Analysing Entity Context in Multilingual Wikipedia to Support Entity-Centric Retrieval Applications

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] L3S Research Center and Leibniz Universität Hannover, Germany
demidova@L3S.de

Abstract. Representation of influential entities, such as famous people 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. A systematic analysis of language-specific entity contexts can provide a better overview of the existing aspects and support entity-centric retrieval applications over multilingual Web data. An important source of cross-lingual information about influential entities is Wikipedia — an online community-created encyclopaedia — containing more than 280 language editions. In this paper we focus on the extraction and analysis of the language-specific entity contexts from different Wikipedia language editions over multilingual data. We discuss alternative ways such contexts can be built, including graph-based and article-based contexts. Furthermore, we analyse the similarities and the differences in these contexts in a case study including 80 entities and five Wikipedia language editions.

1 Introduction

Entities with world-wide influence, such as famous people and multinational corporations, can be represented differently in the news, in Web pages and 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 communities. In order to enable a better representation of the language-specific entity aspects in the information retrieval systems, methods to systematically identify language-specific entity contexts – i.e. the aspects in the entity descriptions typical to a specific language – need to be developed.

For example, in the English news, the entity "Angela Merkel", the Chancellor of Germany, is often associated with US and UK politicians such as Barack Obama, or David Cameron. Also, recent discussions of the European importance, such as Greek financial situation are included. On the contrary, although the German pages 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. For another example, in case of the multinational companies like GlaxoSmithKline (a British healthcare company), the aspects related to the local activities are prevalent in the reporting 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. The knowledge of such language-specific aspects can support entity-centric retrieval applications. These applications can include e.g.: (i) creation of a comprehensive overview of the language-specific entity aspects in the document collection; and (ii) targeted retrieval of entity-centric information in a specific language context.

In order to obtain a comprehensive overview over the language-specific entity aspects and their representations in different languages, language-specific background knowledge about this entity is required. In this paper we choose Wikipedia as a multilingual knowledge base to obtain such background. Wikipedia — a multilingual encyclopaedia available in more than 280 language editions — contains language-specific representation 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. [19]). As entity representations in different Wikipedia language editions can evolve independently, they often include overlapping as well as language-specific entity aspects.

In this paper we focus on the problem of creating language-specific entity contexts to support multilingual entity retrieval applications. We use Wikipedia as a knowledge base to build such language-specific entity contexts. We discuss different ways of building these contexts using Wikipedia, including *article-based* and *graph-based* approaches and propose a measure to compute the context similarity. Furthermore, we use this similarity measure to analyse the similarities and the differences of the language-specific entity contexts in a case study using 80 entities of four different entity types and the representations of these entities in five European languages. Our experiments show that the proposed *graph-based* entity context can effectively provide a comprehensive overview over the language-specific entity aspects.

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 set of aspects $\{a_1, \ldots, a_n\}$ of e in L_i, weighted to reflect the relevance of the aspects in the context: $C(e, L_i) = (w_1 * a_1, \ldots, w_n * a_n)$.*

In this paper, we consider entity aspects being all related entities, concepts and terms. We obtain such aspects automatically by extraction of *noun phrases* that co-occur with the entity in a given language. In addition, we can also consider the names of the linked articles as an additional source of the entity aspects. The weights of the aspects are based on two factors: (1) the *language-specific aspect co-occurrence frequency*, i.e. the frequency of the co-occurrence of the aspect and the entity in a language, and (2) the *cross-lingual aspect co-occurrence frequency* - the number of languages in which the entity context contains 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. Overall, given a multilingual data collection containing the languages $L = \{L_1, \ldots, L_N\}$, 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:

$$w(a_k, e, L_i) = af(a_k, e, L_i) \cdot log \frac{N}{af(a_k, e, L)}, \tag{1}$$

where $af(a_k, e, L_i)$ is the *language-specific aspect co-occurrence frequency* of the aspect a_k in the entity context of e in the language L_i, N is the number of languages in the multilingual collection, and $af(a_k, e, L)$ is the *cross-lingual aspect co-occurrence frequency*.

In case the aspects are represented through the manually-defined links, their relevance can be high, but the frequency low compared to the noun phrases; Therefore, we assign them with an average *language-specific aspect co-occurrence frequency* computed for the noun phrases.

2.2 Context Similarity Measure

In order to compute the similarity between language-specific entity contexts, we use the vector space model. Each axis in the vector space represents the 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 the context of the entity in the language L_i. Then the context similarity of e's contexts in the languages L_i and L_j is computed as the cosine similarity of the context vectors:

$$Sim(C(e, L_i), C(e, L_j)) = \frac{C(e, L_i) \cdot C(e, L_j)}{|C(e, L_i)| \times |C(e, L_j)|}. \tag{2}$$

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 paper, 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 Extraction

Wikipedia articles representing an entity in different language editions (i.e. the articles that contain the entity name as a title) can be a useful source for the creation of the language-specific context vectors. Thus, we first propose the *article-based* context extraction approach, which simply uses the articles representing the entity in different language editions of Wikipedia. We use all sentences from an article representing 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 aspect 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, in the Wikipedia article about "Economic Council Germany", it 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 main English Wikipedia article entitled "Angela Merkel". As this example illustrates, just employing the main Wikipedia article about the entity can not entirely satisfy the need to obtain a comprehensive overview over the language-specific aspects.

