An Event-centric Model for Multilingual Document Similarity

Jannik Strötgen
Institute of Computer Science
Heidelberg University
Heidelberg, Germany
stroetgen@uni-hd.de

Michael Gertz
Institute of Computer Science
Heidelberg University
Heidelberg, Germany
gertz@uni-hd.de

Conny Junghans
Institute of Computer Science
Heidelberg University
Heidelberg, Germany
cj@uni-hd.de

ABSTRACT

Document similarity measures play an important role in many document retrieval and exploration tasks. Over the past decades, several models and techniques have been developed to determine a ranked list of documents similar to a given query document. Interestingly, the proposed approaches typically rely on extensions to the vector space model and are rarely suited for multilingual corpora.

In this paper, we present a novel document similarity measure that is based on events extracted from documents. An event is solely described by nearby occurrences of temporal and geographic expressions in a document's text. Thus, a document is modeled as a set of events that can be compared and ranked using temporal and geographic hierarchies. A key feature of our model is that it is term- and language-independent as temporal and geographic expressions mentioned in texts are normalized to a standard format. This also allows to determine similar documents across languages, an important feature in the context of document exploration. Our approach proves to be quite effective, including the discovery of new similarities, as our experiments using different (multilingual) corpora demonstrate.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Clustering—*Similarity measures*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language models, Text analysis*

General Terms

Algorithms, Languages

1. INTRODUCTION

When are two documents similar? This is the question essential to tasks such as document classification, clustering, or topic detection and tracking. To address this often subjective and application specific question, over the past decades several approaches for modeling the similarity of text documents have been developed (see, e.g., [12]). Respective approaches range from extensions to the vector space model, such as latent semantic analysis [6], to techniques that employ explicit knowledge representation schemes such as ontologies to estimate semantic relatedness (e.g., [9]). In general, documents can be similar according to the words they contain, their structure, the topic they address, or the semantic concepts that have been associated with the documents. This plethora of different similarity aspects clearly does not lead to a single, universally applicable document similarity measure. Instead, different measures may lead to new insights into document similarity that cannot be captured by just one single approach.

In this paper, we present a novel similarity approach that is based on an event-centric document model. Loosely speaking, an event is something that happens at a specific place and time. Our claim is that for many categories of documents, events are essential to describe a topic or theme. This holds, among others, for historical works or biographies, an aspect we will elaborate on later in this paper. Our approach is based on a simple characterization of an event: it is solely described by nearby occurrences of a temporal expression and a geographic expression in a document's text. We thus do not rely on complex statistical NLP-based techniques or rich ontology-based frameworks for extracting event descriptions from documents. In particular, the proposed model leads to a small footprint for documents as only sets of events extracted from documents need to be managed and used for the proposed document similarity measure.

There are two key aspects to temporal and geographic expressions describing events. First, both can be normalized in an explicit way. No matter what original expressions are used in the documents, they can be normalized to the same value represented in a standard date notation (for a temporal expression) or as a coordinate (for a geographic/location expression). Grounding our event-centric model on such types of information makes the model term- and language-independent. The model therefore allows to compare documents written in different languages. Second, with both temporal and location information simple concept hierarchies are associated that specify a containment relationship. For example, the granularity of temporal information (as part of an event) can be of type day, like '2011-01-14', or type month, like in '2011-01', with the former being contained in the latter. Similarly, different granularities can be associated with geographic information, such as city, county, country, etc. Exploiting the different levels of granularity

SIGIR'11, July 24–28, 2011, Beijing, China.

for event specifications provides for an interesting technique to compare events extracted from documents. Thus, two documents may still be similar in terms of the events they contain even though the documents do not describe exactly the same events. Exploring event-based similarities among documents may also lead to new information that was not explicit before. For example, even though two documents talk about completely different topics, they both may mention the same event. This new information then can be used to investigate and establish new cross-document references.

In summary, the paper's main contributions are as follows:

- We describe a concise and succinct way to represent event information extracted from documents in the form of so-called document event profiles.
- We incrementally develop and present a model for an event-based document similarity measure that is term- and language-independent. The measure is based on normalized temporal and geographic expressions that can be compared using simple hierarchies for temporal and geographic concepts, respectively.
- We demonstrate the effectiveness of our new approach using different evaluations. For this, we investigate bilingual corpora (English and German) to show that our model identifies new document similarities across languages and topics, which are usually not detected by term-based methods.

The remainder of the paper is structured as follows. After a review of related work, we detail the extraction and representation of event information in documents in Section 3. In Section 4, we outline a similarity measure for pairs of events and extend this approach in Section 5 to compare event sets extracted from documents. In Section 6, we present a comprehensive evaluation of the proposed approach using different (multilingual) corpora. We conclude our paper in Section 7.

2. RELATED WORK

There are many approaches to the computation of document similarity in different IR related tasks such as document classification and clustering. In standard information retrieval, documents are typically represented as vectors, as are the queries [2, 14]. These vectors consisting of weights, such as tf-idf, for all terms in the documents are then used to calculate the similarity between a query and a document or two documents. There are numerous approaches to improve this standard similarity measure, e.g., by using (probabilistic) latent semantic analysis to analyze conceptual contents [4, 6]. Chim and Deng use phrase-based document similarity for clustering and show that feature vectors of phrase terms can be seen as expanded feature vectors for single-word terms [5]. Because of the large footprint of the vector space model for documents, the MapReduce framework was recently applied to large corpora for calculating document similarity [8].

An interesting empirical evaluation of models for text document similarity was conducted by Lee et al. [12]. They conclude that many automatic models have very good precision, but poor recall. That is, they detect only a subset of highly semantically similar document pairs. This observation is a motivation for our approach, because we do not want to replace existing similarity measures, but we do want to provide a measure for non-standard document similarity to identify new information, that is, event-based similarity relationships between documents.

An area related to our approach is topic clustering and in particular topic detection and tracking (TDT) where items of a document stream (e.g., a news stream) are analyzed. The goals of these approaches are to detect new unreported news events, and to track topics by assigning documents to already detected events [1]. There is a lot of research dealing with TDT for which the identification of events is necessary. Often, the similarity measures use information of named entity recognition, for example, locations, temporal expressions, or person and organization names mentioned in the documents [13, 27]. In contrast to our work, however, TDT systems try to identify a main event that can be associated with documents. Our goal, on the other hand, is to identify as many events in documents as possible, and to use the identified events for calculating document similarity.

Similar to our approach of applying concept hierarchies to temporal and geographic information, Lakkaraju et al. [11] use general concept trees to classify documents according to a taxonomy. Our concept hierarchies, however, are very small (less than 20 concepts in total) and specific to (standard) temporal and geographic aspects. There is also related research on similarities for event identification in social media [3]. In this work, however, they study general types of events and are not specifically concerned with temporal and geographic information. Some work combining geographic and temporal information extracted from documents for search and exploration tasks has been studied in [15, 20] but without focusing on document similarity.

