MultiWiki: Interlingual Text Passage Alignment in Wikipedia

SIMON GOTTSCHALK, L3S Research Center, Hannover, Germany
ELENA DEMIDOVA[1], University of Southampton, UK and L3S Research Center, Hannover, Germany

In this article we address the problem of text passage alignment across interlingual article pairs in Wikipedia. We develop methods that enable the identification and interlinking of text passages written in different languages and containing overlapping information. Interlingual text passage alignment can enable Wikipedia editors and readers to better understand language-specific context of entities, provide valuable insights in cultural differences and build a basis for qualitative analysis of the articles. An important challenge in this context is the trade-off between the granularity of the extracted text passages and the precision of the alignment. Whereas short text passages can result in more precise alignment, longer text passages can facilitate a better overview of the differences in an article pair. To better understand these aspects from the user perspective, we conduct a user study at the example of the German, Russian and the English Wikipedia and collect a user-annotated benchmark. Then we propose MultiWiki – a method that adopts an integrated approach to the text passage alignment using semantic similarity measures and greedy algorithms and achieves precise results with respect to the user-defined alignment. MultiWiki demonstration is publicly available and currently supports four language pairs.

CCS Concepts: •**Information systems** → **Wikis;** *Web applications;*

Additional Key Words and Phrases: Interlingual text alignment, Wikipedia

ACM Reference Format:
Simon Gottschalk, Elena Demidova, 2016. MultiWiki: Interlingual Text Passage Alignment in Wikipedia. *ACM Trans. Web* V, N, Article A (January YYYY), 31 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Articles containing information about entities of public interest become increasingly available in different languages on the Web, within community-created knowledge bases, encyclopedias and on the online news. As these sources evolve independently in each language, they often reflect community-specific points of view [Rogers 2013] and can contain complementary and sometimes contradictory information. This diversity is particularly interesting in the context of events influencing several communities (e.g. the Brexit, the refugee crisis in Europe and the Snowden affair). In order to provide an overview of the language and community-specific facets of the entities, help users to identify overlapping and complementary information and enable quality control in multilingual datasets, methods for effective identification and interlinking of related information across languages are required.

One prominent example of a large community-created interlingual data source is Wikipedia – an online encyclopedia available in more than 290 language editions, counting above 50 million users from all over the world and containing more than

[1]Corresponding author: Elena Demidova, demidova@L3S.de

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

DOI: 0000001.0000001

Table I: A user-aligned pair of text passages from the Wikipedia article "Gohi Bi Zoro Cyriac" in the English and the German Wikipedia, along with a manual English translation of the German text passage. The highlighted information about a move of the football player to Charlton Athletic in 2007 is only present in the German version of the article.

English Text Passage	**German Text Passage**	**German Text Passage (Translated)**
In March 2004, he joined ASEC Mimosas.[2] Cyriac was a topscorer of Côte d'Ivoire Premier Division in 2008 season. On 31 January 2009, he moved to Standard Liège signing a five-year contract with Belgian champions.	2004 wechselte er in die zweite Mannschaft von ASEC Mimosas, welche er 2007 Richtung England verließ. Der Ivorer unterschrieb bei Charlton Athletic, jedoch wurde er von den Addicks an seinen Jugendverein weiterverliehen. Mit den ASEC wurde er Vizemeister, zudem wurde er Torschützenkönig der höchsten ivorischen Liga. Nach einem weiteren halben Jahr in der Heimat wurde er im Januar 2009 von Standard Lüttich verpflichtet.	In 2004 he changed over to the second team of ASEC Mimosas, which he left towards England in 2007. The Ivorian signed at Charlton Athletic, however the Addicks immediately loaned him to his youth club. With ASEC he became vice champion, in addition he became topscorer of the highest Ivorian league. After a further half a year in his home country, he was signed by Standard Liège in January 2009.

30 million articles.[2] In Wikipedia, the articles representing equivalent real-world entities in different language editions become increasingly interlinked. In the following, we refer to such interlinked articles written in different languages as *partner articles*. The independent evolution of the Wikipedia language editions often leads to significant semantic differences across the partner articles. For example, Table I illustrates inconsistencies in the German[3] and English[4] versions of the article "Gohi Bi Zoro Cyriac" caused by the information contained only in the German version and related to the move of this footballer to Charlton Athletic in 2007. On a more general note, previous studies have shown the information asymmetries across Wikipedia language pairs: Although the English Wikipedia is by far the biggest with respect to the number of articles, edits and users[2], it has been shown that for many entities, Wikipedia articles in other languages are much longer than the corresponding descriptions in English and may contain contradictory information [Filatova 2009]. Paramita et al. [2012] conducted a user study on a random sample of 800 cross-lingual partner articles to find out that 28.8% of them are only moderately similar and 18.8% were judged to be different.