2.4 Graph-based Context Extraction

To alleviate the shortcomings of the *article-based* method 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 extraction method. The idea behind this method 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 main Wikipedia article describing the entity and the *language-links* of these articles 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 build language-specific contexts.

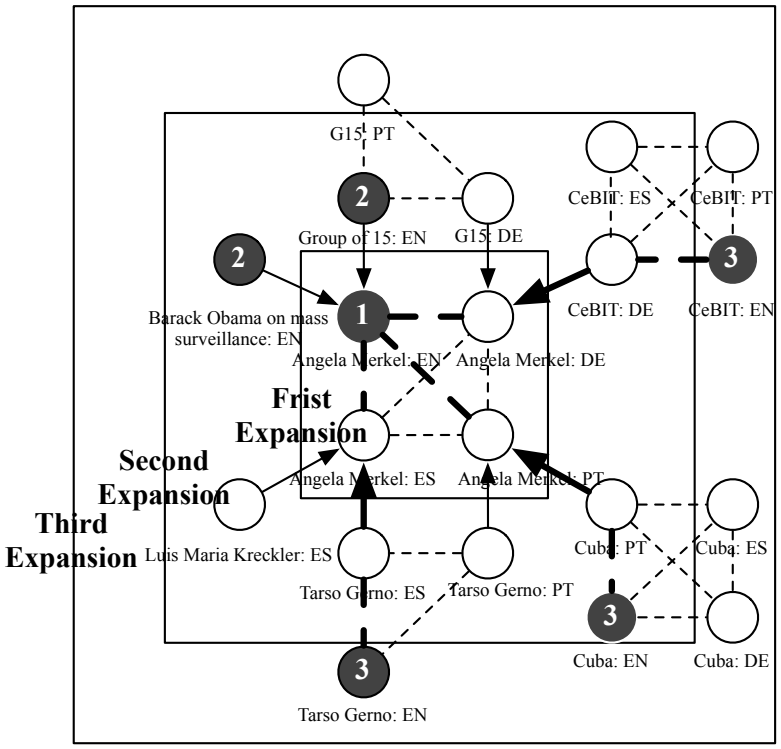

Fig. 1. An Example of the Graph-Based Context Extraction from the English Wikipedia for the entity "Angela Merkel".

To illustrate our approach, we use the extraction of the context in the English edition of Wikipedia for the entity *Angela Merkel* as an example. For the Wikipedia article in English about Angela Merkel, there are several *in-links* from other articles in English that mention the entity to it. Besides that, there are also *language-links* from the articles representing Angela Merkel in other Wikipedia languages 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 about the entity (the central node) to all language editions of this Wikipedia article; Then, we further expand the node set with all the articles having *in-link*s to the nodes in the node set; Finally, we expand the node set with all the articles having *language links* to the existing nodes, 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 one of the language editions has an *in-link* to one of the other language editions' article about Angela Merkel; 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 Entity Context Analysis

The goal of the entity context analysis is to compare the *graph-based* and the *article-based* context creation methods. To this extent we analyse the similarities and the differences of the contexts obtained using these methods in a case study.

3.1 Dataset Description

To facilitate our analysis, we selected a total number of 80 entities with world-wide 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 chosen 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 main Wikipedia article describing the entity using the *article-based* approach is around 50 in our dataset. Using our *graph-based* context extraction method that utilises 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 1000 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* method for each entity processed. The total number of sentences collected by the *graph-based* approach is 419,289 sentences for the whole dataset under consideration.

3.2 Context Similarity Analysis

The cross-lingual context similarity resulting from the *article-based* and the *graph-based* context creation methods, are presented in Table 1 and Table 2, respectively. To enable cross-lingual context similarity computation, we translated all context vectors 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 80 entities in the entire dataset.

Table 1. Cross-lingual context similarity using the *article-based* context creation method. The table presents the similarity for four selected entities of different types, as well as the average similarity and the standard deviation for the whole dataset of 80 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 80	**0.50**	**0.47**	**0.46**	**0.43**	0.38	0.36	0.39
Stdev of 80	0.16	0.20	0.23	0.22	0.18	0.19	0.22

From Table 1, we can observe that using the *article-based* context creation method, the average similarity of the language pairs including English is always higher than that 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, it is possible that the English edition builds a reference for the creation of the articles in other language editions in many cases. This can be further explained by the large size of the English Wikipedia that has the largest number of users, articles, and edits compared to

Table 2. Cross-lingual context similarity using the *graph-based* context creation method. The table presents the similarity for four selected entities of different types, as well as the average similarity and the standard deviation for the whole dataset. 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 80	0.53	**0.60**	0.56	0.52	0.53	0.48	**0.61**
Stdev of 80	0.25	0.22	0.21	0.24	0.24	0.25	0.20

other language 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* method reach at most 0.5, even for the language pairs with 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* method, the *graph-based* method 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 context pair is are the Spanish and the 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* method. We can also observe that the average similarity values significantly increase compared to the *article-based* method and can exceed 0.6 in our dataset.