As already pointed out, our model for document similarity is based on a combination of geographic and temporal information to identify events. We need to clarify that we do not consider events in the same way as it is done in linguistics, e.g., in TimeML, the standard markup language for temporal annotations [24]. Instead, we consider something to be an event if it is described by a temporal and a geographic expression. In the area of cross-language information retrieval, similarities are calculated on multilingual corpora. These calculations are usually based on translations while our approach uses normalized, language-independent information. There is only very few work on multilingual, translation-independent document similarity. One approach that makes use of a multilingual thesaurus for computing similarities has been proposed by Steinberger et al. in [18].

3. GEOGRAPHIC AND TEMPORAL INFORMATION IN DOCUMENTS

In contrast to most other works where a single temporal value (the document timestamp) and no geographic value are associated with a document, we use all geographic and temporal information mentioned in documents to form events. Therefore, we need to extract and normalize all geographic and temporal expressions using a geo-tagger and a temporal tagger. These named entity recognition and normalization tasks are outlined in the following Sections 3.1 and 3.2. In addition, we describe how the granularities of geographic and temporal expressions can be used for comparing expressions of different granularities. In Section 3.3, we combine geographic and temporal information into so-called document event profiles, which form the basis for our analysis of event-based document similarity.

3.1 Temporal Information

Temporal expressions can be grouped into four types, as done in the standard meta language for temporal annotations, TimeML [21]: *date*, *time*, *duration*, and *set*. In this paper, we only consider expressions of type *date*, although our method is suitable to support other types as well. Dates are anchored in a timeline, and a date's position in a timeline is referred to as its normalized value. It is important to note that the normalized value of a temporal expression, called *chronon*, is independent of the language used in the according document. Thus, this representation is perfectly suited for multilingual analysis tasks.

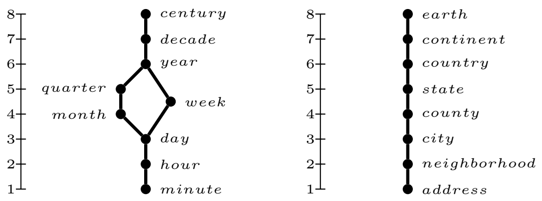
Normalization is done assuming a discrete representation of time, based on the Gregorian Calendar. Temporal expressions can be of different granularities, such as day, week, month, year, or decade. For the representation of these granularities, we assume different timelines, e.g., T_{day} for days and T_{month} for months. For example, 'March 19, 1999' can be anchored in T_{day} while 'May 2010' can be anchored in T_{month}. Timelines can be ordered into a hierarchy, which is a crucial point for our analysis of event-based similarities between documents. A hierarchy of temporal information granularities is given in Figure 1(a). Although there are other possible timelines (finer ones like hours or coarser ones like centuries), we use the following timelines for our hierarchy: $\mathcal{T} = \{T_{day}, T_{month}, T_{year}, T_{decade}\}$. A mapping can be performed from a value of a finer granularity to a coarser granularity, since months consist of days, years consist of months, and decades consist of years. The mapping from one timeline to the next coarser one is defined as one *temporal mapping step* α_t. For example, α_t('1999-03-19') ='1999-03' and $\alpha_t(\alpha_t$('1999-03-19')) = '1999'.

For the extraction and normalization of temporal expressions in documents, so-called *temporal taggers* are employed. Examples include GUTime, which is part of the Tarsqi toolkit [23], or HeidelTime [19], which was the best performing system for identifying and normalizing English temporal expressions in the TempEval 2010 challenge [24]. Using such temporal taggers, the following information about temporal expressions is extracted and used for further analysis: the offset (start and end position) of the expression in the document, the type (date, time, etc.), and chronon (normalized value). In Section 6, we give more details about the specific document processing pipeline employed in our approach.

3.2 Geographic Information

In the same way temporal expressions like dates refer to a point in time, geographic expressions refer to locations. We assume that with each geographic expression a geometry can be associated, which is represented as a latitude/longitude value pair. For the extraction and normalization of geographic expressions in documents, named entity recognition (NER) tools for locations can be used. Such *geo-taggers* usually use gazetteers and context information for the disambiguation of location names. Examples of such tools are MetaCarta [16] and Yahoo Placemaker [26]. Note that normalized location information is language-independent in the same way as temporal information.

Usually, geo-taggers only assign point geometry information to a location, regardless of the actual geographic extent of the location. Nevertheless, containment information, e.g., that a city is in a state or country, is often associated with an extracted expression. For example, the normalized containment information about 'New York City', which can be referred to by expressions such as 'New York, NY, US', 'New York', 'NYC', or 'Big Apple', is 'New York, NY, US'. This containment relationship is very important for geographic information and can be used directly and formalized in the following way: If a location g_i is contained in a location g_j, the relationship is denoted $g_i \subset_G g_j$. For example, the fact that a city g_1 is contained in a country g_2 can be expressed as $g_1 \subset_G g_2$.



(a) Temporal hierarchy. (b) Geographic hierarchy.

Figure 1: Concept hierarchies for geographic and temporal information.

Similar to the hierarchy of timelines for temporal expressions, the hierarchy of locations is crucial for our similarity measure. A general overview of the hierarchy of geographic information is depicted in Figure 1(b). Although there are many other levels (finer ones like addresses or coarser ones like continents), we use the following granularities for our hierarchy: city, state, and country. The mapping from one granularity to the next coarser one is defined as one *geographic mapping step* α_g. For example, α_g('San Diego, CA, USA') ='CA, USA'.

3.3 Document Event Profiles

As mentioned in Section 1, events in our scenario are co-occurrences of geographic and temporal expressions in a document. While there are several ways to form events from such expressions found in documents, we consider a pair of a geographic and a temporal expression to form an event if the two expressions occur in the same sentence. This method is much more precise than forming events from the cross-product of all pairs of geographic and temporal expressions in a document, which results in many false positives. Also, our method is much more efficient than a complete linguistic analysis of the documents, which is infeasible for large corpora.

We summarize the information about the geographic and the temporal dimension of all events in a given document d in a *document event profile*, denoted $dep(d)$. The idea behind document profiles is to describe the information textually mentioned in a document in a concise manner and to make it accessible for further analysis. More precisely, a document event profile $dep(d)$ contains a set of pairs $\langle(t_i, c_i, p(t)_i)$, $(g_i, v_i, p(g)_i)\rangle$ with (1) t_i and g_i being the temporal and geographic expressions occurring at the sentence level, (2) c_i and v_i being their normalized temporal (chronon) and geographic values, and (3) $p(t)_i$ and $p(g)_i$ being the offset of both expressions in document d. Instead of just recording the latitude/longitude information, we include the normalized containment information as the normalized geographic value v_i, since our similarity measure exploits this information. Note that events can be of different geographic granularities and timelines. Nevertheless, they can be compared using geographic and temporal mappings α_t and α_g, which were described in the previous paragraphs.