Precise alignment of text passages containing overlapping information in partner articles can enable users to obtain a comprehensive overview over common entity facets shared across the language editions and their language-specific context. On the one hand, manual alignment of overlapping information like it was performed in [Rogers 2013] can be very precise. However, such manual alignment requires user proficiency in a foreign language and can be a very time consuming and daunting task even for an expert user, especially for longer articles. On the other hand, an automatic alignment of semantically overlapping text passages is a challenging problem. This is due to the varying granularity of the overlapping text passages, differences in the text flow, additional information (facts or intermediate sentences) that does not have a direct correspondence in the other language as well as different linguistic structures used to

[2]https://meta.wikimedia.org/wiki/List_of_Wikipedias
[3]https://de.wikipedia.org/wiki/Gohi_Bi_Cyriac?oldid=136482800
[4]https://en.wikipedia.org/wiki/Gohi_Bi_Zoro_Cyriac?oldid=637509171

express equivalent information. Therefore, it is important to develop automatic methods that are able to identify semantically similar text passages, while being robust against syntactic differences.

We address the problem of the interlingual text passage alignment in order to facilitate a comprehensive overview of the information shared by partner articles. Existing approaches to interlingual text alignment are limited to few specific applications such as alignment of parallel fragments and sentences to support machine translation (e.g. [Chu et al. 2013], [Mohammadi and Ghasem-Aghaee 2010]) and detection of plagiarism cases (e.g. [Sanchez-Perez et al. 2015]). On the one hand, fragment and sentence alignment fail to provide an overview of the overlapping article parts due to their high granularity. On the other hand, interlingual plagiarism detection enforces strict conditions on the overlapping parts and is therefore not directly applicable to align text passages providing complementary information and having significant structural differences. Overall, existing text alignment methods are not suitable to provide a comprehensive overview of the interlingual article overlap.

In this article, we present a novel approach to *interlingual text passage alignment* across partner articles. To facilitate this alignment, we rely on semantic information including overlapping entities, time expressions related to common time intervals and selective terms. Text passages are extracted concurrently from partner articles and aligned based on their interlingual semantic similarity, while simultaneously enforcing granularity-related objectives and taking into account the interlingual context. In summary, the contributions of this article are as follows: (i) We present the problem of interlingual text passage alignment. To the best of our knowledge, this problem is not addressed by any existing approach; (ii) We propose an effective method for interlingual text passage alignment based on semantic similarity measures and greedy algorithms; (iii) We conduct a user study and create a user-annotated benchmark. We make this benchmark publicly available to encourage further research in this area. Our experiments demonstrate that the proposed method is effective with respect to both, precision of the alignment and granularity of the extraction. Text passages aligned by our method closely match the user-defined annotations.

MultiWiki demonstration is currently available in four language pairs: German-English, Dutch-English, Portuguese-English and Russian-English.[5] In this article we provide evaluation results for German-English and Russian-English pairs. The MultiWiki text passage alignment method presented in this article can facilitate a wide range of interlingual applications. For example, in our recent demo paper [Gottschalk and Demidova 2016] we presented a novel application to analyze the interlingual temporal evolution of the article pairs. This application enables users to observe the propagation of information across the language editions of the articles on a timeline and to perform a detailed visual comparison of the article snapshots at a particular point in time. Using the MultiWiki text passage alignment method presented in this article we can facilitate an effective visual comparison of the partner articles in this application.

The rest of this article is organized as follows: In Section 2 we provide a formal definition of the problem of the interlingual text passage alignment and provide an overview of our approach. Then, we present our methodology to address this problem, including: 1) A semantic similarity function described in Section 3; and 2) An interlingual text passage alignment procedure presented in Section 4. Following that, we describe the fine-tuning of the similarity function and its evaluation in Section 5. In order to further fine-tune our interlingual text passage alignment method and to enable its evaluation, we conduct a user study presented in Section 6 and collect a user-annotated bench-

[5] http://multiwiki.l3s.uni-hannover.de/demo.html

mark. Following that, we discuss the evaluation results of the text passage alignment method on German-English and Russian-English article pairs in Section 7. Section 8 provides a related work overview. Finally, we discuss our contributions, limitations of the approach and future plans in Section 9.

