Extracting Descriptions of Location Relations from Implicit Textual Networks

Andreas Spitz
Institute of Computer Science
Heidelberg University
Im Neuenheimer Feld 205
69120, Heidelberg, Germany
spitz@informatik.uni-heidelberg.de

Gloria Feher
Institute of Computer Science
Heidelberg University
Im Neuenheimer Feld 205
69120, Heidelberg, Germany
feher@stud.uni-heidelberg.de

Michael Gertz
Institute of Computer Science
Heidelberg University
Im Neuenheimer Feld 205
69120, Heidelberg, Germany
gertz@informatik.uni-heidelberg.de

ABSTRACT

For the retrieval of concise entity relation information from large collections or streams of documents, existing approaches can be grouped into the categories of (multi-document) summarization and knowledge extraction. The former tend to fall short for this task due to the involved amount of information that cannot be easily condensed, while knowledge extraction approaches are often pattern-based and too discriminative for exploratory purposes. For location relations in particular, this translates to a set of very short relationship descriptors that predominantly encode hierarchical or containment relations such as `located in` or `capital of`. As a result, available knowledge bases that are typically populated through knowledge extraction are limited to these discrete and typed relations. In contrast, the representation of document collections as implicit networks of entities, terms, and sentences has emerged as a way to encode a much wider range of entity relations and occurrences, which can be leveraged for filtering the relevant information and enabling subsequent interactive explorations.

In this paper, we discuss the extraction of descriptive sentences for sets of entities from such implicit networks to support an interactive exploration, and apply them to the extraction of complex location relations that are not hierarchical or containment-based. We introduce and compare efficient ranking methods for sentence extraction that address this entity-centric search task by leveraging entity and term relations in implicit network representations of large document collections. Based on Wikipedia articles and Wikidata as a knowledge base, we demonstrate the extraction of novel location relations that are not contained in the knowledge base.

CCS CONCEPTS

•**Information systems** →*Document representation;* •**Computing methodologies** →*Information extraction;*

KEYWORDS

Implicit entity network; toponym; sentence extraction; ranking

ACM Reference format:
Andreas Spitz, Gloria Feher, and Michael Gertz. 2017. Extracting Descriptions of Location Relations from Implicit Textual Networks. In *Proceedings of 11th Workshop on Geographic Information Retrieval, Heidelberg, Germany, November 30-December 1, 2017 (GIR'17),* 9 pages.
DOI: 10.1145/3155902.3155909

1 INTRODUCTION

The steadily growing amount of available textual information with geographic content has long passed the threshold for unassisted human analysis in many areas. Frequently, such an analysis is focused on certain aspects of the content that can be interpreted as a class of entities, such as person or location mentions, as well as their relations. In those cases, the exploration of new or newly digitized document collections inherently benefits from automated and interactive exploration tools that support an entity-centric focus. It thus comes as no surprise that automated summarization and knowledge extraction approaches are among the most popular topics in information extraction and retrieval. Given a set of documents, extractive summarizers are aimed at composing a comprehensive, readable and short overview of the contained key points [19]. Entity-centric summarization techniques in particular allow for the extraction of brief descriptions of entities or sequences of entities from a text [7]. However, in settings where the document collection is large, such approaches are problematic due to the sheer amount of information, which calls for the preparatory selection of relevant information with a focus on some aspect of the collection. Useful tools in this regard are searching and indexing approaches, external knowledge bases, or question answering systems [12]. However, such systems are not designed for the retrieval of succinct descriptions of aspects, entities, or relations within groups of entities in the collection. Due to their predominantly pattern-based design, fact extraction approaches stand to miss information that is not well structured but potentially suitable to human understanding in an interactive exploration of the documents.

For the extraction of geographic mentions, location attributes, and location relations in particular, this results in an extremely limited scope of attributes and relations that can be retrieved from the texts, due to the fact that more complex relations beyond co-location and containment are often not expressed in simple patterns (or cannot even be expressed in such patterns). Consider, for example, trade relations between countries or cities. While it would be possible to have a sentence that explicitly states such a relation (e.g., *China is the largest trading partner of the European Union*), it is much more likely that this relation is never explicitly stated and can only be inferred from the frequent mention of individual trades. For these types of relations, a pattern-based extraction is ill-suited. Significant co-occurrences of location mentions in a specific context, on the other hand, can be helpful in determining pairs of related locations and subsequently identifying sentences that describe their relation. As a result, such relations are largely missing from established knowledge bases like YAGO [16] and DBpedia [13], which are populated through pattern-based knowledge extraction, and even from the community edited Wikidata [26], which is increasingly populated through automated approaches as well [24]. Thus, these knowledge bases and subsequent applications stand to benefit from an augmentation through linking descriptive sentences to the contained entities or sets of entities.

In this paper, we take an important first step to satisfy this information need and consider the task of *descriptive sentence extraction* for locations and location relations in such an exploratory setting. To support the diverse underlying exploratory queries, we require the representation of large-scale document collections in an efficient data and indexing structure that enables ad-hoc browsing and information extraction around focus entities. Furthermore, the representation should contain relevant identifying passages and documents for all entities that can be extracted and used in subsequent, more specialized analyses. For this purpose, we build on the concept of implicit entity networks, which have recently emerged as useful and language-agnostic tools that provide provenance information with regard to document collections for named entities in general [23] and location mentions in particular [9]. Based on the representation of entities in a document as nodes of a graph and co-occurrence relations as edges, such networks allow a user to search for entity-specific information within a corpus, without introducing external knowledge (which may be rather different from the contained, internal knowledge), and without the need to summarize the entire corpus (which would contain unnecessary information and is technically infeasible). In the following, we investigate how such networks can be used to extract descriptive and explanatory sentences for location entities or sets of such entities from the underlying collection, and apply these techniques to the extraction of complex location relations that go beyond the hierarchical and containment relations that constitute the majority of geographic (spatial) relations in knowledge bases.

Contributions. In summary, our contributions in this paper are as follows. **(i)** We formalize the problem of entity-centric explanatory sentence extraction for the interactive search of large-scale document collections. **(ii)** We describe an efficient model for the extraction of such sentences from an (implicit) entity network representation of such a collection that can serve as a basis for composition techniques in (multi) document summarization. **(iii)** We use an entity network of the English Wikipedia and ground truth from Wikipedia glossary pages for our subsequent evaluation, which we make available as resources[1]. **(iv)** We apply the approach to the extraction of complex location relations from the English Wikipedia to identify relations that are not contained in the underlying Wikidata knowledge base.