To summarize, in our approach, a document is solely described by an event profile. Compared to a vector space representation, our event-centric document model leads to a small amount of information that is used to compute the document similarity based on events. This computation and underlying similarity measure will be discussed next.

4. EVENT SIMILARITY

In this section, we introduce the fundamental step for calculating event-centric document similarity, which is the comparison of two arbitrary events. Specifically, we first describe the problem setting for computing the similarity of two events by specifying the requirements for such a similarity function in Section 4.1. In Section 4.2, we develop the similarity function and verify that all requirements are fulfilled. The similarity of pairs of events is then used in our new approach for calculating event-based document similarities, which will be detailed in Section 5.

4.1 Event Similarity Function Requirements

For comparing any two events, we use the normalized values of their temporal and geographic information as described in Section 3. Thus, an event is described by a chronon c and a normalized geographic value v. We call c and v the *dimensions* of an event $e = \langle c, v\rangle$. Both dimensions can be mapped to coarser granularities using the temporal and geographic mapping steps α_t and α_g introduced in Sections 3.1 and 3.2, respectively.

In order to define the requirements for the similarity function for events, denoted sim_e, we first list all possible similarity cases that can occur. In the listed cases, we do not distinguish the number of granularity mapping steps that have to be applied for one dimension, but use c^* for c and v^* for v being mapped to any coarser granularity. Given two events $e_1 = \langle c_1, v_1\rangle$ and $e_2 = \langle c_2, v_2\rangle$, the following similarities can occur:

1. Values of both dimensions of the events are identical.
   (1.1) $c_1 = c_2$ and $v_1 = v_2$

2. The values of one dimension have to be mapped to a coarser granularity.
   (2.1) $c_1^* = c_2$ and $v_1 = v_2$ (2.3) $c_1 = c_2$ and $v_1^* = v_2$
   (2.2) $c_1^* = c_2^*$ and $v_1 = v_2$ (2.4) $c_1 = c_2$ and $v_1^* = v_2^*$

3. The values of both dimensions have to be mapped to a coarser granularity.
   (3.1) $c_1^* = c_2$ and $v_1^* = v_2$ (3.4) $c_1^* = c_2$ and $v_1^* = v_2^*$
   (3.2) $c_1^* = c_2$ and $v_1 = v_2^*$ (3.5) $c_1^* = c_2^*$ and $v_1^* = v_2^*$
   (3.3) $c_1^* = c_2^*$ and $v_1^* = v_2$

4. It is not possible to map the values of one dimension to a coarser granularity to achieve equality.
   (4.1) $c_1^* \neq c_2^*$ and $v_1 = v_2$ (4.4) $c_1 = c_2$ and $v_1^* \neq v_2^*$
   (4.2) $c_1^* \neq c_2^*$ and $v_1^* = v_2$ (4.5) $c_1^* = c_2$ and $v_1^* \neq v_2^*$
   (4.3) $c_1^* \neq c_2^*$ and $v_1^* = v_2^*$ (4.6) $c_1^* = c_2^*$ and $v_1^* \neq v_2^*$

5. It is not possible to map the values of both dimensions to a coarser granularity to achieve equality.
   (5.1) $c_1^* \neq c_2^*$ and $v_1^* \neq v_2^*$

Using the listed set of possible cases, we now detail the requirements for the similarity function sim_e. For this, we state requirements followed by their detailed description.

R1: The more similar e_1 and e_2, the higher sim_e.

The more similar two events e_1 and e_2 are, the higher should be their similarity $sim_e(e_1, e_2)$, with $sim_e(e_1, e_2)$ being maximal if both events are identical. In (1.1), the events are identical and thus should have the highest possible similarity value.

R2: The fewer values of the same dimension need to be mapped, the higher sim_e.

In the second group, either at least one of the chronons has to be mapped to a coarser timeline (2.1 and 2.2) or at least one normalized geographic value has to be mapped to a coarser granularity (2.3 and 2.4). If only one value has to be mapped (2.1 and 2.3), the similarity score should be higher than if both values have to be mapped (2.2 and 2.4). This reflects the fact that the former ones can be correct (although described in an imprecise way) while the latter ones cannot be correct. For example, the sentences *he visits NYC in May 2010* and *he visits NYC on May 4, 2010* can talk about the same event. In contrast, the sentences *he visits NYC in May 2010* and *he visits NYC in April 2010* cannot talk about the same event. Nevertheless, there still is a similarity between e_1 and e_2 in the latter example, since both events happened at the same location and temporally close to each other (both in 2010). Consequently, sim_e should be higher for (2.1) and (2.3) and penalize (2.2) and (2.4).

The same cases occur in the third group: If only one value of each dimension has to be mapped (3.1, 3.2), the similarity score should be higher than if both values of the same dimension are involved (3.3, 3.4, and 3.5).

R3: If mapping leads to no equality, sim_e should be 0.

In the fourth group, either the chronons or the normalized geographic expressions cannot be mapped to a coarser granularity to achieve equality. In the fifth group both types of normalized values cannot be mapped sufficiently. Note that whether such cases can occur depends on the used hierarchies for geographic and temporal mappings (cf. Figure 1). For example, if 'Earth' is used as the top level of the geographic hierarchy, then every geographic expression can be mapped to the top of the hierarchy. However, if 'country' is at the top level, then, e.g., cities located in different countries cannot be mapped to a coarser granularity to achieve equality. Thus, if no sufficient mapping can be found, the assigned similarity score sim_e is 0, even though there may be a temporal or geographic similarity when using a different hierarchy. Instead, such unmatched events influence the aggregated similarity score when comparing two documents, as detailed in Section 5.

R4: The fewer mapping steps are needed, the higher sim_e.

The similarity score additionally depends on the differences of granularity between either c and c^* or v and v^*. The granularities are defined in Sections 3.1 and 3.2, and are represented by the timelines for temporal information and by the containment hierarchy for geographic expressions. The larger the differences, the less precise the information.

R5: The finer the granularities, the higher sim_e.

So far, the original granularities of the values, i.e., before they are mapped to coarser granularities, are not taken into account. For example, if there are two events $e_1 = \langle(2006),$ (Germany)$\rangle$ and $e_2 = \langle$(2006-07-09), (Berlin,Germany)$\rangle$, then

$sim_e(e_1, e_1)$ should not equal $sim_e(e_2, e_2)$, as the similarity score should be sensitive to the original granularities of the events in the documents. An event that is mentioned more fine-grained in the document should be weighted higher than a coarser one, i.e., $sim_e(e_1, e_1) < sim_e(e_2, e_2)$ should hold.