2. PROBLEM STATEMENT AND AN OVERVIEW OF THE MULTIWIKI APPROACH

The goal of the *interlingual text passage alignment* is to extract and align text passage pairs containing overlapping information from partner articles to facilitate an effective overview of the interlingual similarities and differences across these articles. To illustrate the interlingual text passage alignment from the user perspective, in Fig. 1 we present a user-annotated example from the partner articles "Ironworkers Memorial Second Narrows Crossing" – representing a bridge in Canada – in the English[6] and the German[7] Wikipedia. This article pair has been manually annotated by a user to identify overlapping interlingual text passages.[8] In this example, the user manually identified and labeled three interlingual text passage pairs, such as "Appearance", "Official Opening" and "Collapse" of the bridge in June 1958.

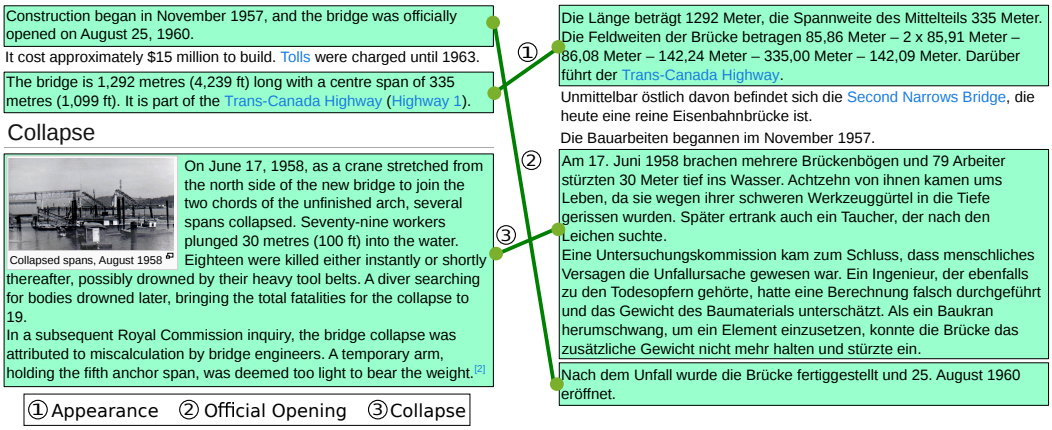


Fig. 1: A user-annotated example illustrating an extract from the partner articles entitled "Ironworkers Memorial Second Narrows Crossing" from the English and German Wikipedia language editions (as of the 1st October, 2015). Aligned text passages are enclosed by the bounding boxes, highlighted in green, connected via the green lines and manually annotated. Included photography: © Ron B. Thomson, licensed under CC BY-SA 3.0[9].

2.1. Problem Statement

In this section we first define the notions of a text passage, text passage alignment and the interlingual text passage similarity. Following that we discuss the interplay of the interlingual text passage similarity and the granularity of the extraction. Finally, we introduce the interlingual text passage alignment as an optimization problem.

[6]https://en.wikipedia.org/wiki/Ironworkers_Memorial_Second_Narrows_Crossing?oldid=674828683
[7]https://de.wikipedia.org/wiki/Ironworkers_Memorial_Second_Narrows_Crossing?oldid=130806835
[8]Our user study and benchmark creation is presented in Section 6 in more detail.
[9]https://creativecommons.org/licenses/by-sa/3.0/deed.en

A **Text Passage** is a non-empty list of consecutive sentences in an article. In the context of text passage alignment, we assume that text passages are topically coherent.

Definition 2.1 (*Text Passage*). Let $A_P = (s_1, \cdots, s_N)$ represent an article A_P from the language edition P through its sentence list $(s_1, \cdots, s_N)$, $N \geqslant 1$. Then, a *text passage* $T_{P_j} \subseteq A_P$ is a consecutive non-empty fragment of the sentence list A_P, such that all sentences in T_{P_j} are related to a common latent topic.

Text Passage Extraction and Alignment: Text passage extraction (i.e. the identification of the text passage borders within an article) and alignment (i.e. the identification of the most relevant text passage in the partner article) both depend on the interlingual context given by the partner article. To enable an efficient overview of the overlapping article parts, we define text passages in an article as mutually exclusive, i.e. containing non-overlapping sentence sequences. The alignment of text passages is mutually exclusive as well, i.e. a text passage is aligned to the most relevant text passage in the partner article.

