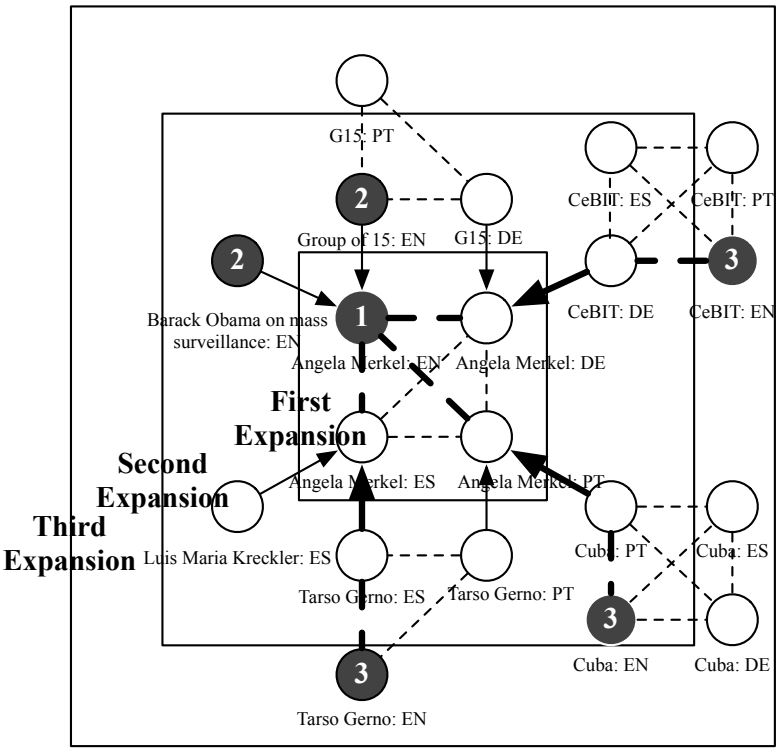


Fig. 1: An Example of the Graph-Based Context Creation from the English Wikipedia for the entity "Angela Merkel".

in the node set, if they have not been included into the node set yet. Different types of edges are also added between the nodes based on the *in-link* and the *language-link* relationships.

2. *Article Extraction.* To efficiently extract as many mentions of Angela Merkel from the English Wikipedia as possible, we first extract the article of the central node (with number 1 in Fig. 1), and then start the graph traversal from it.

   Second, all the articles in the graph that have paths of length 1, and the path types are *in-link* to the central node (with number 2 in Fig. 1), are extracted;

   Third, all the articles in the graph that have paths of length 3, the articles are in English, and the path types are *language-link* — *in-link* — *language-link* (marked as bold lines in Fig. 1) to the central node (with number 3 in Fig. 1), are also extracted. These articles, although they do not have the direct *in-link* paths to the central node, are in English and their other language editions have *in-links* to articles describing "Angela Merkel" in other languages; Therefore, these articles are likely to mention "Angela Merkel".

3. *Context Construction.* We employ DBpedia Spotlight [13] to annotate the extracted articles to identify the sentences mentioning "Angela Merkel". All the noun phrases extracted from sentences mentioning "Angela Merkel", form the English *Graph-based* context of "Angela Merkel". In addition, we add the names of the linked articles to the entity context.

## 3 News Retrieval using Language-Specific Entity Context

In this section we present several retrieval scenarios for answering entity-centric queries over news collections in a common language. Then we describe our approach to address selected scenarios using language-specific entity contexts presented in Section 2.

### 3.1 Entity-centric News Retrieval

When users are interested in the current news about a named entity, they could simply provide the entity name as the query to a retrieval application. On a daily basis, only a limited number of news articles that explicitly mention this named entity are published. However, one named entity is typically related to various other named entities and events, as we observed during the context creation using the Wikipedia link structure. This kind of relationships among entities and events is represented by the entity contexts, the creation of which we described in Section 2. Our intuition is that by using these contexts, we could significantly increase recall of retrieved documents for the entity-centric queries in a news retrieval application, while keeping high precision. Moreover, some documents could only marginally mention an entity, without providing any comprehensive information for the specific entity; In these cases, the entity's contexts can help retrieval application to focus on more relevant documents.

When only using an entity name as the query, traditional information retrieval systems can only return news articles with the named entity's occurrence, which can barely satisfy the users' needs of a comprehensive knowledge about the named entity. For example, when using "Angela Merkel" as the query, it would be beneficial to return news articles like `http://www.thelocal.de/20151202/germany-fear-terrorism-if-army-fights-in-syria`, which describe the current situation in Germany. Although the content of this article contains neither the term "Angela" nor the term "Merkel", it reports about an event that has potentially large impact on her political decisions. In order to tackle this problem, our context-based information retrieval model incorporates the entity's contexts from Wikipedia into the search and ranking process. As a result, the articles that are similar to the entity context will obtain higher ranks, even if the entity is not mentioned explicitly.