Structure. The remainder of this paper is structured as follows. In Section 2, we give an overview of related work. We discuss the representation of a document collection as an implicit network of entities and introduce the sentence extraction approaches in Section 3. In Section 4, we evaluate the different sentence ranking functions on a set of glossary entities, before we perform the sentence extraction for location relations in Section 5. Finally, we discuss implications and future work in Section 6.

2 RELATED WORK

Since the existing work on descriptive sentence extraction in general and the description of location relations in particular is fairly limited, we give a brief overview of related areas and the works that are conceptually the most similar.

Document Summarization. The summarization of entire documents (or sets of documents) has been well researched in the past and several recent surveys exist on the matter [15, 19]. In particular graph-based approaches to coherence and composition are highly successful, which were pioneered by contributions such as as LexRank [8] and TextRank [18] that use graph centrality to identify relevant components of a summary. These methods have been continuously improved and more recent additions include novel techniques such as topic signatures [1], vector embeddings [17], or word associations relative to a background corpus [10]. In contrast to document summarization, we focus on the retrieval and exploration of location entity information from large document collections on a scale for which traditional multi-document summarization approaches are ill-suited.

Knowledge Extraction and Question Answering (QA). A traditional source of entity-centric information are knowledge bases such as DBpedia [13] or YAGO [16], which are constructed in part from structured knowledge and in part from unstructured text. This extraction of knowledge from unstructured text is closely related to information extraction on a web scale, such as Open Information Extraction [3]. Structured knowledge and the patterns that are used in its extraction are then useful tools in question answering (QA) systems (for an overview of such systems, see [12]). The extraction step in both knowledge base construction and question answering is largely pattern-based. Ambiguous information is often discarded, even though it might be sufficient to satisfy a user's information need. Here, we focus on a less restricted approach that uses sentence retrieval to avoid pattern-based and language-specific extraction.

Summarization and QA for Geographic IR. Text summarization and question answering with a focus on geographic concepts have first been extensively addressed in GeoCLEF[2] as part of the Cross Language Evaluation Forum (CLEF). Among the main tasks, the focus has been put on NLP-based geographic search, i.e., on a hybrid approach between text summarization and QA, but neither on the extraction of relationships between locations (or geographic entities in general) nor on location summarization. Similarly, Chen et al. [6] also focus on geographic QA but do not address location summarization or relations between locations. In addition to geographic QA, the more recent NTCIR GeoTime track[3] included the temporal dimension for determining answers to NLP-based geographic search queries. Different aspects of summaries for geographic IR have been analyzed and studied by Perea-Ortega et

[1] https://dbs.ifi.uni-heidelberg.de/resources/data/

[2] http://www.uni-hildesheim.de/geoclef/

[3] http://metadata.berkeley.edu/NTCIR-GeoTime/

al. [20], who focus on sentences in a document containing geographic entities but do not specifically study location summaries or extracted location relations.

Similar Approaches. Probably the most popular works on enriching location information are based on a method proposed by Rattenbury and Naaman [21], in which image tags in Flickr are used to derive semantically rich geospatial image and location descriptions. Similarly, Tardi et al. [25] outline an approach to identify the characteristics of locations from tags associated with photographs. However, these frameworks neither employ text rich sources to further improve location descriptions, nor do they investigate descriptive relations between locations.

Some works in other domains focus on similar, entity-centric tasks and sentence extraction. Kim et al. consider the extraction and ranking of sentences based on their usefulness for understanding sentiments in a document for opinion summarization [11]. Biadsy et al. extract summaries targeted at the creation of person biographies by focusing on sentences that are biographical in nature [4]. Amitay and Paris summarize websites from external descriptions by looking at sentences that contain hyperlinks to these websites [2], which is conceptually similar to our focus on entities, but they utilize patterns that are specific to hyperlink anchors. The above approaches are focussed on a specific type of entity or notion (sentiments, persons, or links). In contrast, we consider a broader approach that is useful for entity-centric explanation in general.

For the extraction of support sentences, Blanco and Zaragoza aim to identify descriptive sentences from a document collection that describe the entities contained in a textual query [5]. In contrast to our approach, they require a text query as input and exclude sentences that do not contain an input entity, which is an issue that frequently arises when multiple entities are used as input. Based on an implicit network of entities, Spitz and Gertz introduce rudimentary sentence ranking for a network of locations, organizations, actors, and dates [23], that they later extend to include a sentence-centric exploration of large document collections [22], which we use as a baseline for our approach.

3 ENTITY GRAPH AND QUERY MODEL

We first give an intuitive definition of entity-centric sentence extraction before introducing the implicit entity network model for representing a document collection as a graph structure. Then, we formalize the extraction operations based on this underlying graph representation. In the following, we consider sets of documents $\mathcal{D}$, sentences $\mathcal{S}$ and words $\mathcal{W}$, for which we have a simple containment relation $\sqsubseteq$. That is, for a word w contained in a sentence s we write $w \sqsubseteq s$. Likewise, the sentence is contained in some document d and we write $s \sqsubseteq d$. We partition the set of words into entities $\mathcal{E}$ and terms $\mathcal{T}$. As entities $\mathcal{E} \subseteq \mathcal{W}$ we consider a subset of words (e.g., words that can be linked to a knowledge base or a gazetteer). The remaining words then constitute the terms $\mathcal{T} := \mathcal{W} \setminus \mathcal{E}$. Intuitively, entities are thus terms with specific characteristics, such as named entities. In Section 5 we focus locations as entities, but use the more general definition here to describe the model.

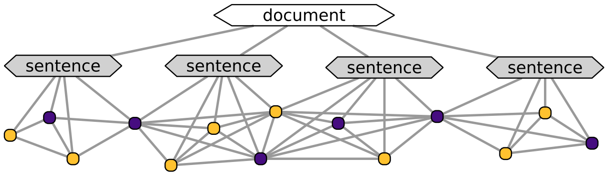

Figure 1: Schematic view of the entity-term graph for one document (graphs are then aggregated over all documents).

3.1 Descriptive Sentence Extraction

To introduce the sentence extraction tasks, we begin with single-entity sentence extraction and then extend it to the more general case of multi-entity sentence extraction.