4.2 Event Similarity Function

Having defined the requirements for a similarity measure for pairs of events, we now formalize a function $sim_e(e_1, e_2)$ that fulfills these requirements. As stated above, $sim_e(e_1, e_2)$ > 0 only holds if equality of two events $e_1 = \langle c_1, v_1 \rangle$ and $e_2 = \langle c_2, v_2 \rangle$ can be achieved, namely by applying a certain number of mapping steps to the geographic and temporal dimensions of the events. We define $\alpha = \alpha_t + \alpha_g$ as the number of mapping steps that are needed to achieve equality for events e_1 and e_2 in both dimensions. Specifically, α_t is the sum of the number of temporal mapping steps that need to be applied to c_1 and c_2, respectively, in order to achieve equality in the temporal dimension. α_g is the corresponding sum of the number of mapping steps applied to v_1 and v_2 in the geographic dimension. That is, $\alpha \in \{0, \ldots, 2k\}$ with k being the total number of possible geographic and temporal mapping steps. Furthermore, we define β to be the maximum of the number of values per dimension that are involved in the mapping, thus $\beta \in \{0, 1, 2\}$. We tentatively define the event-centric similarity $sim_e(e_1, e_2)$ to be calculated in the following way:

$$sim_e(e_1, e_2) = \frac{1}{(1+\alpha)^{\beta}} \qquad (1)$$

While α is used to moderately decrease $sim_e(e_1, e_2)$, β increases the denominator exponentially, thus penalizing the similarity score stronger than α. This is motivated by requirement R2 that e_1 and e_2 can refer to the same event if $\beta = 1$, but cannot if $\beta = 2$ – no matter how large α is.

Equation 1 fulfills requirements R1 through R4 as will be shown below. However, it does not yet support R5 (the finer the original granularities, the higher sim_e). Thus, we additionally consider a parameter α_{poss}, which is the number of mapping steps (both temporal and geographic) that are still possible for e_1 and e_2 after both events have been mapped to be of equal granularity in both dimensions. That is, for example, if e_1 and e_2 were mapped to both represent $\langle(c = 2006\text{-}06, v = \text{Germany})\rangle$, then $\alpha_{poss} = 2$, as no further mapping step is possible for v and two more mapping steps are possible for c (year and decade). By weighting sim_e with $\alpha_{poss} + 1$, R5 is supported by our similarity function. Adding 1 to α_{poss} is necessary as the similarity of the coarsest granularity would be 0 otherwise.

$$sim_e(e_1, e_2) = \frac{1}{(1+\alpha)^{\beta}} \times (\alpha_{poss} + 1) \qquad (2)$$

Equation 2 fulfills all requirements R1 through R5, and can thus be used for calculating the similarity of two events. To exemplarily verify this, we calculate the similarity scores between four events (Table 1) and show that all five requirements are met. For better readability, we demonstrate this example using only the granularities day, month, and year for the temporal dimension, and city and country for the geographic dimension.

id	events
(e_1)	$\langle$(2006),(Germany)$\rangle$
(e_2)	$\langle$(2006-07),(Stuttgart,Germany)$\rangle$
(e_3)	$\langle$(2006-06),(Berlin,Germany)$\rangle$
(e_4)	$\langle$(2006-06-09),(Berlin,Germany)$\rangle$

	e_1	e_2	e_3	e_4
e_1	1	0.33	0.33	0.25
e_2		3	0.04	0.03
e_3			3	1.5
e_4				4

Table 1: Events (left) and similarity scores between them calculated using Equation 2 (right).

Although R1 (the more similar e_1 and e_2, the higher sim_e) is a subjective formulation, there are some examples in Table 1 for which this formulation is obvious, e.g., e_4 is more similar to e_3 than to e_1. This example shows that sim_e is calculated correctly with respect to R1, since $sim_e(e_3, e_4) >$ $sim_e(e_1, e_4)$. R4 (the fewer mapping steps are needed, the higher sim_e) is considered by sim_e since, e.g., $sim_e(e_3, e_4)$ (one mapping step is needed) is higher than $sim_e(e_1, e_4)$ (three mapping steps are needed). The fact that R2 is taken into account can be shown directly using Equation 2. If zero, one, or two values of the same dimension need to be mapped, then β equals 0, 1, or 2, respectively. For $\beta = 0$, the denominator of Equation 2 equals 1. If $\beta > 0$, then $\alpha > 0$ and thus, $(1+\alpha) < (1+\alpha)^2$, i.e., R2 is fulfilled since $sim_e(\beta = 1) > sim_e(\beta = 2)$. The consideration of R5 (the finer the granularities, the higher sim_e) is already achieved by the modification from Equation 1 to Equation 2. Finally, if no equality can be achieved with any number of mappings, Equation 2 is defined as $sim_e(e_1, e_2) = 0$, i.e., R4 is fulfilled.

5. EVENT-CENTRIC DOCUMENT SIMILARITY

Defining the similarity of just two events is already not trivial and many requirements need to be satisfied by the similarity function, as discussed in the previous section. However, aggregating the similarity of two sets of events in a meaningful way is even more challenging. Therefore, before defining how to calculate a respective aggregation, we first give the problem statement and define some requirements for this aggregation in Section 5.1. Then, an event-based document similarity model satisfying these requirements is incrementally developed in Section 5.2.

5.1 Problem Statement

Informally, the problem of computing the event-centric document similarity can be stated as follows: Given two documents d_1 and d_2 represented through their document event profiles $dep(d_1)$ and $dep(d_2)$, with each profile describing a multiset of events. Using the profiles, compute the event-centric document similarity $sim_e(d_1, d_2)$ in a concise and meaningful way.

To be able to specify a suitable similarity function, we first formalize some requirements for the aggregation of event similarities that need to be fulfilled:

A1: the more matching events are in d_1 and d_2, the higher $sim_e(d_1, d_2)$

A2: the more non-matching events are in d_1 and d_2, the more $sim_e(d_1, d_2)$ should be penalized

A3: if only one document contains additional events, this should not be penalized as much as if both documents contain additional non-matching events

In addition, all the requirements formulated for event similarities apply here, too. That is, requirements R1 to R5 described in Section 4.1 can be summarized as:

A4: the more similar the events in d_1 and d_2, the higher $sim_e(d_1, d_2)$

5.2 Aggregation of Event Similarities

Given a document, the objective now is to determine a ranked list of most similar documents using the information given by their document event profiles. The simplest way to calculate this similarity is to view all events as terms. For every document, these terms then form a vector so that the similarity between two documents can be calculated by comparing their vectors with, e.g., the cosine similarity function. This simple approach satisfies A1. However, other requirements are not fulfilled, in particular A4 is not taken into account. Therefore, we base our model on the similarity function between events introduced in Section 4.2 (Equation 2). For this, we use the following methods:

- *granularity mapping* (M): both dimensions of an event are mapped to all coarser granularities to enable the comparison between two different events
- *granularity weighting* (W): two matching events are weighted according to their granularities

To satisfy A1 and A3, we furthermore utilize

- *quantity normalization* (N): the similarity score calculated for two documents is normalized with respect to the number of events in their document event profiles

For realizing these methods, the vector approach is not applicable since not only exact matches are considered but matches between events after granularity mapping. Therefore, instead of comparing vectors, we perform event alignment by building the cross-product of the document event profiles to compare all event pairs. If two events are not equal, they will be mapped to coarser granularities until equality is reached or no further mapping is possible. The similarity score for every pair is calculated according to $sim_e(e_1, e_2)$ (cf. Equation 2) and aggregated to $sim_e(d_1, d_2)$.