While using entity contexts for retrieval applications, the relevance of a news article to a named entity may be controversial among people with different language backgrounds. For example, a news article containing information about a recent VW scandal affecting the biggest German

car production company `http://www.thelocal.de/20151202/what-the-vw-scandal-means-for-germanys-economy` can be considered as relevant by most German people, as they can think that the German Chancellor should take direct measures to boost the national economy hurt by the scandal. However, the relevance of this article to the query "Angela Merkel" can be considered to be low among the English-speaking communities. These users could think this to be a company problem, and it could be hard for them to understand if this scandal has a big impact at the national level.

We tackle this problem by using language-specific entity contexts in news retrieval. The users of the retrieval application can select their preferred language-specific entity context when searching for a named entity. The returned news articles and their ranks are then language-specific, based on the background knowledge from the corresponding language edition of Wikipedia.

Besides the retrieval of relevant articles, it is also useful to provide information regarding the aspects of the entity influencing their relevance. That is particularly important in case the entity itself is not mentioned in the article. Our context-based information retrieval model addresses this problem by creating an overview of each news article discussing language-specific entity aspects and uses the most relevant entity aspects to annotate it.

### 3.2 Approach to Context-based Information Retrieval

For each document $d_i$ from document set $\mathbb{D}$, we extract all the *noun phrases* in the text as potential related *aspects* to query entities (the named entities whose names are provided as the queries), and then index all the documents by the *aspects*. For a query entity $e$, and a chosen $l_i$, we represent each document $d_i$ by the same set of aspects in $C(e, l_i)$, with each aspect $a_k$ weighted by:

$$w(a_k, d_i) = af(a_k, d_i) \cdot log \frac{N}{lf(a_k, e, \mathbb{L})}, \tag{3}$$

where $af(a_k, d_i)$ is the number of matches of aspect $a_k$ with the noun phrases from $d_i$. In this way, $d_i$ can be represented as $C(d_i, e, l_i) = (w(a_1, d_i), \ldots, w(a_M, d_i))$.

We apply the same vector space model and similarity metric as in Section 2.2 to compute the similarity between $C(d_i, e, l_i)$ and $C(e, l_i)$. $Sim(C(d_i, e, l_i), C(e, l_i))$ will be used to measure the level of relevances between the query entity and documents. All the documents have similarities with $C(e, l_i)$ higher than a threshold will be returned. Their ranks will be decided based on $Sim(C(d_i, e, l_i), C(e, l_i))$.

The top weighted aspects in $d_i$ will also be returned to provide an overview of what aspects $d_i$ is discussing about $e$.

## 4 Entity Context Analysis

The goal of the entity context analysis is to compare the *Graph-based* and the *Article-based* context creation approaches. To this extent we analyse the simi-

larities and the differences of the contexts obtained using these approaches in a case study.

### 4.1 Dataset Description

To facilitate our analysis, we selected a total number of 220 entities with worldwide influence that evenly come from four categories as our target entities. These categories include: multinational corporations, politicians, celebrities and sports stars. For our study we selected five European languages: English, German, Spanish, Portuguese and Dutch, as our target languages. For each category, we included entities originating from the countries that use one of these target languages as official languages. As our approach requires machine translation of the contexts to enable cross-lingual context similarity computation, we chose Google Translate — a translation service that provides good quality for the involved language pairs.

Based on the approach described in Section 2, we created the entity-centric contexts for the entities in our dataset from the five Wikipedia language editions listed above using the *Graph-based* and the *Article-based* approach. The average number of sentences extracted from the Wikipedia article describing the entity using the *Article-based* approach was around 50 in our dataset. With our *Graph-based* context creation approach that utilised Wikipedia link structure to collect sentences mentioning the entity from multiple articles, the number of sentences referring to an entity was increased by the factor 20 to more than 1,000 sentences per entity in a language edition, on average. This factor reflects the effect of the additional data sources within Wikipedia we use in the *Graph-based* approach for each entity processed. The total number of sentences collected by the *Graph-based* approach is 1,196,403 for the whole dataset under consideration.

### 4.2 Context Similarity Analysis

The cross-lingual context similarity resulting from the *Article-based* and the *Graph-based* context creation approaches are presented in Table 1 and Table 2, respectively. To enable cross-lingual context similarity computation, we translated all entity contexts to English. Due to the space limitations, we present example similarity values for four selected entities (one per entity type) for the seven language pairs. In addition, we present the average similarity and the standard deviation values based on all 220 entities in the entire dataset.

From Table 1, we can observe that using the *Article-based* context creation approach, the average similarity values of the language pairs including English are always higher than those of the other language pairs. Using these computation results, we can make several observations: First, as the *Article-based* contexts are more similar to English than to other languages, the English edition builds a reference for the creation of the articles in other language editions. This can be further explained by the fact that the English Wikipedia has the largest number of users, articles, and edits compared to other language

Table 1: Cross-lingual context similarity using the *Article-based* context creation approach. The table presents the similarity values for four selected entities of different types, as well as the average similarity and the standard deviation for the whole dataset of 220 entities. The language codes representing the original context languages are as follows: "NL" — Dutch, "DE" — German, "EN" — English, "ES" — Spanish, and "PT" — Portuguese.