From a single entity perspective, some entities may achieve higher than the average similarities for a language pair, when more common aspects are included in the context on the 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 similarities for some language pairs, especially when distinct aspects are included into the corresponding languages 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 context vectors of the entity "Angela Merkel" constructed using the *graph-based* method, 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",

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

"Kingdom" and "Dilma Rousseff" — are more relevant to the Angela Merkel's international affairs in corresponding 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 the Angela Merkel's domestic activities.

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

Overall, our observations confirm that the *graph-based* context provides a better overview of the different entity aspects than the *article-based* method. The *graph-based* method can determine the similarities and the differences of the language contexts, independent of the coverage and completeness of any dedicated Wikipedia article. The results of the t-test confirm the statistical significance of the context similarity differences between the *article-based* and the *graph-based* methods for all language pairs except the EN-DE. This exception can be explained by a relatively high coverage of the German Wikipedia pages 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 method is capable of capturing these differences from different language editions by creation of comprehensive language-specific entity contexts.

4 Related Work

Due to its coverage and diversity, Wikipedia has been acting as an outer knowledge source to build semantic representations of entities in various areas. Examples include information retrieval [4, 10, 14, 17, 22], named entity disambiguation [1, 2, 8, 11], text classification [6, 20] and text clustering [9].

To extract an entity context, many researches directly used the Wikipedia article about the entity [1, 2, 4, 7–10, 14, 17, 20–22]; some works extended the article with all the other Wikipedia articles linked to the Wikipedia article of the entity [6]; Others extended the article with the anchor texts and words in a fixed-sized window around such anchor text [12]; Further works only considered the first paragraph of the article [2] . Different from these approaches, our *graph-based* method do not only employ in-links and language links to broaden the article set that is likely to mention the entity, but also performs a finer-grained process: Annotate 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 of the entity, [1, 11] defined it as the tf-idf/word count/binary occurrence values of all the vocabulary words in the target entity content; [2] defined it as the binary occurrences of entities; [18] defined it as as number of occurrences of entities; [4–7, 9, 10, 14, 17, 20, 22] defined it as the tf-idf similarities between the target entity and some entities in Wikipedia. In this paper, 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.

Following [21], some researches [9, 12, 14] also employed category links from the entity's Wikipedia article. Since the category structure is language-specific, it does not provide enough insides to compute similarity scores directly.

With the development of multilingual Wikipedia, researchers have been employing it in many multilingual applications [3,15,16]. Similar to the English-only contexts, each dimension in a multilingual context represented the relatedness of the target entity with the set of entities in the corresponding language. As different language editions of Wikipedia express different aspects of the entity, in our research, we take a step further to analyse the differences in the language-specific entity contexts.

5 Conclusions and Outlook

In this paper we presented the notion of the language-specific entity context and proposed a measure to compute the context similarity across different languages. Furthermore, we compared different ways of context creation including the article-based and the graph-based methods. A Wikipedia article describing the entity in a certain language can be seen as the most straightforward source for the language-specific entity context creation. Nevertheless, such context can be incomplete, lacking important entity aspects. Therefore, in this paper we proposed an alternative method to collect data for the context creation, i.e. the graph-based method. This method enables us to obtain a more comprehensive overview of the entity representation in a Wikipedia language edition as a whole. Our evaluation results show significant differences between the contexts obtained using different context creation methods and suggest that the graph-based method is a promising approach to obtain a comprehensive overview of the language-specific entity representation independent of the Wikipedia article describing the entity. In our future work, we plan to focus on the development of the context-aware retrieval applications from multilingual data sources.

Acknowledgments

This work was partially funded by the COST Action IC1302 (KEYSTONE) and the European Research Council under ALEXANDRIA (ERC 339233).

References

1. R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16.
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. V. Nastase and M. Strube. Transforming wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85, 2013.
16. 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.
17. A. Otegi, X. Arregi, O. Ansa, and E. Agirre. Using knowledge-based relatedness for information retrieval. *Knowledge and Information Systems*, pages 1–30, 2014.
18. D. Ploch. Exploring entity relations for named entity disambiguation. In *Proceedings of the ACL 2011 Student Session*, pages 18–23. Association for Computational Linguistics.
19. R. Rogers. *Digital Methods*, chapter Wikipedia as Cultural Reference. The MIT Press, 2013.
20. P. Wang, J. Hu, H.-J. Zeng, and Z. Chen. Using wikipedia knowledge to improve text classification. 19(3):265–281.
21. 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.
22. 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.