However, requirement A2 is not fulfilled so far. Therefore, we have to normalize $sim_e(d_1, d_2)$ according to the number of events in the documents. For two documents d_1 and d_2 containing n and m events, respectively, using the sum $n+m$ violates A3. Thus, we use $min(n, m)$ for normalization and $sim_e(d_1, d_2)$ is thus calculated as follows:

$$sim_e(d_1, d_2) = \frac{\sum_{i=0}^{n} \sum_{j=0}^{m} sim_e(e_i, e_j)}{min(n, m)} \tag{3}$$

This equation together with the methods M, W, and N represents our full document similarity model (FM). To analyze the influence of the different methods, we use three further models in our evaluation: a model without granularity mapping (FM-M), a model without granularity mapping and granularity weighting (FM-MW), and additionally a model without normalization (FM-MWN).

To calculate $sim_e(d_1, d_2)$ efficiently, we map and materialize the events to all coarser granularities before comparing pairs of events. This way, the mapping does not have to be done every time an event is compared with another one. An example showing how the granularities are used for the calculation is given in Table 2. At the top of Table 2(a) and 2(b), the original events of document d_1 and document d_2 are given. Below them, the original events and their mappings are grouped by α_{poss} since two events with different α_{poss} values cannot be equal, and thus do not have to be compared. Note that for better clarity, we only use day, month and year, and city and country as temporal and geographic concepts in the hierarchies, respectively. In Table 2(c), we show the steps for comparing some events.

Algorithm 1 CalculateSimE($dep1, dep2$)

```
1: $sim = 0$
2: $aposs = (3, 2, 1, 0)$ {allowed mapping steps}
3: $mappings1 = create\_mappings(dep1)$
4: $mappings2 = create\_mappings(dep2)$
5: for all $e1$ in $dep1$ do
6:   for all $e2$ in $dep2$ do
7:     todo.add($e1.id\_e2.id$) {store ids in a hash set}
8: for all $poss$ in $aposs$ do
9:   for all $e1$ in $mappings1(poss)$ do
10:    for all $e2$ in $mappings2(poss)$ do
11:      if $e1.term = e2.term$ then
12:        for all $id1$ in $e1.getidset$ do
13:          for all $id2$ in $e2.getidset$ do
14:            if $todo(id1\_id2)$ then
15:              $todo.remove(id1\_id2)$
16:              $sim$ += $get\_sim(id1.t, id1.g, id2.t, id2.g, poss)$
17:              if $todo.length = 0$ then
18:                return $sim/min(dep1.length, dep2.length)$
19: return $sim/min(dep1.length, dep2.length)$
```

The pseudocode for computing the similarity of two documents represented by their event profiles $dep1$ and $dep2$ is shown in Algorithm 1. In line 2, all values of α_{poss} are listed in descending sort order, with 3 being the highest value in our example. In lines 3 and 4, the possible mappings of all events in the document event profiles of d_1 and d_2 are precalculated. This results in the mappings grouped by α_{poss} as shown in Table 2(a) and 2(b). Using the ids of the events, the cross-product is calculated in lines 5 to 7 and stored in the hash set *todo*. Then, the real calculation starts with iterating over the sorted α_{poss} and the events contained in the mappings of the respective α_{poss}. If two events match (line 11), all ids of the events are combined by iterating over the events' ids (lines 12-13). In the example (Table 2), four event pairs are compared for $\alpha_{poss} = 3$ with one matching pair (a,f). Once $\alpha_{poss} = 0$, one event pair is left containing several ids for each document. However, not all event pairs are still in *todo*, i.e., their similarity might already have been calculated (e.g., for a,f). If the id pair is still in *todo* (line 14), the pair is removed from the hash set (line 15) and the event similarity score is added to the total score *sim* (line 16). The method *get_sim* calculates the similarity between two events according to Equation 2 (Section 4.2). Method *get_sim* accesses values α_t and α_g of events through their ids, e.g., *c.t* represents the value of α_t for event *c*. For better efficiency, the results can directly be stored in a hash map, since they are limited to the combinations of $id1.t$, $id1.g$, $id2.t$, $id2.g$, and *poss*. Finally, if the *todo* hash set is empty (lines 17-18) or if all α_{poss} are processed (line 19), the similarity score is normalized by dividing *sim* by the minimum of the lengths of $dep1$ and $dep2$ and returned. In the example, the aggregated similarity score is divided by 3, i.e., the number of events extracted from document $d2$.

In summary, using the above approach, a document similarity measure can effectively be computed solely based on document event profiles. Several optimizations can be applied to this computation (e.g., binning of events based on some granularity such as year and then only computing cross-products for same-year bins). Due to space constraints and to allow for a concise representation of our approach, these optimizations are not discussed in this paper. In the next section, we demonstrate that this new event-based similarity measure for documents is meaningful and also applies to multilingual corpora.

(a) Document d_1.

Original Events		
a	2006-07-09	Berlin, Germany
b	2006-06-09	Munich, Germany
c	2006-06	Germany
d	2006	Germany
Mappings of events		
$\alpha_{poss} = 3$		
$a_{0,0}$	2006-07-09	Berlin, Germany
$b_{0,0}$	2006-06-09	Munich, Germany
$\alpha_{poss} = 2$		
$a_{1,0}$	2006-07	Berlin, Germany
$b_{1,0}$	2006-06	Munich, Germany
$a_{0,1}$	2006-07-09	Germany
$b_{0,1}$	2006-06-09	Germany
$\alpha_{poss} = 1$		
$a_{2,0}$	2006	Berlin, Germany
$b_{2,0}$	2006	Munich, Germany
$a_{1,1}$	2006-07	Germany
$b_{1,1}$, $c_{0,0}$	2006-06	Germany
$\alpha_{poss} = 0$		
$a_{2,1}$, $b_{2,1}$, $c_{1,0}$, $d_{0,0}$	2006	Germany

(b) Document d_2.