| | ***Article-based* cross-lingual similarity** | | | | | | |
|---|---|---|---|---|---|---|---|
| **Entity** | **EN-DE** | **EN-ES** | **EN-PT** | **EN-NL** | **DE-ES** | **DE-NL** | **ES-PT** |
| GlaxoSmithKline | 0.43 | 0.34 | 0.29 | 0.29 | 0.31 | 0.22 | 0.26 |
| Angela Merkel | 0.68 | 0.66 | 0.84 | 0.54 | 0.60 | 0.59 | 0.66 |
| Shakira | 0.71 | 0.58 | 0.84 | 0.75 | 0.48 | 0.64 | 0.58 |
| Lionel Messi | 0.71 | 0.86 | 0.81 | 0.89 | 0.71 | 0.68 | 0.82 |
| **Average of 220** | **0.50** | **0.47** | **0.43** | **0.45** | 0.38 | 0.38 | 0.37 |
| **Stdev of 220** | 0.16 | 0.17 | 0.19 | 0.19 | 0.15 | 0.16 | 0.17 |

Table 2: Cross-lingual context similarity using the *Graph-based* context creation approach. The table presents the similarity values for four selected entities of different types, as well as the average similarity and the standard deviation for the whole dataset of 220 entities. The language codes representing the original context languages are as follows: "NL" — Dutch, "DE" — German, "EN" — English, "ES" — Spanish, and "PT" — Portuguese.

| | ***Graph-based* cross-lingual similarity** | | | | | | |
|---|---|---|---|---|---|---|---|
| **Entity** | **EN-DE** | **EN-ES** | **EN-PT** | **EN-NL** | **DE-ES** | **DE-NL** | **ES-PT** |
| GlaxoSmithKline | 0.72 | 0.73 | 0.59 | 0.61 | 0.63 | 0.62 | 0.55 |
| Angela Merkel | 0.64 | 0.62 | 0.42 | 0.60 | 0.75 | 0.82 | 0.51 |
| Shakira | 0.91 | 0.94 | 0.90 | 0.88 | 0.94 | 0.91 | 0.94 |
| Lionel Messi | 0.63 | 0.76 | 0.77 | 0.68 | 0.70 | 0.62 | 0.76 |
| **Average of 220** | 0.52 | **0.59** | 0.54 | 0.50 | 0.56 | 0.52 | **0.64** |
| **Stdev of 220** | 0.24 | 0.22 | 0.21 | 0.23 | 0.23 | 0.23 | 0.19 |

editions[3]. Second, as other language pairs are less similar, the overlapping aspects between the English edition and the other language editions appear to be language-dependent. Finally, although the cosine similarity values can be in the interval [0,1], the absolute similarity values achieved by the *Article-based* approach reach at most 0.5, even for the language pairs which are supposed to have relatively high similarity, such as English and German. Such relatively low absolute similarity values indicate that although the articles contain some overlapping entity aspects, they also include a significant proportion of divergent aspects.

In contrast to the *Article-based* approach, the *Graph-based* approach collects a more comprehensive overview of the entity aspects spread across different articles in a language edition. From Table 2, we can see that the most similar

[3] `http://en.wikipedia.org/wiki/List_of_Wikipedias`

context pair is Spanish and Portuguese. Intuitively, this could be explained by the closeness of the cultures using these two languages, and a more comprehensive overview of the covered entity aspects in both languages compared to the *Article-based* approach. We can also observe that the average similarity values significantly increase compared to the *Article-based* approach and can exceed 0.6 in our dataset.

From a single entity perspective, some entities may achieve higher similarity values than the average similarity for some language pairs, when more common aspects are included in the contexts on both sides. For example, this is the case for EN-NL, DE-ES and DE-NL pairs for the entity "Angela Merkel". Other entities may have lower similarity values for some language pairs, especially when distinct aspects are included into the corresponding contexts, such as the EN-DE, ES-ES, and EN-PT pairs for "Lionel Messi".

To illustrate the differences in the language-specific *Graph-based* entity contexts, we select the highly weighted aspects from the contexts of the entity "Angela Merkel" constructed using the *Graph-based* approach, as shown in Table 3. In this table, the unique aspects that appear with high weights in all contexts of the entity "Angela Merkel" are underlined. We can observe that the aspects that appear with high weights only in the non-German context — e.g. "England", "Kingdom" and "Dilma Rousseff" — are more relevant to her international affairs in corresponding language-speaking countries. In contrast, the aspects that appear with high weights only in the German context — such as "German children" and "propaganda" — are more relevant to her domestic activities.

Overall, our observations confirm that the *Graph-based* context provides a better overview of the different entity aspects than the *Article-based* context. The *Graph-based* approach can determine the similarity values and the differences of the language-specific contexts, independent of the coverage and completeness of any dedicated Wikipedia article. The results of the t-test confirm the statistical significance of the differences in context similarity values between the *Article-based* and the *Graph-based* approaches for all language pairs except the EN-DE. This exception can be explained by a relatively high coverage of the German Wikipedia articles with respect to the aspects of the represented entities in our dataset.

The analysis results also confirm our intuition that, although the editors of different Wikipedia language editions describe some common entity aspects, they can have different focus with respect to the aspects of interest. These differences are reflected by the complementary information spread across the Wikipedia language editions and can probably be explained by various factors including the culture and the living environment of the editors, as well as the information available to them. Our *Graph-based* context creation approach is capable of capturing these differences from different language editions by creating a comprehensive language-specific aspects overview.