Definition 2.2 (*Text Passage Alignment*). We use the notation $T_{P_j} \leftrightarrow T_{G_k}$ to represent the alignment of the text passages $T_{P_j} \subseteq A_P$ and $T_{G_k} \subseteq A_G$ from the partner articles A_P and A_G, respectively. The following conditions apply: (1) Text passages in an article are mutually exclusive, i.e. a sentence can belong to at most one text passage: $\forall T_{P_j}, T_{P_k} \subseteq A_P : T_{P_j} \bigcap T_{P_k} = \emptyset$; (2) Text passage alignment is mutually exclusive, i.e. one text passage is aligned to at most one text passage in the partner article; and (3) All sentences in $T_{P_j} \leftrightarrow T_{G_k}$ are related to a common latent topic.

Interlingual Text Passage Similarity: An interlingual text passage pair is *semantically similar* if these text passages share similar information nuggets. An information nugget can be an entity, a fact, a similar piece of information, or an answer to a question [Clarke et al. 2008]. The interlingual text passage similarity can be estimated using a similarity function based on semantic and syntactic features. The value of the similarity function should correlate with the overall similarity of the information nuggets contained in the text passage pair.

Definition 2.3 (*Interlingual Similarity Function*). Let $T_{P_j} \subseteq A_P$, $T_{G_k} \subseteq A_G$ be two text passages in the partner articles A_P and A_G, respectively. Then, $Sim_F(T_{P_j}, T_{G_k}) \in [0, 1]$ is the function that estimates an interlingual similarity of these text passages using a set $F = \{f_1, \ldots, f_N\}$ of semantic and syntactic features. Sim_F is monotonically increasing, with "1" corresponding to the highest similarity.

A Trade-off between the Similarity and Granularity of Aligned Text Passages: One way to maximize the similarity of the aligned text passages is to increase their granularity (i.e. to extract text passages that contain less sentences). In an extreme case this naive approach results in a large number of short text passages (e.g. text passages consisting of a single sentence each). However, such high-resolution alignment fails to provide a comprehensive overview of the common facets covered in the article pair. At the other extreme, in case of a low-granularity alignment, an entire article could be considered as one long text passage. Such alignment can likely result in low similarity due to the potentially high proportion of dissimilar information nuggets in the text passage pair, failing to meet the overview goal either. Therefore, an effective text passage alignment method should concurrently enforce the objectives related to the semantic similarity and the granularity of extracted text passages.

The Objectives of the Alignment: The *interlingual text passage alignment* aims at the following objectives:

1: Maximize the similarity of the aligned text passages in an article pair.
2: Minimize the overall number of the extracted text passages.

2.2. An Overview of the Text Passage Extraction and Alignment in MultiWiki

Given the optimization problem of the interlingual text passage alignment defined in Section 2.1, our method relies on the two key components: 1) A semantic similarity function that enables precise assessment of the interlingual text passage similarity for text passages containing overlapping information nuggets; and 2) A greedy algorithm that incrementally extracts similar text passages from the interlingual article pairs using their mutual context to create an effective alignment.

To enable fine-tuning and evaluation of the proposed method, we create two user-annotated benchmarks: 1) $Sim - B$ that provides continuous similarity values for text passage pairs at the sentence level and thus facilitates efficient fine-tuning and evaluation of the similarity function; and 2) $Align - B$ that contains aligned interlingual text passage pairs extracted and annotated by the users to facilitate fine-tuning and evaluation of the text passage extraction and alignment algorithm. To facilitate further research in this area, our benchmarks are publicly available.[10]

3. INTERLINGUAL TEXT PASSAGE SIMILARITY

In order to facilitate text passage alignment we need to estimate an interlingual similarity of text passages by instantiating the similarity function introduced in Definition 2.3. This function uses a set $F = \{f_1, \ldots, f_N\}$ of semantic and syntactic features. The choice of the features in this article is driven by two factors: 1) The availability of interlingual translation services and extractors that enable effective and efficient extraction of feature values in the interlingual settings; and 2) The intuition that the features correlate with the overall interlingual text passage similarity.

Intuitively, co-occurring selective terms and semantic annotations such as named entities and time expressions can substantially contribute towards precise text passage alignment, in particular in the case of partial information overlap. In order to facilitate computation of the term-based similarity, in this article we use English as a pivot language due to the relatively high availability of the translation services. In particular, in our experimental evaluation we use the Bing translation API that enables high quality machine translation in more than 50 languages.[11] Semantic features such as named entities and time expressions can be efficiently extracted and co-referenced in a number of languages using state-of-the-art tools such as DBpedia Spotlight [Daiber et al. 2013] and HeidelTime [Strötgen and Gertz 2013]. The feature set in our model is easily extendable, such that in case further interlingual semantic information extractors become proficient, new features can be added. For example, open relation extraction (e.g. in [Faruqui and Kumar 2015]) is an interesting direction to add more semantic information to the model in the future.