Original Events		
e	2006-07-08	Bonn, Germany
f	2006-07-09	Berlin, Germany
g	2006	Germany
Mappings of events		
$\alpha_{poss} = 3$		
$e_{0,0}$	2006-07-08	Bonn, Germany
$f_{0,0}$	2006-07-09	Berlin, Germany
$\alpha_{poss} = 2$		
$e_{1,0}$	2006-06	Bonn, Germany
$f_{1,0}$	2006-07	Berlin, Germany
$e_{0,1}$	2006-06-09	Germany
$f_{0,1}$	2006-07-09	Germany
$\alpha_{poss} = 1$		
$e_{2,0}$	2006	Bonn, Germany
$f_{2,0}$	2006	Berlin, Germany
$e_{1,1}$, $f_{1,1}$	2006-07	Germany
$\alpha_{poss} = 0$		
$e_{2,1}$, $f_{2,1}$, $g_{0,0}$	2006	Germany

(c) Similarity Calculation.

pair	match	α	β	α_{poss}
a, e	$a_{1,1}$, $e_{1,1}$	4	2	1
a, f	$a_{0,0}$, $f_{0,0}$	0	0	3
a, g	$a_{2,1}$, $g_{0,0}$	3	1	0
...				
c, g	$c_{1,0}$, $g_{0,0}$	1	1	0
...				
d, g	$d_{0,0}$, $g_{0,0}$	0	0	0

Single similarities:

$$sim_e(a, e) = 0.08$$
$$sim_e(a, f) = 4$$
$$sim_e(a, g) = 0.25$$
$$\ldots$$
$$sim_e(c, g) = 0.5$$
$$\ldots$$
$$sim_e(d, g) = 1$$

Aggregated similarity:

$$sim_e(d_1, d_2) = \frac{\sum_i \sum_j sim_e(e_i, e_j)}{min(n,m)}$$
$$= \frac{0.08+4+0.25+\cdots+0.5+\cdots+1}{3}$$
$$= 2.23$$

Table 2: The calculation of $sim_e(d_1, d_2)$ for two example documents. Original events and their mappings contained in d_1 (Table (a)) and d_2 (Table (b)) are grouped by α_{poss}. Indices of event ids represent α_t and α_g. In Table (c) some examples for matching events and the similarity calculation are shown.

6. EVALUATION

In this section, we first discuss the evaluation objectives and scenarios followed by a description of the corpora used for evaluation and our document processing pipeline. Then, we present the evaluation results and compare event-centric document similarity to a simple term-based similarity measure, namely tf-idf combined with cosine similarity.

6.1 Objectives and Scenarios

Manual Evaluation. Evaluating event-centric similarity is a challenging task since no adequate gold standard is available. We cannot use standard similarity evaluation corpora as our goal is not to identify documents as similar that talk about the same topic in general, but only documents that contain similar events. Although there are evaluation corpora for related tasks such as topic detection and tracking (TDT), these are not suitable due to the different goals of TDT and our similarity model. While TDT systems associate a main event with documents and cluster incoming news articles according to these events, we take into account all events extracted from documents to calculate event-centric similarity scores. Thus, a straightforward way to evaluate our model is to use a corpus, calculate the similarities between all documents and manually check whether two documents are similar or not. However, this scenario is very labor-intensive and can only be done for a small subset of documents. The results of such a manual evaluation are presented in Section 6.6.

Cross-language Evaluation. Another way to evaluate our model is based on a multilingual corpus containing cross-language links between related documents from different languages. Intuitively, documents written in different languages having the same content (e.g., about the same person) can be assumed to be similar in an event-centric way. For example, the English and the German version of a biography can obviously be regarded as similar with respect to the mentioned events (e.g., birth, death, travels) – no matter whether or not the two documents are (partial) translations of each other. Using a multilingual corpus containing cross-language links, we can evaluate how often cross-language linked documents are the top-k most similar documents for each other. Note that the cross-language links are only used for evaluation purposes, and not considered for calculating the similarities. This evaluation scenario allows for a large-scale evaluation, and the results are described in Section 6.4.

6.2 Corpora Description

We created two corpora for our evaluation. The smaller corpus contains the English Wikipedia featured articles [25] and the corresponding German articles linked to the English ones through cross-language links. This corpus contains 3,124 English and 1,825 German articles. Note that the German articles are not necessarily featured. The reasons for choosing the Wikipedia featured articles are that (i) they are determined by the editors to be of high quality, and (ii) they are grouped into 30 categories and 13 biography subcategories. The latter fact allows for a detailed analysis which documents contain many events and for which topics our similarity model is particularly suitable. Table 3 shows some numbers for the categories. Topics such as history, warfare, or biographies tend to contain many events on average. We expect to achieve better results for them compared to categories containing only few events, such as computing or physics. The differing numbers for English and German can be explained by the different length of the documents. Featured articles tend to be very long.

As a second test set, we built a larger corpus to evaluate our model with more documents taken into account. We used the publicly available Wikipedia XML Corpus [7], containing Wikipedia articles as XML files. We selected the main collections of English and German articles consisting of 659,388 and 305,099 articles, respectively, and created a subset of all document pairs for which the English and the German article are available, resulting in 94,348 pairs.

	English		German	
Category	Documents	Events	Documents	Events
geography, places	178 (-/-)	198	114 (15/26)	31
history	115 (-/2)	186	57 (8/16)	30
education	37 (-/-)	140	15 (0/4)	8
sport, recreation	151 (-/3)	138	65 (7/16)	38
warfare	254 (-/3)	119	147 (14/51)	20
transport	107 (2/3)	115	32 (5/12)	21
all biographies	717 (2/11)	105	496 (47/81)	23
...				
video gaming	128 (12/32)	21	81 (24/33)	9
computing	18 (2/3)	17	15 (8/11)	2
physics, astronomy	92 (11/24)	14	81 (38/54)	2

Table 3: Statistics of the featured articles corpus grouped by category. In parentheses are the numbers of documents with no/less than 3 events.

	Featured articles			Wiki XML		
	English	German	Pairs	English	German	Pairs
Docs	3124	1825	1825	94348	94348	94348
Events	87	17		9	5	
$e \geq 1$	3059	1455	1447	59404	52843	44109
$e \geq 3$	2953	1173	1157	43516	35671	26909
$e \geq 5$	2868	985	972	34198	26280	18244
$e \geq 10$	2690	669	660	21054	14784	8518

Table 4: Statistics of both corpora with $e \geq x$ being the number of documents with at least x events.

Some details on both corpora, which will be of interest in the result sections, are given in Table 4. The large differences in the number of average events per document can mainly be explained by the different lengths of the documents. The featured articles tend to be much longer than other articles. In addition, the Wiki XML corpus contains several very short documents with just a couple of sentences. For both corpora, we only use the main parts of the articles, i.e., reference sections are ignored. These sections often contain bibliographic information including publication date and place, which are not relevant for the article itself.

6.3 Document Processing Pipeline

For the evaluation, we built a UIMA-based text mining pipeline [22]. We developed collection readers to access both corpora. Linguistic preprocessing, i.e., sentence splitting, tokenization, and part of speech tagging is done using components of the DKPro repository [10] and the Tree Tagger [17]. As temporal tagger, we chose HeidelTime [19] due to its best performance in the TempEval-2 challenge [24] for the extraction and normalization of temporal expressions from English documents. For identifying locations, we run Yahoo Placemaker [26] since it returns detailed containment information about extracted locations, which is crucial in terms of using the geographic hierarchy. While Yahoo Placemaker directly handles English and German documents, we had to develop a rule set for the temporal tagger to process German documents. Finally, we identify events by extracting co-occurrences based on the annotations of the temporal tagger, the geo-tagger and the sentence splitter. By applying our document processing pipeline, we obtain all the needed information to create document event profiles and thus to perform our event-centric document similarity analysis.