Table 3: Top-30 highly weighted aspects of the entity "Angela Merkel" in language-specific *graph-based* contexts.

| | |
|---|---|
| **English** | angela merkel, battle, berlin, cdu, chancellor, chancellor angela merkel, church, edit, election, emperor, empire, england, france, george, german, german chancellor angela merkel, germany, government, jesus, john, kingdom, merkel, minister, party, president, talk, union, university, utc, war |
| **German** | academy, angela merkel, article, berlin, cdu, cet, chancellor, chancellor angela merkel, csu, election, example, german, german chancellor angela merkel, german children, germany, government, kasner, merkel, minister, november, october, office, party, president, propaganda, ribbon, september, speech, time, utc |
| **Spanish** | administration, angela merkel, berlin, cdu, chancellor, chancellor angela merkel, coalition, council, country, december, decommissioning plan, decreed, election, energy, france, german, german chancellor angela merkel, german federal election, germany, government, government coalition, grand coalition, merkel, minister, october, party, president, spd, union, year |
| **Portuguese** | ali, angela merkel, bank, cdu, ceo, chairman, chancellor, chancellor angela merkel, china, co-founder, coalition, csu, dilma rousseff, german chancellor angela merkel, germany, government, government merkel, koch, leader, merkel, minister, november, october, party, petroleum, president, saudi arabia, state, union, york |
| **Dutch** | angela merkel, angela dorothea kasner, bundestag, candidate, cdu, chancellor, chancellor angela merkel, coalition, csu, december, fdp, fist, french president, german, german chancellor angela merkel, german christian democrat politician, german federal election, germany, government, majority, merkel, minister, november, october, party, president, right, spd, state, union |

## 5 Language-specific Retrieval of News Articles for Entity-centric Queries

In this section we discuss the impact of the language-specific entity contexts on a news retrieval application. Since results following the same patterns can be observed across all named entities, we randomly selected two named entities, one originated from an English speaking country, and the other one originated from a non-English speaking country, as examples to demonstrate the effectiveness of the context-based information retrieval model.

### 5.1 Dataset Description

The two named entities we chose were: "Angela Merkel" originated from Germany and "David Cameron" originated from Great Britain. To enable the comparison among different language-specific contexts of an entity, we built two datasets each containing daily news from different sources: the German media

news dataset and the British media news dataset. For the German media news dataset, we randomly sampled 300 news articles from three mainstream online English news websites' RSS feeds published on December 2nd, 2015 in Germany. These websites were: Deutsche Welle[4], Spiegel Online[5] and The Local[6]. Regarding the British media news dataset, we randomly sampled 300 news articles, from two mainstream British online English news websites' RSS feeds on December 10th, 2015 in Great Britain. These websites were: The Guardian[7] and Daily Express[8]. Then, we analysed the performance of English and German contexts of the entities "Angela Merkel" and "David Cameron" on the retrieved results for these two datasets, respectively.

We expected that there were only few news articles per day mentioning a specific entity, even if this entity was prominent. Nevertheless, daily news can contain many relevant articles that discuss events related to the entity. Therefore, we used the following criteria to annotate the articles as "Relevant":

1. *Is the named entity involved in this event?*
2. *Is the named entity one of the direct causes of this event?*
3. *Will the named entity be directly impacted by this event?*

After the annotation, 51 news articles in the German media news dataset were annotated as "Relevant" for the query "Angela Merkel". In the British media news dataset, 71 news articles were annotated as "Relevant" for the query "David Cameron"[9].

### 5.2 Precision-Recall Analysis

We used a state-of-the-art information retrieval model, BM25 [21] as a baseline. The baseline model retrieved the documents using the original query containing the entity name without expansion (i.e. "Angela" and "Merkel", "David" and "Cameron").

In Fig. 2, we present the interpolated precision achieved by the baseline and the context-based information retrieval model using different contexts at different recall levels for query entity "Angela Merkel". As we can observe in Fig. 2, although the traditional ranking algorithm based on the BM25 scores of the news articles given a query entity can maintain a relatively high precision, the highest recall it can achieve is about 0.45. That is because a lot of news articles, such as `http://www.thelocal.de/20151202/germany-to-send-1200-troops-to-aid-isis-fight`, `http://www.spiegel.de/international/europe/paris-attacks-pose-challenge-to-european-security-a-1063435.html`

[4] http://www.dw.com/en/
[5] http://www.spiegel.de/international/
[6] http://www.thelocal.de/
[7] http://www.theguardian.com/
[8] http://www.express.co.uk/
[9] The annotated datasets are accessible at: `https://github.com/zhouyiwei/WIKIIRDATA`.