Single-Entity Sentence Extraction. Given a collection of documents $\mathcal{D}$ containing sentences $\mathcal{S}$, a set of entities $\mathcal{E}$, and a query entity $e \in \mathcal{E}$, we let $\mathcal{S}_e := \{s \in \mathcal{S} \mid e \sqsubseteq s\}$ be the set of sentences in which entity e is mentioned. The aim is then to identify the sentence $s_e \in \mathcal{S}_e$ that generally best describes e. Extending this notion, we can then also consider the summarization of relationships between entities by focusing on sentences that jointly describe the occurrences of a set of entities.

Multi-Entity Sentence Extraction. Given a collection of documents $\mathcal{D}$, sentences $\mathcal{S}$, entities $\mathcal{E}$ and a subset of entities $\mathcal{U} \subseteq \mathcal{E}$, we let $\mathcal{S}_{\mathcal{U}} := \bigcup_{u \in \mathcal{U}} \{s \in \mathcal{S} \mid u \sqsubseteq s\}$ be the set of sentences in which at least one of the entities is mentioned. A descriptive sentence for the set of entities $\mathcal{U}$ then is the sentence $s_u \in \mathcal{S}_{\mathcal{U}}$ that best describes the joint occurrences of entities from $\mathcal{U}$ in $\mathcal{D}$. From this definition, it is clear that single-entity sentence extraction is a special case of multi-entity sentence extraction for $|\mathcal{U}| = 1$. In the following, we thus focus on the more general task.

3.2 Entity Graph Extraction

In order to approach a solution to the ad-hoc extraction of entity-centric sentences, we rely on an efficient representation of the document collection as an implicit entity graph $G = (V, E)$ that is generated from entity mentions in the text. We use a model that is similar to the LOAD graph for event exploration [23], but do not distinguish between different classes of entities.

Graph Nodes. As nodes V of the graph, we include the set of entities of interest $\mathcal{E}$ and the remaining terms $\mathcal{T}$. Furthermore, we also include sentences $\mathcal{S}$ and the documents $\mathcal{D}$ for provenance. Thus, we define the set of nodes as $V := \mathcal{D} \cup \mathcal{S} \cup \mathcal{E} \cup \mathcal{T}$.

Graph Edges. Edges E of the graph are based on the containment relation. Two nodes are connected if one is contained in the other (e.g., a term in a sentence) or if they co-occur in at least one instance (e.g., two terms occurring in a sentence). Formally, for $v \in \mathcal{E} \cup \mathcal{T}$, $s \in \mathcal{S}$ and $d \in \mathcal{D}$ we have edges of the containment type $(v, s) \in E :\Leftrightarrow v \sqsubseteq s$, and equivalently $(s, d) \in E :\Leftrightarrow s \sqsubseteq d$. For $v, w \in \mathcal{E} \cup \mathcal{T}$ we also have edges of the co-occurrence type $(v, w) \in E :\Leftrightarrow \exists s \in S : v \sqsubseteq s \wedge w \sqsubseteq s$. For a schematic overview of the resulting network, see Figure 1. In the following, let $N(v)$ denote the neighbourhood of a node, i.e., the set of nodes that are connected to v by an edge.

Edge Weights. While the multiplicity of edges of the containment type is largely irrelevant (with the exception of stop words, sentences rarely include the same word twice), edges of the co-occurrence type are likely to have a higher multiplicity since each sentence in which two entities or terms co-occur induces such an edge. To aggregate these edges over the entire document collection and obtain a simple graph, we assign this multiplicity as a weight m to the edge. Formally, for $x, y \in \mathcal{E} \cup \mathcal{T}$, we let

$$m(x, y) := |\{s \in S \mid x \sqsubseteq s \wedge y \sqsubseteq s\}|. \qquad (1)$$

The resulting weight is symmetric and encodes the frequency at which x and y co-occur. Based on this multiplicity, we can induce directed edge weights that include a normalization for the overall number of co-occurrences. For two entities or terms $x \in X$ and $y \in Y$ with $X, Y \in \{\mathcal{E}, \mathcal{T}\}$, we define a weight $\omega : (\mathcal{E} \cup \mathcal{T})^2 \to \mathbb{R}$ as

$$\omega(x \mid y) := m(x, y) \log \frac{|Y|}{|N(x) \cap Y|}. \qquad (2)$$

Since only the neighbourhood information of the two nodes is used, this weight can be computed efficiently. Due to the second factor, weights are directed such that $\omega(x|y) \neq \omega(y|x)$. The resulting weighted network can be seen as a type of embedding for entities and terms. However, this does not imply that neural word to vector embeddings would be equally useful if used instead of the network. As we show in the following, discrete term representations are required for the propagation of importance from terms to sentences, which vector representations cannot support.

Note that the specific construction of the graph that we utilize here is not a necessary condition for the following ranking approaches, which are compatible with any implicit entity graph of similar structure. The only requirement are edge weights that encode a measure of semantic relatedness between the entities.

3.3 Graph-based Sentence Extraction

Based on the implicit entity graph, we now introduce realizations of sentence extraction methods. We treat the task as a sentence ranking problem, in which we rank sentences according to their relevance for a set of input query entities $Q \subseteq \mathcal{E}$ and then select the top-ranked sentence. Formally, we use scoring functions $r : \mathcal{S} \to \mathbb{R}$ that allow a ranking of sentences in the collection by their descriptiveness for the input entities. The answer to a query then is a sentence $s \in S$ such that $r(s) \geq r(s') \,\forall s' \in S \backslash \{s\}$. In the following, we describe four different scoring methods in an iterative manner. Thus, with the exception of the last method M4, each subsequent method includes and builds upon all previous methods.

Entity Count (M1). As a baseline r_1, we include the method that was originally proposed for event description extraction in the LOAD model [23]. It counts the number of entities from the query set that occur in a sentence. For each sentence s, we obtain

$$r_1(s, Q) := |N(s) \cap Q|. \qquad (3)$$

For descriptive sentence extraction, this method can only serve as a simple baseline since it suffers from two flaws. First, since the method is designed for larger sets of query entities, it performs poorly in the extraction of descriptive sentences for single query entities, as it assigns equal rank to all sentences that contain the entity. Second, it does not consider the context of a sentence beyond the contained entities. Thus, ties between sentences with the same number of entities cannot be broken, which is especially of interest if no sentence exists that contains all query entities Q.