6.4 Cross-language Evaluation

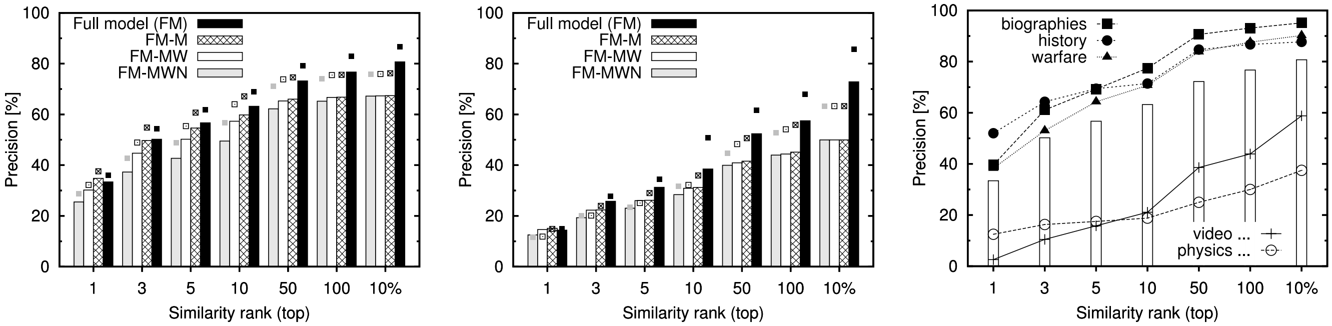
The hypothesis for the cross-language evaluation is that two documents linked by a cross-language link are similar in an event-centric way. Although we do not expect the linked documents to be the most similar ones for each other in all cases, we assume that they are among the most similar ones. We compare the four models introduced in Section 5.2, i.e., our full model (FM) to the model without granularity mapping (FM-M), to the model without granularity mapping and weighting (FM-MW), and to the one additionally without normalization according to the number of events (FM-MWN). For every document, its similarity to each other document in the corpus is calculated. Obviously, if one document does not contain any events, no similarity score can be calculated for this document. In such cases, the score is set to zero. In addition, the similarity score can be zero if no similarity between the events of two documents is identified after the mapping process. For all models, given one document, we determine if its cross-language linked document belongs to the x most similar ones. We set x to 1, 3, 5, 10, 50 and 100. In addition, we set x to 10% of all compared documents.

Figure 2 shows the results for the Wikipedia featured articles corpus (2(a)) and for the Wiki XML corpus (2(b)). For both corpora, we show results for document pairs where both documents contain at least 1 event (bars) and at least 3 events (points). In addition, we group the results for the Wikipedia featured articles by the category of the documents (2(c)) using only the full model.

For the featured articles corpus, using the simplest model FM-MWN already results in 26% (29%) precision that a cross-language linked document is the most similar one if both documents contain at least 1 (3) events. This indicates that, in general, events can be helpful for identifying similar documents. Adding the event-occurrences normalization, weighting of different granularities, and the mapping of fine grained events to coarser granularities significantly improves the results in all cases. The full model achieves a precision of 36%, 54%, 62%, and 69% (top 1, 3, 5, and 10, respectively) for document pairs containing at least 3 events. A further advantage of the full model is that due to the mapping of the events to coarser granularities, more similarities can be calculated. While the full model identifies similarities for cross-language linked document pairs in 86% and 94% of the cases, the other three models identify only 67% and 76% for pairs containing at least 1 and 3 events, respectively.

In Figure 2(b), the results for the Wiki XML corpus are given. Note that due to the much larger number of documents in the corpus, the chance for a cross-language linked document to be within the x most similar documents decreases. In addition, there may be more documents in the corpus being very similar in an event-centric way. This explains the lower precision results for the Wiki XML corpus compared to the smaller corpus. Nevertheless, using the full model (FM) for documents with at least three events, in 15% (28%, 34%, 51%, 62%, 68%, 86%) of the cases cross-language linked documents are within the top x most similar documents, with x being 1, 3, 5, 10, 50, 100, and 10%, respectively. In addition, the results show that the full model outperforms the other models significantly. Especially the mapping of events to coarser granularities provides a substantial feature to discover new similarities.

Figure 2(c) shows the results grouped by the categories of the documents. Categories rich in events clearly outperform categories containing just few events (cf. Table 3), with biographies performing best with 69% (77%) for the cross-language linked documents being within the top 5 (top 10) most similar documents for each other.



(a) Featured articles corpus. (b) Wiki XML corpus. (c) Results grouped by category.

Figure 2: Results of the cross-language evaluation. Figures (a) and (b) show results of the four models for documents containing at least one event (bars) and at least three events (points). Figure (c) shows the results of the full model for different categories (points) with average results for all categories (bars).

6.5 Comparison with Term-based Similarity

Although there are several advanced approaches to calculate document similarities as pointed out in Section 2, most of them are extensions to the vector space model and thus term-based. In comparison, our model is based on the document event profiles described in Section 3.3. Hence, we expect to find other kinds of documents to be similar compared to term-based models. Therefore, for comparison with term-based models, we do not have to use highly sophisticated methods such as latent semantic analysis, but can choose a simple model as representative for all term-based approaches. For this, we select the tf-idf measure combined with the cosine similarity denoted sim_t.

To evaluate the differences between sim_e and sim_t, we analyze pairs of documents (d_1, d_2) according to their rank for both similarity scores. This results in four categories:

c1. (d_1,d_2) is similar for both scores

c2. (d_1,d_2) is similar for sim_e, but not for sim_t

c3. (d_1,d_2) is similar for sim_t, but not for sim_e

c4. (d_1,d_2) is not similar for either scores

This evaluation is performed using the Wikipedia featured articles corpus. We use the top-n ranked documents for sim_e, with $n \in \{1, 3, 5, 10\}$, i.e., $rank_e(d_1, d_2) \leq n$ and calculate the ratio of documents that are similar using sim_t, i.e., for which $rank_t(d_1, d_2) \leq m$, with $m \in \{1, 3, 5, 10, 100, 1000\}$. As Figure 3 indicates, about 48% (56%) of the top-5 (top-10) ranked documents using sim_e are within the top-10 ranked documents using sim_t, and the best ranked document is the same for both measures in about 15% of the cases. This indicates that using sim_e leads to the discovery of new similarity relationships, which are hidden in the geographic and temporal information in documents. These cannot be discovered using standard similarity measures.