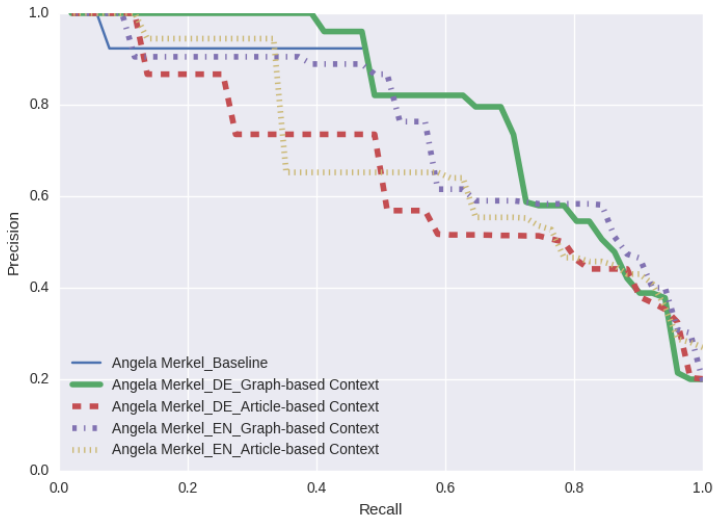


Fig. 2: Precision-Recall curve of language-specific context for "Angela Merkel".

and `http://www.thelocal.de/20151029/germany-maintains-record-low-unemployment`, report events either directly driven by "Angela Merkel", or would directly impact her. Although these articles do not mention the query entity by name, they provide indispensable insights on the query entity's current focus or past achievements, such that the users issuing this query would consider them to be relevant, especially when the number of articles mentioning the query entity is small. The context-based information retrieval model using all the contexts, no matter they are *Article-based* or *Graph-based*, no matter they are extracted from English Wikipedia or German Wikipedia, achieved higher recall for this query.

We can also observe that German(DE) *Graph-based* context achieves the overall best performance. For most of the time, it achieves higher precision than other contexts, while achieving the same recall. This is because the German(DE) *Graph-based* context provides a more comprehensive overview of the aspects of "Angela Merkel".

Moreover, the model utilising the German(DE) *Graph-based* context outperforms the baseline with respect to precision at all recall levels for this query entity. This is because "Angela" is quite a common term. By incorporating the background information from Wikipedia, the model can differentiate the Chancellor of Germany from other celebrities, such as Angela Gossow (German singer) and Angela Maurer (German long-distance swimmer), which helps to increase the precision of retrieved results.

The baseline approach ranks the news articles mostly based on the occurrences of the entity name. In contrast, our model considers all the aspects men-

tioned in the news articles about the named entity. The ranks are generated based on the similarity values between the articles' aspects and the named entity's language-specific Wikipedia context, such that news articles that provide a more comprehensive overview of language-specific aspects are prompted to higher ranks.

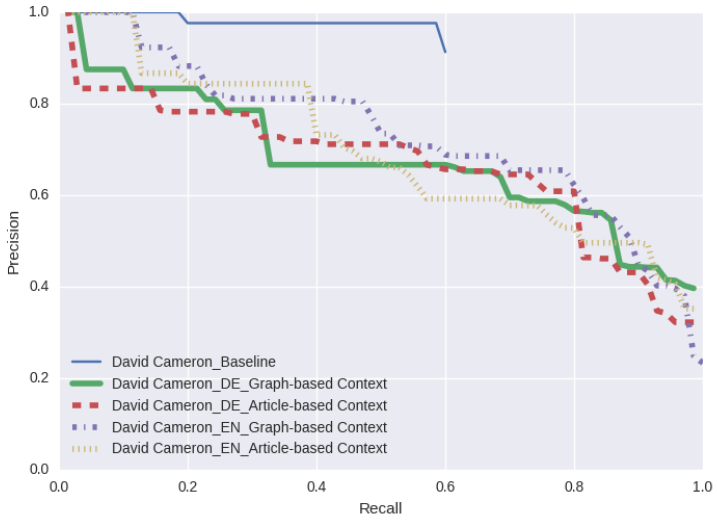


Fig. 3: Precision-Recall curve of language-specific context for "David Cameron".

The effectiveness of the context-based information retrieval model can also be observed for the query "David Cameron", presented in the Fig. 3. As shown in Fig. 3, the proposed model can achieve a much higher recall than the baseline for this query as well, while maintaining high precision. As expected, the English(EN) *Graph-based* context, which is local for this query, performs overall better than the other contexts.

We did not observe significant differences among the rest of the contexts for the query "David Cameron". One of the reasons can be the numbers of aspects in the corresponding entity contexts. The English *Graph-based* context of the entity "Angela Merkel" contains 7,317 non-zero weighted aspects, the German one contains 6,614. Both of them are much larger than her English and German *Article-based* contexts, which contain 562 and 1,069 non-zero weighted aspects, respectively. Resulting from that, the German and English *Graph-based* contexts for "Angela Merkel" are much more powerful than its German and English *Article-based* contexts. For "David Cameron", the most populated context is the English *Graph-based* context, which contains 10,365 non-zero weighted aspects, whereas other contexts have much smaller and comparable sizes. The German

*Graph-based* context of the entity "David Cameron" only has 1,627 non-zero weighted aspects; the numbers for his English and German *Article-based* contexts are 1,143 and 291. Although all of these contexts can still help to greatly improve the recall while maintaining relatively high precision, their effectiveness is somewhat limited because of their sizes.