Term Influence (M2). Based on the above observations, we suggest an improved two-component ranking. The number of entities in the sentence is kept as the first component, while we derive the second component from the set of terms that are most relevant to the query entities. To this end, consider a ranking of terms in the neighbourhood of entity e by ω and let t_n be the n-th ranked term. Let $T_n(e) := \{t \in \mathcal{T} \mid \omega(t|e) \geq \omega(t_n|e)\}$, then this set $T_n(e)$ contains the n highest ranked terms in the graph with regard to e. We obtain the most relevant terms for a set of query entities Q as

$$T_n(Q) := \bigcup_{e \in Q} T_n(e) \qquad (4)$$

which we then use to represent the content and as possible placeholders for query entities in a sentence. We still rank by the number of query entities in the sentence first, but break ties by using the n most relevant terms for each entity. Formally, we let

$$r_2(s, Q, n) := |N(s) \cap Q| + \frac{|N(s) \cap T_n(Q)|}{|T_n(Q)| + 1}. \qquad (5)$$

Since the first component is an integer and the second component is strictly less than one, we obtain a ranking based on the two-component intuition discussed above. We find that identifying and using such relevant terms in addition to the query entities works well for sentence selection, but suffers from sentence length. While short and compact sentences are preferable descriptions in practice, both r_1 and r_2 assign a higher weight to sentences that contain more entities (and terms) and thus favor longer sentences. In the following, we consider two normalization schemes.

Normalization by Length (M3). One way of normalizing for the length of a sentence s is by directly using the length in characters $l(s)$. Thus, we introduce the normalized score r_3 as

$$r_3(s, Q, n) := \frac{1}{\log l(s)} \left[|N(s) \cap Q| + \frac{|N(s) \cap T_n(Q)|}{|T_n(Q)| + 1} \right] \qquad (6)$$

Since we found in our empirical evaluation that we would otherwise give preference to extremely short sentence fragments that contain little more than the entity itself, we use the logarithm of the length (an alternative would be the length of a sentence in words). While this scheme normalizes based on the length of a sentence, it does not account for the number of entities and terms in the sentence overall and does not distinguish between terms and entities.

Normalization by Count (M4). As a final method, we thus include a two-factor normalization that is based on the number of entities and terms per sentence. We define r_4 as

$$r_4(s, Q, n) := \frac{|N(s) \cap Q|}{|N(s) \cap \mathcal{E}|} + \frac{|N(s) \cap T_n(Q)|}{|T_n(Q)| \cdot (|N(s) \cap \mathcal{T}| + 1)} \qquad (7)$$

by normalizing the two components separately. The addition of 1 in the second denominator is due to the border case of sentences that contain no terms. We obtain a two-component ranking that effectively measures the fraction of relevant entities and relevant terms occurring in a sentence. The factor $|T_n(Q)|$ in the second term also ensures that the contribution of entities to the final score is much larger than the contribution of relevant terms.

In the following, we utilize and compare these four methods for the description of entity relations in general and relations between geographic locations in particular.

4 SENTENCE EXTRACTION EVALUATION

We discuss the construction of the large implicit network of entities against which we run the evaluation queries, as well as the extraction of the ground truth data, before proceeding to the evaluation.

4.1 Evaluation Data and Network Construction

The evaluation performance of entity-centric approaches is always influenced by the quality of entity annotations in the available data. Since we rely on both the recognition and disambiguation of named entities for a large document collection, manual annotations are prohibitively expensive. Thus, we use Wikipedia as an evaluation resource, which allows us to recognize entity mentions due to embedded links between pages and disambiguate them through connections to the underlying knowledge base Wikidata.

Entity Network Construction. To obtain a large-scale network of entities from text, we use the English Wikipedia (dump of December 1, 2016) as a document collection and restrict the content to unstructured text (we exclude lists, references, and info boxes). As an entity, we consider any surface string that covers an embedded link to another Wikipedia page. Thus, we use embedded links to identify entities and directly link them to Wikidata identifiers (that can be found on the target page). According to Wikipedia rules, entities are linked only once per page and we thus also use a string search of cover texts and Wikidata entity labels to tag subsequent mentions. We exclude all links that have no associated Wikidata identifier (i.e., links that lead to Wikipedia pages with no associated Wikidata entity). To generate the network and exclude word fragments, we split the documents into sentences that are then tokenized. We restrict the terms to a minimum length of 4 characters before stemming and removing stop words. The set of edges is generated as described in Section 3. The resulting implicit network is then constructed from 4.9M pages with 53.2M sentences and contains 3.6M entities, 5.8M terms, and 2.8B edges.

Query Response Time. Since the network data is massive (400GB in JSON format), supporting efficient query processing is a valid concern. However, despite the size of the data, we achieve average query response times in the order of seconds when using a classic database architecture on secondary storage. When using an optimized, in-memory representation with collapsed edge attributes (i.e., sentence information is merged into the edge data structure) the runtime can be reduced to milliseconds, thus allowing real-time and interactive use of the approach even on very large data sets.

Ground Truth Data. We evaluate the performance of all methods on single-entity sentence extraction since data from Wikipedia is available that can be used to evaluate this task. While an additional multi-entity evaluation would be beneficial, we are unaware of any labelled data sets that are usable for such an evaluation. To obtain single-sentence descriptions of a variety of entities, we use the Wikipedia glossary pages for astronomy[4], biology[5], chemistry[6], and geology[7]. All four pages have a list structure with items and brief explanations in the form of a few sentences that can be automatically extracted. For some examples from the glossary page of geology, see Table 1. To link the entities to nodes in the network, we use embedded links in a similar approach to the initial entity recognition. We extract all items from these lists that have an associated Wikidata identifier and use only single-sentence descriptions for the evaluation. The sizes of the resulting evaluation sets are given in Table 2. Note that the exact sentences do not occur in the collection of documents that we use for the network construction since the glossaries are formatted as lists (and are therefore structured). Thus, we evaluate on the task of extracting descriptive similar sentences, not on retrieving exact matching sentences.