We assume a linear dependency between the features under consideration and the overall text passage similarity. The motivation for the linear combination is its simplicity, efficiency of training and computation as well as its effectiveness, as demonstrated by our results. Therefore, we model the similarity function as a linear combination. Using this modeling, feature weights can be efficiently learned from annotated datasets, e.g. using linear regression. The importance of the features in the context of the inter-

[10] http://multiwiki.l3s.uni-hannover.de/benchmark.html
[11] https://www.microsoft.com/en-us/translator/translatorapi.aspx

lingual text passage alignment is represented using the *feature importance factors* (or weights) $\beta_i \in [0,1]$, with $\sum_i \beta_i = 1$:

$$Sim_F(T_{P_j}, T_{G_k}) = \sum_{i=1}^{N} \beta_i \times sim(T_{P_j}, T_{G_k}, f_i), \quad (1)$$

where $sim(T_{P_j}, T_{G_k}, f_i) \in [0,1]$ is the similarity of text passages T_{P_j} and T_{G_k} computed using feature f_i.

3.1. Text Passage Alignment Features

In the following, we present the features that we found to be effective for the interlingual text passage alignment and the corresponding similarity computation in more detail.

Cosine Similarity (Co): *Cosine Similarity* measures the similarity of text passages using terms translated to a pivot language, while taking term frequency (tf) and selectivity (idf) of the terms into account [Manning et al. 2008]. To increase the precision of the alignment, the terms are pre-processed using stemming and stop word removal. Finally, the text passages are represented as vectors of $tf - idf$ term weights and the cosine similarity of the vectors is computed. Using *Cosine Similarity* text passage pairs containing selective terms, i.e. the terms that can distinguish a particular text passage pair from the rest of the corpus, are prioritized.

Text passage alignment using *Cosine Similarity* taken in isolation works well for parallel text passages that contain equivalent information. For example, this can be the case if one article is created as a translation of the other. In case of the partial overlap, term-based alignment is not sufficient to distinguish between semantic similarity, i.e. common information nuggets, and simple overlap in selective terms. Therefore, we do not expect *Cosine Similarity* taken in isolation to precisely distinguish between text passage pairs containing common information nuggets and text passage pairs containing parallel fragments. In order to enable for precise interlingual alignment of partially overlapping text passages, we use further features such as *Entity Annotations* and *Time Annotations*.

Entity Annotations (E): *Entity Annotations* are references to named entities mentioned in the text passages. Named entities are one of the key semantic components to support an effective alignment of text passages containing common information nuggets across languages. In order to enable effective usage of *Entity Annotations* for text passage alignment, interlingual entity co-referencing and sparsity of entity annotations need to be addressed.

Within a particular language named entity references can be extracted and co-referenced using existing annotation tools (e.g. DBpedia Spotlight [Daiber et al. 2013]). Also, interlingual annotation tools such as Babelfy become recently available [Moro et al. 2014]. In this work, we annotate the entities using DBpedia Spotlight in the original language versions of the articles and then establish interlingual links between the *Entity Annotations* using Wikipedia language links (i.e. the links connecting partner articles in Wikipedia). Another possible solution would be to use a machine translation service before named entity disambiguation is applied. However, we observed that machine translation services (such as Bing Translator API) often fail to correctly translate named entity labels.

Due to the sparsity and the distribution of the *Entity Annotations* in text passages, directly applying cosine similarity measure to these annotations does not lead to a very precise text passage alignment. For example, if two text passages in the English and the German article “Japan” only have a single *Entity Annotation* “Tokyo” each,

their similarity shall not be very high because of the high frequency of this annotation within the article. However, cosine similarity returns the maximum similarity of 1 because the vector representations of the text passages are identical in this case. This problem has also been observed in the related approaches that use sparse annotations and short texts as features (e.g. in [Duh et al. 2013]). Therefore, our approach is to put an additional emphasis on the highly selective entity annotations. To this extent, we compute the text passage similarity using the cosine of the vectors containing annotations and add a smoothing factor $\vec{n}$, further emphasizing selectivity of the entities in an article pair.

For each *Entity Annotation*, the smoothing factor shall be equal to 1 if this annotation is unique in both articles and shall approximate 0 if the annotation is very frequent. Moreover, the function should quickly decrease with the decreasing selectivity of the annotations. These conditions are fulfilled by a function for exponential decay. Therefore, we create a vector $\vec{n}$, where n_i is the weight of the *Entity Annotation* i computed as:

$$n_i = e^{-\frac{df_i}{\alpha}}, \tag{2}$$

where df_i denotes the number of sentence pairs (i.e. the shortest text passages) in the article pair containing the annotation i. Note that this smoothing factor does not take the length of the document into account. The weights computed by Equation 2 are in the interval $n_i \in (0,1]$ with the lower weights corresponding to the more common annotations. The greater α, the slower the decay: If α is very large, highly frequent annotations are assigned a greater smoothing value and vice versa. With $\alpha = 12.2$ we maximized the correlation between the proposed similarity function and the similarity derived from the user annotations on the training dataset TD_1 described later in Section 5 as measured using the Pearson Correlation Coefficient (PCC).