### 5.3 Analysis of Language-specific Results

Table 4 presents the Top-8 results returned by the news retrieval application applying German and English *Graph-based* contexts of the query "Angela Merkel". As we can observe, when using the German context, German local news such as `http://www.thelocal.de/20141001/german-cabinet-agrees-cap-on-rent-rises-cities` and `http://www.thelocal.de/page/view/german-astronaut-calls-for-peace-and-tolerance` are included in the top-ranked results, which is not the case when using the English context. Nevertheless, as the key aspects of the entity are shared across both contexts, the results at the top of the both rankings are similar.

To better understand the impact of the language-specific contexts on the retrieved results, we define a measure: *Result Specificity*. *Result Specificity* ($RS$) is the percentage of unique documents in top-$K$ results retrieved using two contexts:

$$RS(R(e, C(e, l_i)), R(e, C(e, l_j))) = 1 - \frac{|R(e, C(e, l_i)) \cap R(e, C(e, l_j))|}{2 \times K}, \quad (4)$$

where $R(e, C(e, l_i))$ is the set of the results retrieved for the entity $e$ using the language context $C(e, l_i)$. The higher the *Result Specificity*, the less overlapping in the retrieved results using language-specific contexts, and the more language-specific the retrieved documents will be.

Fig. 4 illustrates the trends of the *Result Specificity* with an increasing number of returned results, when using the German and English *Graph-based* contexts for the entity "Angela Merkel". Whereas the most relevant results are very similar for both German and English *Graph-based* contexts, the *Result Specificity* of this pair reaches its maximum of 0.7 when $K$=15. This means that these language-specific contexts can retrieve distinct and relevant news articles at the higher ranks. Then, with an increasing $K$, both relevance and distinctiveness of the retrieved results drop, but the *Result Specificity* still stays above 0.5. On one hand, this is because many aspects in these two contexts overlap (as shown in Table 2, the EN-DE *Graph-based* context similarity value of "Angela Merkel" is 0.64). On the other hand, the most relevant news articles have been included in the retrieved results already by lower $K$ values. With an increasing $K$, divergent articles with lower relevance are further retrieved.

Similar trends can be observed in Fig. 5 for the query "David Cameron".

Table 4: Top-8 results for the query "Angela Merkel" retrieved using German(DE) and English(EN) *Graph-based* contexts.

| Rank | URL & Aspects overview(DE) | URL & Aspects overview(EN) |
|---|---|---|
| 1 | http://www.spiegel.de/international/germany/angela-merkel-changes-her-stance-on-refugee-limits-a-1063773.html (minister, idea, germany, merkel, chancellor) | http://www.spiegel.de/international/germany/angela-merkel-changes-her-stance-on-refugee-limits-a-1063773.html (minister, idea, germany, merkel, chancellor) |
| 2 | http://www.thelocal.de/20151130/we-owe-future-generations-a-climate-deal-merkel (prosperity, time, percent, paris, merkel) | http://www.thelocal.de/20151202/german-forces-will-back-france-in-syria-fight (bundeswehr, france, germany, thursday, syria) |
| 3 | http://www.thelocal.de/20151202/german-forces-will-back-france-in-syria-fight (bundeswehr, france, germany, thursday, syria) | http://www.thelocal.de/20151130/we-owe-future-generations-a-climate-deal-merkel (prosperity, time, percent, paris, merkel) |
| 4 | http://www.thelocal.de/20151202/no-better-life-for-afghans-in-germany-merkel (merkel, migration, dec, security, afghanistan) | http://www.thelocal.de/20151030/the-sailors-who-brought-down-the-german-empire (revolt, attack, government, battle, wilhelmshaven) |
| 5 | http://www.thelocal.de/page/view/hamburg-bids-farewell-to-its-most-famous-son (merkel, chancellor, schmidt, flag, terror) | http://www.spiegel.de/international/germany/editorial-on-anti-refugee-sentiment-in-germany-a-1062442.html (hitler, culture, germany, time, country) |
| 6 | http://www.spiegel.de/international/germany/editorial-on-anti-refugee-sentiment-in-germany-a-1062442.html (hitler, culture, germany, time, country) | http://www.thelocal.de/20151202/no-better-life-for-afghans-in-germany-merkel (merkel, migration, dec, security, afghanistan) |
| 7 | http://www.thelocal.de/20141001/german-cabinet-agrees-cap-on-rent-rises-cities (percent, average, law, oct, property) | http://www.thelocal.de/page/view/hamburg-bids-farewell-to-its-most-famous-son (merkel, chancellor, schmidt, flag, terror) |
| 8 | http://www.thelocal.de/page/view/german-astronaut-calls-for-peace-and-tolerance (publicity, vogel, space, space station, photo) | http://www.thelocal.de/20151202/less-than-half-of-german-jets-ready-for-action (report, syria, germany, wednesday, dec) |