Table 1: Example of ground truth entities with Wikidata identifiers and descriptions from the glossary for geology.

entity	wikidata	description
archipelago	Q33837	a chain or cluster of islands
mineralization	Q6864409	hydrothermal deposition of economically important metals in the formation of ore bodies or lodes
tectonics	Q193343	large-scale processes affecting the structure of the earth's crust

4.2 Evaluation

We briefly introduce the evaluation setup and the employed evaluation metric, before discussing the results.

Evaluation Metric. The evaluation of extractive summarization and single-sentence descriptions is notoriously difficult due to the lack of measures with a genuine semantic comparison for short texts. However, since we are only interested in the relative performance of the four methods, this problem is less severe. To obtain some measure of comparison between the four ranking methods, we thus employ the accepted standard evaluation metric ROUGE [14]. We use the RxNLP Java implementation[8] for our evaluation, with enabled stemming and stop word removal. Due to the limited size of sentences in comparison to summaries, higher-order n-grams do not occur with meaningful frequency and we thus focus on ROUGE-1 as a performance measure.

Evaluation Setup. For each entity in each of the four ground truth data sets, we identify the corresponding node in the implicit entity network by using the Wikidata identifier. We then perform a ranking of all sentences that contain the entity and extract the top-ranked sentence according to each of the four methods. We compute the ROUGE-1 scores and calculate the macro-averages for each of the four data sets as well as for the entire ground truth.

Evaluation Results. In Table 2, we show the average ROUGE-1 precision, recall, and F-scores of the ranking methods M1 - M4 on the four evaluation sets and on the combined set of all entities. We use the $n = 5$ most relevant terms for selecting the best context of entities in a sentence for methods M2, M3 and M4. The best performing score for each evaluation set and each metric is highlighted in boldface. The overall performance is expectedly low due to the strictness of the task evaluation, but we find several clear patterns. Method M2 consistently has the best recall across all data sets, which we attribute to the missing normalization by sentence length.

[4] https://en.wikipedia.org/wiki/Glossary_of_astronomy
[5] https://en.wikipedia.org/wiki/Glossary_of_biology
[6] https://en.wikipedia.org/wiki/Glossary_of_chemistry_terms
[7] https://en.wikipedia.org/wiki/Glossary_of_geology
[8] http://www.rxnlp.com/rouge-2-0/

Table 2: ROUGE-1 precision, recall, and F-scores of all four sentence ranking methods at a relevant term count of $n = 5$ for each evaluation set, as well as evaluation set sizes. The best values for each metric and set are highlighted in bold.

set	#entities	M1			M2			M3			M4		
		prec	rec	F1	prec	rec	F1	prec	rec	F1	prec	rec	F1
astronomy	18	0.069	0.207	0.099	0.064	**0.248**	0.096	0.078	0.184	0.097	**0.084**	0.199	**0.109**
biology	167	0.086	0.181	0.105	0.075	**0.302**	0.106	**0.212**	0.133	0.127	0.160	0.179	**0.151**
chemistry	177	0.039	0.180	0.062	0.044	**0.316**	0.074	0.082	0.149	0.093	**0.084**	0.187	**0.107**
geology	225	0.053	0.144	0.072	0.061	**0.215**	0.090	**0.114**	0.129	0.100	0.105	0.150	**0.111**
all	587	0.059	0.167	0.079	0.060	**0.271**	0.090	**0.131**	0.138	0.105	0.113	0.171	**0.121**

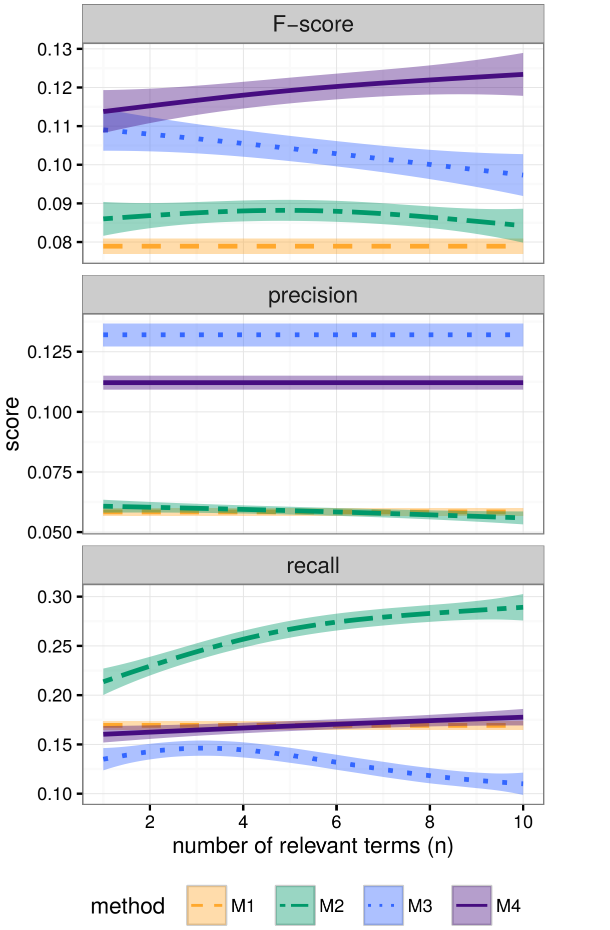

Figure 2: Average ROUGE-1 performances of the four sentence ranking methods for various relevant term counts n. Shaded areas denote 0.95% confidence intervals.

As a result, M2 favors longer sentences that are thus more likely to contain the key terms from the evaluation descriptions. The best precision is split between methods M3 and M4, depending on the data set. However, the difference in performance is fairly small in cases where M4 performs better and more pronounced in cases where M3 performs better, which indicates that M3 has a slightly higher performance with regard to recall for these data sets. Using the F-score as an overall measure, methods M3 and M4 consistently outperform the other two methods, with M4 performing best overall, as evident by its superior recall in comparison to M3. Despite the difficulty of the task, the relative gain in performance that is achieved by normalization by length is noteworthy, as we find a 33% performance increase of M4 over M2. Given that Wikipedia contains some extremely long sentences and that overly long sentences may occur due to errors in the sentence splitting step, using such a normalization is clearly favorable to obtain readable results.