When calculating the similarity $sim_E(T_{P_j}, T_{G_k})$ of two text passages T_{P_j} and T_{G_k} based on the *Entity Annotations*, $tf-idf$ weights of the annotations are adjusted by $\vec{n}$:

$$sim_E(T_{P_j}, T_{G_k}) = \frac{\sum_{i=1}^{N} w_{i,T_{P_j}} w_{i,T_{G_k}} n_i}{\sqrt{\sum_{i=1}^{N} w_{i,T_{P_j}}^2} \sqrt{\sum_{i=1}^{N} w_{i,T_{G_k}}^2}}, \tag{3}$$

where w_{i,s_j} is the $tf-idf$ weight of the annotation i in the text passage s_j and N is the number of distinct aligned annotations in both articles.

Time Annotations (T): *Time Annotations* are normalized time expressions extracted from text passages. These annotations are an important factor to support an effective interlingual alignment, in particular with regard to the temporal facts. As Wikipedia is an encyclopedic text collection, time expressions play an important role: In a subset of English and German Wikipedia articles, we observed that on average 36% of the sentences contain time expressions. When analyzing time expressions mentioned in the text passages it is important to take their semantic similarity into account. First, extraction and normalization of time expressions allows a more accurate comparison of the described time intervals than a pure syntactic similarity. Second, time expressions describing longer time intervals, such as a year or a month are less precise and thus contribute less to the overall text passage similarity than more concrete time points such as a date. Thus, it is important to enable an accurate comparison of the (partially) overlapping time intervals taking the length of the interval into account.

Therefore, we assign relevance values to the time intervals ta according to their length: The longer the time interval, the smaller the relevance value. Following this intuition, we set the weight of the time interval $w(t_i) = 1$, if t_i represents a particular date, $w(t_i) = 0.85$ for a month and $w(t_i) = 0.6$ for a year. We experimentally observed

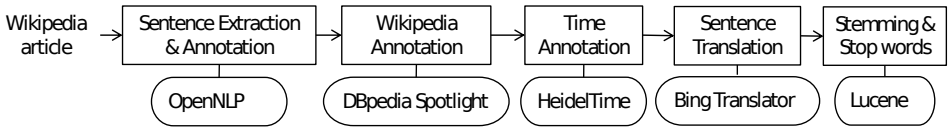


Fig. 2: Processing Pipeline for Feature Extraction

that the function configuration using these weights outperforms other configurations, such as equal weights for the time intervals of different length.

To compute the similarity based on the *Time Annotations* $sim_T(T_{P_j}, T_{G_k})$ between two text passages T_{P_j} and T_{G_k}, we align each *Time Annotation* $t_i \in ta_{T_{P_j}}$ with its best matching counterpart $t_j \in ta_{T_{G_k}}$ (if any) in these text passages and sum up the minimum relevance values of the aligned annotations to obtain a time overlap value $tovl$:

$$tovl(T_{P_j}, T_{G_k}) = \sum_{t_i \in ta_{T_{P_j}}} \sum_{t_j \in ta_{T_{G_k}}} \begin{cases} \min(w(t_i), w(t_j))\ ^* \\ 0,\ otherwise \end{cases} \tag{4}$$

*if t_i, t_j refer to an overlapping time interval, and there is no other overlapping $t_{j'} \in ta(T_{G_k})$ with a higher weight for $\min(w(t_i), w(t_{j'}))$.

If, for example, the annotations "2011/03/20" and "2011/03" are aligned, the relevance weight for a month is taken. The time overlap is computed for both directions, summed up and then normalized by the total number of the *Time Annotations* $|ta_{T_{P_j}}| + |ta_{T_{G_k}}|$ in the text passages T_{P_j} and T_{G_k}:

$$sim_T(T_{P_j}, T_{G_k}) = \frac{tovl(T_{P_j}, T_{G_k}) + tovl(T_{G_k}, T_{P_j})}{|ta_{T_{P_j}}| + |ta_{T_{G_k}}|}. \tag{5}$$

For example, if a text passage T_{P_j} contains the *Time Annotations* "2011/03/20" and "2011" and text passage T_{G_k} contains "2011/03", the similarity is calculated as $sim_T(T_{P_j}, T_{G_k}) = \frac{(0.85+0.6)+(0.85)}{2+1} \approx 0.767$.