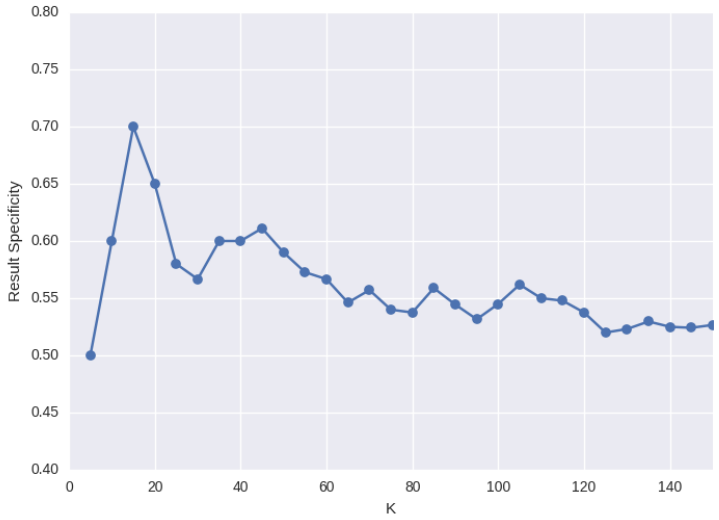


Fig. 4: *Result Specificity* of the top-$K$ retrieved results of German and English *Graph-based* contexts for the query "Angela Merkel" .

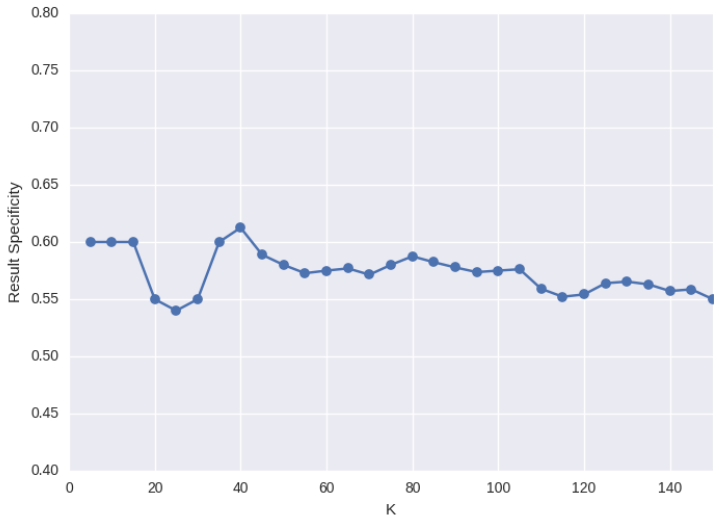


Fig. 5: *Result Specificity* of the top-$K$ retrieved results of German and English *Graph-based* contexts for the query "David Cameron".

## 6 Related Work

Due to its coverage and diversity, Wikipedia has been acting as an outer knowledge source to build semantic representations of entities/documents in various areas. Examples include information retrieval [4, 14, 18], named entity disambiguation [1, 2, 7, 8, 11, 12], text classification [25] and entity ranking [10].

To extract the content of an entity context, many researches directly used the Wikipedia article describing the entity [1, 2, 8, 9, 14, 25–27]; some works extended the article with all the other Wikipedia articles linked to the Wikipedia article describing the entity [6, 7, 12]; while some only considered the first paragraph of the Wikipedia article describing the entity [2]. Different from these approaches, our *Graph-based* approach not only employs *in-links* and *language-links* to broaden the article set that is likely to mention the entity, but also performs a finer-grained process: extracting the sentences that mention the entity, such that all the sentences in our context are closely related to the target entity.

As to the context-based representation vector of the entity, [1, 11] defined it as the tf-idf/word count/binary occurrence values of all the vocabulary words in the context content; [2,19] defined it as the word count/binary occurrence values of other entities in the context content; [5, 6, 9, 14, 25] defined it as the tf-idf similarity values between the target entity's context content and other entities' context contents from Wikipedia; [27] defined it as the visiting probability from the target entity to other entities from Wikipedia; [7, 26] used a measurement based on the common entities linked to the target entity and other entities from Wikipedia. Different from all former researches, we employ *aspect weights* that have a different interpretation of the frequency and selectivity than the typical tf-idf values and take co-occurrence and language specificity of the aspects into account.

Some researches [1, 2, 9, 12, 14, 25] also employed *category-links* to the Wikipedia article describing the entity. Since the category structure of Wikipedia is language-specific, it is hard to gain insights about cross-lingual context similarity for our case.

With the development of multilingual Wikipedia, researchers have been employing it in many multilingual applications [3, 16, 17, 20, 23, 24]. Similar to the English-only contexts, each dimension in a multilingual context representation vector represented the relatedness of the target entity with a set of entities/words in the corresponding language. However, none of these researches paid attention to the language-specific bias of multilingual Wikipedia, which has been proposed and verified in [28–30]. As different language editions of Wikipedia express different aspects related to the entity, in our research, we take a step further to analyse the differences in the language-specific entity contexts, and realised language-specific information retrieval through language-specific contexts.

As for incorporating the Wikipedia knowledge in information retrieval applications, [4, 15, 18] applied concept-based approaches that mapped both the documents and queries to the Wikipedia concept space; [14, 23] focused only on query extension; [20, 24] focused only on mapping documents to Wikipedia

concept space. To retrieve documents that did not explicitly mention the query entity by name, but were still relevant to the query entity, we chose to map both the query and the documents to the aspect space. As for the evaluation metrics of these information retrieval models, all these researches used the occurrence of the query entity as a prerequisite of one document to be relevant. Our research, on the other hand, excluded this condition. A document would be annotated as "Relevant" as long as it can satisfy any one of the three criteria in Section 5.1. When facing a dataset without enough documents mention the query entity explicitly, our context-based information retrieval model would still be able to return most relevant documents, thus achieved higher recall than former researches under this setting.