To analyze the difference in performance of the four methods more closely, we also consider precision, recall, and F-score as we vary the number of relevant terms n (see Figure 2). Note that the performance of M1 is constant as it does not account for the occurrence of relevant terms. While M4 and M3 initially have similar F-scores for very low values of n, M4 benefits more from using additional relevant terms. For recall, the performance of M2 visibly exceeds all other methods by a large margin, which we attribute to the limitation of sentence length for M3 and M4. In contrast, the precision of the normalized methods is higher since the shorter sentences contain less noise. We find that increasing the number of relevant terms has no visible effect on the precision of M3 and M4, and even decreases the precision of M2. For recall on the other hand, M2 and M4 benefit from additional relevant terms, while the performance of M3 decreases. As a result, we find that M4 is best suited to obtain concise, descriptive sentences of moderate length and that it benefits the most from adding context through additional relevant terms. However, the tradeoff between the methods can be used to select an appropriate ranking approach that is tailored to the preferred result.

5 LOCATION RELATION EXTRACTION

We now apply the findings of the previous section to the extraction and description of novel relations between locations. Specifically, we investigate relations between pairs of locations that are significant in the implicit network created from Wikipedia, but not contained in Wikidata. Before we evaluate the coverage of discovered location relations with regard to the knowledge base, we first provide exploratory results as described in the following.

5.1 Exploration of Location Relations

The focus on entities of the type location requires an adjustment of the used data sets, which we describe first before presenting and discussing the exploratory results.

Data Preparation. As data set for this investigation, we utilize the same implicit network of entities from Wikipedia, but also consider a classification of entities according to the LOAD scheme into

Table 3: Example of descriptive sentences for pairs of locations that have no connecting relation in Wikidata, extracted from the text of the English Wikipedia. Shown are the two highest ranked sentences for five pairs of European cities.

Berlin (Q64)	Hamburg (Q1055)
1.	Berlin being the largest and Hamburg being the second largest city in Germany, they are also German states in their own right, having made both Wowereit and von Beust also state premiers.
2.	Two cities in Germany, namely Berlin and Hamburg, are considered city-states (German: "Stadtstaaten").
Berlin (Q64)	Vienna (Q1741)
1.	In the same way that Vienna was the center of Austrian operetta, Berlin was the center of German operetta.
2.	Robert Bodanzky, also known as Danton (born 18 March 1879 in Vienna, Austria-Hungary as Isidor Bodanskie, died 2 November 1923 in Berlin, Germany), was an Austrian journalist, playwright, poet and artist.
Athens (Q1524)	Sparta (Q5690)
1.	Although Thebes had traditionally been antagonistic to whichever state led the Greek world, siding with the Persians when they invaded against the Athenian-Spartan alliance, siding with Sparta when Athens seemed omnipotent, and famously derailing the Spartan invasion of Persia by Agesilaus.
2.	The Greek historian Thucydides wrote in his History of the Peloponnesian War of how, in 416 BC, Athens attacked Milos for refusing to submit tribute and refusing to join Athens' alliance against Sparta.
Athens (Q1524)	Corinth (Q103011)
1.	During the first centuries of the city's existence, imported Greek articles predominated: pottery (see Kerch Style), terracottas, and metal objects, probably from workshops in Rhodes, Corinth, Samos, and Athens.
2.	In the wake of this battle, Athens, Thebes, Corinth, and Argos joined together to form an anti-Spartan alliance, with its forces commanded by a council at Corinth.
Rome (Q220)	Milan (Q490)
1.	It was set up in 1958 in Rome and now is settled in Milan and represents all the highest cultural values of Italian Fashion.
2.	Italian fashion is dominated by Milan, Rome, and to a lesser extent, Florence, with the former two being included in the top 30 fashion capitals of the world.

locations, organizations, actors, and dates. Since the classification of entities in Wikidata can be difficult due to the ever-changing hierarchies, we instead use YAGO classes. To this end, we first merge the YAGO class hierarchy to the set of entities by matching the corresponding Wikipedia page URLs of items in both knowledge bases and subsequently use it to assign entities into classes. For locations in particular, we utilize all entities that either have the class `yagoGeoEntity`, or are located in its subtree. The resulting data set has 242.8K entities of type location.

Exploration Results. In Table 3, we show examples of descriptive sentences for pairs of European cities as extracted with ranking method M4. Since we do not aim to recreate relations that are already available in knowledge bases, we restrict these examples to pairs of cities that are not connected by a relation in Wikidata. For each pair of cities, we show only the two highest ranked sentences (in the case of ties, we randomly selected two sentences among the top-ranked sentences). We find that the sentences describe interesting relations between the cities that would be hard to qualify as discrete relations in a knowledge base but are well captured by descriptive sentences. For example, we find a variety of cultural relations, such as the importance of Berlin and Vienna for operetta, or Rome and Milan as centers of Italian fashion. The examples also include hierarchical similarities such as the city states Hamburg and Berlin (note, however, that the sentence as contained in Wikipedia is not entirely correct since Bremen is commonly considered to be the third city state of Germany). Interestingly, the approach also identifies historic relations that are unlikely to be covered in contemporary knowledge bases, such as the relation of ancient Greek city states in the Peloponnesian War. A common occurrence is the extraction of sentences that pertain to persons moving between two cities (e.g., the second-ranked sentence for Berlin and Vienna). These sentences are artefacts of the predominance of biographical data in Wikipedia, which puts an emphasis on terms concerning the births, deaths, and movements of persons in relation to their locations of residence. We discuss possible avenues to address or even utilize this phenomenon in Section 6.

5.2 Evaluation of Coverage

Based on the premise that we can extract descriptive sentences for novel relations that are outside of the scope of knowledge bases, we also need to investigate the performance of the implicit entity network to identify and rank such relations in the first place. To this end, we extract ranked lists of locations in the network neighbourhood of given input locations by applying entity ranking methods designed for implicit networks. We then compare the ranked list of related locations for each input location to the knowledge base relations of the input entity.