3.2. Feature Extraction Pipeline

In order to facilitate interlingual text passage alignment we apply a processing pipeline including the following steps: Sentence splitting, entity annotation and annotation of time expressions in the original language; Sentence translation to a pivot language; Stemming and stop word removal from the translated sentences. This easily reproducible pipeline is implemented using state-of-the-art tools and services and is presented in Fig. 2.

3.3. Feature Weights

In order to obtain the feature weights, we created a benchmark $Sim - B$ described in Section 5 and performed function tuning. As a result, our similarity function (***CoET***) has the following feature weights: $\beta_{Co} = 0.69$, $\beta_T = 0.2$ and $\beta_E = 0.11$. As to not penalize the sentences with missing features, we have trained additional similarity function configurations for these cases: ***CoE*** ($\beta_{Co} = 0.89$, $\beta_T = 0$ and $\beta_E = 0.11$), ***CoT*** ($\beta_{Co} = 0.8$, $\beta_T = 0.2$ and $\beta_E = 0$) and ***Co*** ($\beta_{Co} = 1.0$, $\beta_T = 0$ and $\beta_E = 0$).

4. THE ALIGNMENT PROCEDURE

The alignment of text passages is an optimization problem that strives for high similarity of the extracted text passages and their low granularity simultaneously. The

brute-force approach to this problem is not feasible as the number of possible text passages and their alignments grows exponentially with the number of sentences in an article pair. Therefore, we propose a greedy approximation algorithm.

Intuitively, our method works bottom-up as follows: The algorithm starts with the alignment of the seed sentences in the partner articles with the similarity above a predefined threshold th. Then, it iteratively expands the alignment. As long as the similarity $Sim_F(T_{P_j}, T_{G_k})$ of a text passage pair can be increased by extending it with a close-by text passage pair, we merge them, such that the overall similarity of the aligned text passages increases and the granularity of the aligned text passages decreases. The overall similarity is measured as $\sum_{T_{P_j} \leftrightarrow T_{G_k} \in S_{A_P, A_G}} Sim_F(T_{P_j}, T_{G_k})$, where S_{A_P, A_G} is the set of all aligned text passages in the article pair A_P, A_G.

Due to the interlingual differences, aligned text passages can contain different number of sentences. Consequently, it does not suffice to merge directly neighbored text passage pairs. Therefore, we propose two options to incrementally extend a text passage pair:

— Merging with neighbored sentences: One of the text passages in a pair can be extended by a single neighbored sentence.
— Merging with (nearly) neighbored text passage pairs: Two text passage pairs can be merged if they are located in a close neighborhood. This implies the inclusion of intermediate sentences if the text passage pairs are not directly adjacent.

Based on an initial alignment of similar sentences, we propose a greedy algorithm that extends the currently most similar text passage pair in each step until no extension is possible any more (i.e. until no text passage pair can be merged with

To extract topically coherent text passages we rely on two estimates: 1) Interlingual context: Due to the different structure of the partner articles, text passages related to different topics are unlikely to come in the same order; Therefore if by adding more sentences to a text passage the interlingual similarity drops, the topics are likely to drift. 2) Text structure: Wikipedia articles are arranged in a hierarchical structure. Articles consist of *sections* that may contain sub-sections, sub-subsections and so forth. At the lowest level of the hierarchy, the text is split into *paragraphs*. The end of such (sub-)sections or paragraphs provides an indication of a possible topic drift; We introduce a parameter (Structure freedom (sf)) that allows different degrees of freedom with respect to this structure.

In the rest of this section, we explain the merging methods in more detail and then show how the algorithm utilizes these methods to create an alignment of text passages.

4.1. Merging of Text Passage Pairs with Neighbored Sentences

The information nuggets of one sentence can be scattered over a few sentences in the partner article, such that it is necessary to merge the sentences from the partner article to perform the alignment. Fig. 1 contains an example where this merging step is necessary: While the English article describes the bridge's collapse and the plunge of the workers in two sentences, the German article sums that information up just in one sentence: "Am 17. Juni 1958 brachen mehrere Brückenbögen und 79 Arbeiter stürzten 30 Meter tief ins Wasser." (Translated: "On June 17, 1958, several bridge arches collapsed and 79 workers plunged 30 meters into the water"). In this case, the content of the German sentence is scattered among two directly adjacent English sentences. Thus, the English sentences are merged to form a larger text passage that can be aligned to the single German sentence. More formally, if two consecutive sentences $s_i \in A_P$ and $s_{i+1} \in A_P$ both (partially) overlap with the same sentence $s_j \in A_G$ in the

other article, they can be merged into the text passage $(s_i, s_{i+1}) \subseteq A_P$, which is aligned to form the text passage pair $(s_i, s_{i+1}) \leftrightarrow (s_j)$.