To demonstrate that the identification of similarities with both measures is still valuable, we give examples for pairs of documents for the categories (c1) to (c3). As the reference document d_1 for which similar documents are analyzed, we use the featured article '7 World Trade Center'. A document d_2 for which $rank_t(d_1, d_2) < 10$ and $rank_e(d_1, d_2) \gg rank_t(d_1, d_2)$, i.e., category (c3), is 'Tower of London'. Obviously, both documents do not talk about same events, but are similar with respect to the content, namely the topic of both documents is a tall building, its construction and design. A document for which $rank_t(d_1, d_2) < 10$ and $rank_e(d_1, d_2) < 10$, is 'American Airlines Flight 11'. Both documents are similar with respect to the topic of the terror attack and thus talk about same events, too. Finally, the article 'Ian Thorpe' is said to be similar in terms of mentioned events, i.e., $rank_e(d_1, d_2) < 10$ (with $rank_t \gg rank_e$). This surprising result is valid, as the article states that Ian Thorpe *was present at the World Trade Center on the morning of 11 September 2001, having stopped there on his jog, before returning to his hotel after forgetting his camera.*

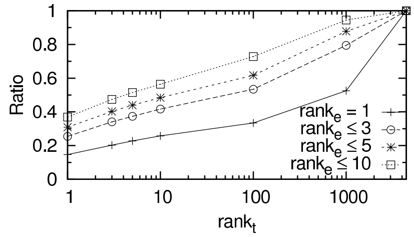


Figure 3: Ratio of documents that are ranked top n according to sim_t for different $rank_e$ values.

For further validation of the utility of sim_e, we perform a manual evaluation, which is described next.

6.6 Manual Evaluation

The objective of the manual evaluation is to validate the precision of the event-centric similarity model. For this, we use the Wikipedia featured articles corpus and randomly select 40 articles from the categories history, warfare, and biographies as source documents. These categories are especially suitable for event-centric document similarity as the cross-language evaluation showed since their documents contain many events (c.f. Table 3). We select the five most similar documents for each source document and evaluate whether they contain exactly the same events, belong to the same category, have a similar main topic, and are written in the same language.

The precision at 1, 3, and 5 for a document to be similar is 65%, 70%, and 64%, respectively. As expected, the results are independent of the language since 53% of the similar documents are not in the language of the source document.

Although there were many documents belonging to the same category (69%) and having a similar main topic (78%), there were several examples showing that sim_e identifies similarities across topics as well.

An error analysis showed that documents were wrongly identified to be similar mainly in the following cases: (i) one of the documents contains person names that were wrongly tagged as locations by the geo-tagger, and (ii) the source document contains no fine-grained events. While the first reason can be faced by using a high confidence value of the geo-tagger, the second one indicates that documents containing only coarse grained events are less suitable for our similarity model than documents with fine-grained events.

7. CONCLUSIONS AND ONGOING WORK

In this paper, we presented a novel model for the event-centric computation of document similarity. The model uses normalized event information, that is, pairs of temporal and geographic expressions extracted from documents to determine document similarity solely on the basis of events. Using this approach, it is possible to identify documents as being similar if they contain similar events although they might be different in other aspects, e.g., the main topic of the documents. In particular, because of the normalization of temporal and geographic expressions, including their containment relationships based on concept hierarchies, our model is applicable to multilingual corpora, an important feature not supported by other document similarity models.

We are currently extending our framework to include more languages, with a focus on Spanish and French, and to extend our corpora for evaluations correspondingly. We are also investigating structural aspects in terms of the order in which events are mentioned in a document. Such an order is obvious, for example, for the biography of a person written in two languages but might also lead to other interesting new information for other categories of documents. Finally, we are extending the types of events we consider by non-trivial events, such as names of holidays and their normalized temporal values.

8. REFERENCES

[1] J. Allan (Ed.). *Topic Detection and Tracking: Event-based Information Organization.* Kluwer Academic Publishers, USA, 2002.

[2] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley, USA, 1999.

[3] H. Becker, M. Naaman, and L. Gravano. Learning Similarity Metrics for Event Identification in Social Media. In *WSDM '10*, 291–300, 2010.

[4] T. Brants and R. Stolle. Finding Similar Documents in Document Collections. In *Proc. LREC-2002 Workshop on Using Semantics for Information Retrieval and Filtering*, 2002.

[5] H. Chim and X. Deng. Efficient Phrase-based Document Similarity for Clustering. *IEEE Trans. on Knowledge and Data Eng.*, 20(9):1217–1229, 2008.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. *J. American Society for Information Science*, 41:391–407, 1990.

[7] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 40(1):64–69, 2006.

[8] T. Elsayed, J. Lin, and D. W. Oard. Pairwise Document Similarity in Large Collections with MapReduce. In *ACL-HLT '08*, 265–268, 2008.

[9] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proc. Int. Joint Conf. on Artificial Intelligence*, 1606–1611, 2007.

[10] I. Gurevych, M. Mühlhäuser, C. Müller, J. Steimle, M. Weimer, and T. Zesch. Darmstadt Knowledge Processing Repository Based on UIMA. In *Proc. 1st Workshop on UIMA, Biannual Conf. Society for Comp. Ling. and Lang. Techn.*, 2007.

[11] P. Lakkaraju, S. Gauch, and M. Speretta. Document Similarity Based on Concept Tree Distance. In *Proc. ACM Conference on Hypertext and Hypermedia*, 127–132, 2008.

[12] M. D. Lee, B. Pincombe, and M. Welsh. An Empirical Evaluation of Models of Text Document Similarity, In *Proc. Annual Conf. Cognitive Science Society*, 1254–1259, 2005.

[13] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi. Topic Detection and Tracking with Spatio-Temporal Evidence. In *ECIR'03*, 251–265, 2003.

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

[15] B. Martins, H. Manguinhas, and J. Borbinha. Extracting and Exploring the Geo-Temporal Semantics of Textual Resources. In *Proc. International Conference on Semantic Computing*, 2008.

[16] MetaCarta Inc. http://www.metacarta.com/.

[17] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. International Conference on New Methods in Language Processing*, 1994.

[18] R. Steinberger, B. Pouliquen, and J. Hagman. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. In *3rd Int. Conf. on Computational Linguistics and Intelligent Text Processing*, 415–424, 2002.

[19] J. Strötgen and M. Gertz. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In *SemEval'10*, 321–324, 2010.

[20] J. Strötgen, M. Gertz, and P. Popov. Extraction and Exploration of Spatio-Temporal Information in Documents. In *Proc. 6th Workshop on Geographic Information Retrieval*, 2010.

[21] TimeML. http://www.timeml.org/.

[22] UIMA. http://uima.apache.org.

[23] M. Verhagen and J. Pustejovsky. Temporal Processing with the TARSQI Toolkit. In *Coling 2008: Companion volume: Demonstrations*, 189–192, 2008.

[24] M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. SemEval-2010 Task 13: TempEval-2. In *SemEval'10*, 57–62, 2010.

[25] Wikipedia Featured Articles. http://en.wikipedia.org/wiki/Wikipedia:Featured_articles.

[26] Yahoo Placemaker. http://developer.yahoo.com/geo/placemaker/.

[27] K. Zhang, J. Zi, and L. G. Wu. New Event Detection Based on Indexing-tree and Named Entity. In *SIGIR '07*, 215–222, 2007.