Evaluation Data. To obtain evaluation data for this task, we again turn to Wikipedia lists that we annotate through embedded links as described in the previous sections. Here, we use lists of locations as seeds, for which we then rank adjacent locations in the networks to identify relations that can be matched against Wikidata as a knowledge base. Specifically, we consider the list of largest German cities[9] (79 locations) and the list of international capitals[10] (250 locations) as input locations. We extract each of the

[9] https://en.wikipedia.org/wiki/List_of_cities_in_Germany_by_population
[10] https://en.wikipedia.org/wiki/List_of_national_capitals_in_alphabetical_order

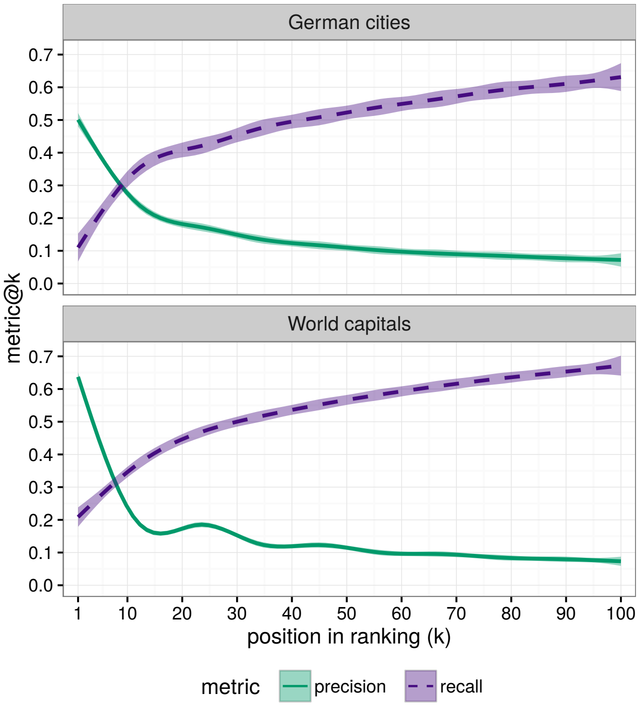

Figure 3: Precision@k and recall@k for the location relation extraction from the implicit Wikipedia network as compared to the location relations stored in Wikidata.

mentioned cities along with its Wikidata identifier and map it to the corresponding node in the implicit network.

Evaluation Setup. Since our aim is a comparison of related locations in the implicit entity network to linked locations in the knowledge base, we require a ranking function that extracts and ranks such locations from the network, given an input location. Here, we use the method described in the LOAD approach [23]. Specifically, we employ the two-tiered ranking approach that was introduced for entity ranking tasks [22] identical to the location relation extraction we are considering here. For a given input entity, the method computes a relatedness score for all connected entities of a specified type. We restrict the output entities to the type location to obtain a ranked list of locations that share a contextual relation to the input location.

Based on these rankings, we perform the comparison to Wikidata relations. That is, we label each extracted location in the ranked list as *positive* if the corresponding location is linked to the input location in Wikidata, and we label it as *negative* if no such link exists. From these labels, we then compute precision and recall values for different positions in the ranked list that indicate how similar or dissimilar the top-ranked items of the list are in comparison to the content of the knowledge base. Thus, a method with high precision and recall would identify exactly the same relations that are contained in the knowledge base, whereas a method with lower precision and recall would identify different relations. For the following evaluation, we limit the set of relations to pairs of locations that are mentioned at least once in a common context.

Evaluation Results. In Figure 3, we show the resulting precision and recall for the two evaluation data sets. In both cases, we find that the top-ranked locations from the implicit network differ strongly from the contained relations in the knowledge base. While the initial precision for low values of k shows that 50% to 60% of the top-ranked relations from the network also occur in the knowledge base, this rapidly drops to 10% at $k = 50$, thus indicating a different type of relation. However, the relations that are contained in the knowledge base are also contained in the ranked lists, but appear slightly further down the list as evident from the recall values. Since the implicit network is based on textual co-occurrences of location mentions, this is a reasonable observation. Hierarchical and containment relations that are typically stored in knowledge bases are much less likely to be mentioned frequently. As a result, the co-occurrences that are observable describe different, more complex relations. Thus, location rankings extracted from implicit networks can be used to identify pairs of locations sharing novel relations that are not contained in knowledge bases. These pairs of locations can thus serve as input for the extraction of descriptive sentences pertaining to location relations.

6 SUMMARY AND OUTLOOK

We give a summary of the presented work, before discussing implications and directions for future research.

6.1 Summary

In this paper, we introduced the task of entity-centric descriptive sentence extraction from implicit entity network representations of large document collections. We presented novel ranking functions for the extraction of location-centric descriptive sentences that utilize both entity and contextual term information. Due to their reliance on the local graph structure around input entities, the functions can be implemented efficiently and thus allow an interactive exploration of location relation descriptions in the underlying document collection. Using a ROUGE evaluation and empirical tests, we compared the proposed ranking functions on descriptive glossary entries from four different areas on an implicit network of entities constructed from Wikipedia. We found that a normalization by sentence length improves the perceived ranking results due to the elimination of overly long sentences, as well as the overall quality of the extracted sentences. In an application to the extraction of descriptive sentences for complex location relations, we found that an implicit network representation allows us to extract sentences that describe meaningful relations between pairs of locations. Furthermore, such networks also enable the extraction of location relations that are not contained in traditional knowledge bases with a high degree of certainty. In summary, our evaluations show that the ranking of location relations in implicit networks can leverage textual co-occurrences to identify and extract descriptive sentences for pairs of locations that are jointly mentioned in an underlying document collection.

6.2 Discussion

Two issues that we uncovered during this project deserve additional investigation, which we discuss in the following.

Effects of Sentence Length. As described in Section 4, a normalization of the sentence rankings by sentence length directly influences the performance, regardless of whether this normalization is based on character or term counts. Specifically, a normalization by length increases the overall precision (i.e., the words per sentence that occur in the evaluation description), but decrease the recall. While the overall benefit outweighs the drawbacks, as evident by the obtained F-scores, such a restriction may not always be desirable. For example, the extraction of concise sentences is paramount for extractive summarizers, but such limitations may not apply to a composition technique for summarization that first extracts multiple sentences covering different aspects in the context of a set of focus entities and then combines them. As a result, a less restrictive extraction may be favorable for such approaches. Thus, the selection of the proper sentence ranking method should be based on the given task at hand and on subsequent uses of the extracted sentences.

Further Entity Types. In our extraction of example sentences for location relations (see Section 5), we found that a substantial number of location relations pertains to the movement of people between places of residence. While we see this largely as an artefact of the amount of biographical data that is stored in Wikipedia and our focus on cities as examples, we also find that this can be exploited with the inclusion of additional entity types. The most direct approach would be to simply exclude or limit sentences that mention person movements from the results. However, additional entities could also be used to focus on certain aspects of relations for a subset of locations. For example, the exchange of scientists between universities can be considered through the extraction of substantiating sentences that explain the transfers. The addition of dates as entities would then even allow an analysis of the flow of persons from one place to another, given an appropriate document collection. Such an inclusion of temporal data would be a step towards an analysis of the evolution of relations between locations in documents with spatio-temporal content.