## 7 Conclusions and Outlook

In this article, we proposed context creation approaches for named entities, and used language-specific contexts to support entity-centric information retrieval. We compared different ways of context creation including the *Article-based* and the *Graph-based* approaches. A Wikipedia article describing the entity in a certain language can be seen as the most straightforward source for the language-specific entity context. Nevertheless, such context can be incomplete, lacking important entity aspects. Therefore, in this article we proposed an alternative approach to collect data for the context creation, i.e. the *Graph-based* approach. Our evaluation results showed significant differences between the contexts obtained using different context creation approaches. We suggested that the *Graph-based* approach was a promising way to obtain a comprehensive, language-specific overview of the entity independent of the Wikipedia article describing the entity. Furthermore, we proposed a context-based information retrieval model that applied such language-specific entity contexts to improve the recall of entity-centric information retrieval applications, while keeping high precision. Our case study illustrated that this model can retrieve documents that contain entity-related information, such as relevant events in the current news articles, even if the entity was not mentioned explicitly. And by selecting contexts of different language editions, our context-based information retrieval model made language-specific results possible.

Even though in this article we used limited number of named entities and languages as examples, our proposed approach and model can be easily extended to all other languages and named entities. In the future work, we plan to improve the context-based information retrieval model, so that it can process multilingual documents; we also plan to apply the model on other domains, to evaluate its effectiveness at a larger scale.

## Acknowledgments

This work was partially funded by the COST Action IC1302 (KEYSTONE), the ERC under ALEXANDRIA (ERC 339233) and H2020-MSCA-ITN-2014 WDAqua (64279).

## References


1. R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16, 2006.
2. S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716.
3. J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 121–124, New York, NY, USA, 2013. ACM.
4. O. Egozi, S. Markovitch, and E. Gabrilovich. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, 29(2):8, 2011.
5. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611.
6. E. Gabrilovich and S. Markovitch. Wikipedia-based semantic interpretation for natural language processing. pages 443–498.
7. X. Han, L. Sun, and J. Zhao. Collective entity linking in web text: a graph-based method. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 765–774. ACM.
8. X. Han and J. Zhao. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 215–224. ACM.
9. J. Hu, L. Fang, Y. Cao, H.-J. Zeng, H. Li, Q. Yang, and Z. Chen. Enhancing text clustering by leveraging wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 179–186. ACM.
10. R. Kaptein and J. Kamps. Exploiting the category structure of wikipedia for entity ranking. *Artificial Intelligence*, 194:111–129, 2013.
11. 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*, pages 1037–1045. ACM.
12. S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466. ACM.
13. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*, pages 1–8, 2011.
14. D. N. Milne, I. H. Witten, and D. M. Nichols. A knowledge-based search engine powered by wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 445–454. ACM.
15. C. Müller and I. Gurevych. Using wikipedia and wiktionary in domain-specific information retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 219–226. Springer, 2009.
16. V. Nastase and M. Strube. Transforming wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85, 2013.

17. J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175, 2013.
18. A. Otegi, X. Arregi, O. Ansa, and E. Agirre. Using knowledge-based relatedness for information retrieval. *Knowledge and Information Systems*, pages 1–30, 2014.
19. D. Ploch. Exploring entity relations for named entity disambiguation. In *Proceedings of the ACL 2011 Student Session*, pages 18–23. Association for Computational Linguistics.
20. M. Potthast, B. Stein, and M. Anderka. A wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval*, pages 522–530. Springer, 2008.
21. S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, Apr. 2009.
22. R. Rogers. *Digital Methods*, chapter Wikipedia as Cultural Reference. The MIT Press, 2013.
23. P. Schönhofen, A. Benczúr, I. Bíró, and K. Csalogány. Cross-language retrieval with wikipedia. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 72–79. Springer, 2008.
24. P. Sorg and P. Cimiano. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74:26–45, 2012.
25. P. Wang, J. Hu, H.-J. Zeng, and Z. Chen. Using wikipedia knowledge to improve text classification. 19(3):265–281.
26. I. Witten and D. Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30.
27. M. Yazdani and A. Popescu-Belis. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3185–3189. AAAI Press, 2013.
28. Y. Zhou, A. I. Cristea, and Z. Roberts. Is wikipedia really neutral? A sentiment perspective study of war-related wikipedia articles since 1945. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015*, 2015.
29. Y. Zhou, E. Demidova, and A. I. Cristea. Analysing entity context in multilingual wikipedia to support entity-centric retrieval applications. In *Proceedings of the 1st International KEYSTONE Conference (IKC 2015)*, pages 9–16, 2015.
30. Y. Zhou, E. Demidova, and A. I. Cristea. Who likes me more? analysing entity-centric language-specific bias in multilingual wikipedia. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '16, 2016.