6.3 Ongoing and Future Work

In addition to the possible extension discussed above, we see future applications in the multi-sentence summarization of entities, where a balance of recall and brevity stands to be investigated for an optimal coverage of contexts. Given the fast response times of our approach, we see such an extension towards near real-time document summarization techniques as the natural next step. Our ongoing research includes the addition of more entity types beyond locations to work towards a general descriptive sentence extraction for (named) entities in large document collections. While such an extension seems natural, the relations between other types of entities may behave very differently from relations between locations. Thus, the validity of descriptive sentence extraction techniques and the quality of their results in such a setting remains to be explored.

REFERENCES

[1] Mohammed Al-Dhelaan. 2015. StarSum: A Simple Star Graph for Multi-document Summarization. In *SIGIR'15*. 715–718. DOI:http://dx.doi.org/10.1145/2766462.2767790

[2] Einat Amitay and Cécile Paris. 2000. Automatically Summarising Web Sites - Is There A Way Around It?. In *CIKM'00*. 173–179. DOI:http://dx.doi.org/10.1145/354756.354816

[3] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In *IJCAI'07*. 2670–2676. http://ijcai.org/Proceedings/07/Papers/429.pdf

[4] Fadi Biadsy, Julia Hirschberg, and Elena Filatova. 2008. An Unsupervised Approach to Biography Production Using Wikipedia. In *ACL'08*. 807–815. http://www.aclweb.org/anthology/P08-1092

[5] Roi Blanco and Hugo Zaragoza. 2010. Finding Support Sentences for Entities. In *SIGIR'10*. 339–346. DOI:http://dx.doi.org/10.1145/1835449.1835507

[6] Wei Chen, Eric Fosler-Lussier, Ningchuan Xiao, Satyajeet Raje, Rajiv Ramnath, and Daniel Sui. 2013. A Synergistic Framework for Geographic Question Answering. In *ICSC'13*. 94–99. DOI:http://dx.doi.org/10.1109/ICSC.2013.25

[7] Shruti Chhabra and Srikanta Bedathur. 2014. Towards Generating Text Summaries for Entity Chains. In *ECIR'14*. 136–147. DOI:http://dx.doi.org/10.1007/978-3-319-06028-6_12

[8] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *J. Artif. Intell. Res.* 22 (2004), 457–479. DOI:http://dx.doi.org/10.1613/jair.1523

[9] Johanna Geiß, Andreas Spitz, Jannik Strötgen, and Michael Gertz. 2015. The Wikipedia Location Network: Overcoming Borders and Oceans. In *GIR'15*. 2:1–2:3. DOI:http://dx.doi.org/10.1145/2837689.2837694

[10] Oskar Gross, Antoine Doucet, and Hannu Toivonen. 2014. Document Summarization Based on Word Associations. In *SIGIR'14*. 1023–1026. DOI:http://dx.doi.org/10.1145/2600428.2609500

[11] Hyun Duk Kim, Malú Castellanos, Meichun Hsu, ChengXiang Zhai, Umeshwar Dayal, and Riddhiman Ghosh. 2013. Ranking Explanatory Sentences for Opinion Summarization. In *SIGIR'13*. 1069–1072. DOI:http://dx.doi.org/10.1145/2484028.2484172

[12] Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A Survey on Question Answering Technology from an Information Retrieval Perspective. *Inf. Sci.* 181, 24 (2011), 5412–5434. DOI:http://dx.doi.org/10.1016/j.ins.2011.07.047

[13] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195. DOI:http://dx.doi.org/10.3233/SW-140134

[14] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL'04 Workshops*, Vol. 8. Barcelona, Spain.

[15] Elena Lloret and Manuel Palomar. 2012. Text Summarisation in Progress: a Literature Review. *Artif. Intell. Rev.* 37, 1 (2012), 1–41. DOI:http://dx.doi.org/10.1007/s10462-011-9216-z

[16] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2015. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR'15*. http://cidrdb.org/cidr2015/Papers/CIDR15_Paper1.pdf

[17] Mohsen Mesgar and Michael Strube. 2016. Lexical Coherence Graph Modeling Using Word Embeddings. In *NAACL HLT'16*. 1414–1423. http://aclweb.org/anthology/N/N16/N16-1167.pdf

[18] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In *EMNLP'04*, Vol. 4. 404–411.

[19] Ani Nenkova and Kathleen McKeown. 2012. A Survey of Text Summarization Techniques. In *Mining Text Data*. 43–76. DOI:http://dx.doi.org/10.1007/978-1-4614-3223-4_3

[20] José M. Perea-Ortega, Elena Lloret, Luis Alfonso Ureña López, and Manuel Palomar. 2013. Application of Text Summarization Techniques to the Geographical Information Retrieval Task. *Expert Syst. Appl.* 40, 8 (2013), 2966–2974. DOI:http://dx.doi.org/10.1016/j.eswa.2012.12.012

[21] Tye Rattenbury, Nathaniel Good, and Mor Naaman. 2007. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *SIGIR'07*. 103–110.

[22] Andreas Spitz, Satya Almasian, and Michael Gertz. 2017. EVELIN: Exploration of Event and Entity Links in Implicit Networks. In *WWW'17 Companion*. 273–277. DOI:http://dx.doi.org/10.1145/3041021.3054721

[23] Andreas Spitz and Michael Gertz. 2016. Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events. In *SIGIR'16*. 503–512. DOI:http://dx.doi.org/10.1145/2911451.2911529

[24] Thomas Steiner. 2014. Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A Global Study of Edit Activity on Wikipedia and Wikidata. In *OpenSym'14*. 25:1–25:7. DOI:http://dx.doi.org/10.1145/2641580.2641613

[25] Camille Tardy, Gilles Falquet, and Laurent Moccozet. 2016. Semantic Enrichment of Places with VGI Sources: a Knowledge Based Approach. In *GIR'16*. 6:1–6:2. DOI:http://dx.doi.org/10.1145/3003464.3003470

[26] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85. DOI:http://dx.doi.org/10.1145/2629489