In Algorithm 1, the procedure `mergeWithSentences` takes a text passage pair as an input and searches through all neighbored and unaligned sentences to merge them with it. When multiple text passages can be created that way, the one with the highest similarity Sim_F is returned.

4.2. Merging Nearly Neighbored Text Passages

Until now, we discussed the alignment with the help of single sentences. This is especially important at the beginning of the procedure to obtain a starting point for text passage extraction: Such aligned sentences constitute initial text passage pairs. In the next step, we merge text passage pairs in a close neighborhood. Due to the goal of similarity maximization, this merge may only be done if the similarity of the resulting text passage pair exceeds the similarity of the initial text passage pair. This condition ensures that the alignment simultaneously strives for low granularity and high overall similarity. Under that condition, two text passage pairs $T_{P_i} \leftrightarrow T_{G_k}$ and $T_{P_j} \leftrightarrow T_{G_l}$ can be merged, such that they are replaced by a single text passage pair $T_{P_{i'}} \leftrightarrow T_{G_{k'}}$, where $T_{P_{i'}}$ contains the sentences from T_{P_i}, T_{P_j} and, potentially, the *intermediate sentences* between them. Such intermediate sentences can provide complementary information and while included within the aligned text passages, they are put in context.

Whether the similarity score can grow by merging text passages depends on how similar information is fragmented in both languages. In cases where there is no 1:1 correspondence at the sentence level, merged text passages help to better assimilate fragmented parts to match the information available on both sides, overall resulting in higher similarity after merging. Although the inclusion of intermediate sentences in text passages can potentially result in lower similarity values, it is not necessarily always the case. As long as the similar parts in the merged text passages overweight, similarity values of the resulting alignment will be higher.

We enable this extension method by the function `mergeWithPassagePair`: Given a text passage pair $T_{P_i} \leftrightarrow T_{G_k}$, all text passage pairs $T_{P_j} \leftrightarrow T_{G_l}$ in the close neighborhood are chosen as candidates and the one that results in the highest similarity after merging is returned.

4.3. The Alignment Algorithm

With the help of the two merging functions `mergeWithSentence` and `mergeWithPassagePair`, we now define our algorithm ***MultiWiki*** to extract and align a precise and low-granular set of text passage pairs in a bottom-up manner. As shown in Algorithm 1, the input is are two articles, A_P, A_G, the similarity threshold th, and the structure freedom parameter sf.

In order to speed up the alignment process, the algorithm operates in a greedy manner: In each step it selects the currently most similar text passage pair $T_{P_i} \leftrightarrow T_{G_j}$ (line 5 - line 6). $T_{P_i} \leftrightarrow T_{G_j}$ is either merged with a neighbored sentence or – if this was already tried – with a neighbored text passage pair (lines 7 - 11). If the extension is successful and the merged text passage pair $T_{P_{i'}} \leftrightarrow T_{G_{j'}}$ achieves a higher similarity score than $T_{P_i} \leftrightarrow T_{G_j}$, the set of text passage pairs is updated accordingly (line 14 - 16). This procedure terminates when no text passage pair can be extended such that its similarity increases: In this case, S_{A_P,A_G} remains unchanged and the parameter $foundChanges$ stays `false`.

ALGORITHM 1: Text Passage Alignment Algorithm

Input: Articles $A_P,\ A_G$, similarity threshold th, structure freedom parameter sf.
Output: The set $S_{A_P,\ A_G}$ of the aligned text passages.

```
S_{A_P,A_G} = alignSentences(A_P, A_G, th);
foundChanges := true;
while foundChanges do
    foundChanges := false;
    sort(S_{A_P, A_G});
    for each T_{P_i} ↔ T_{G_j} in S_{A_P, A_G} do
        if not mergedWithSentence(T_{P_i} ↔ T_{G_j}) then
            T_{P_i'} ↔ T_{G_j'} = mergeWithSentence(T_{P_i} ↔ T_{G_j}, sf);
        end
        else
            T_{P_i'} ↔ T_{G_j'} = mergeWithPassagePair(T_{P_i} ↔ T_{G_j}, sf);
        end
    end
    if Sim_F(T_{P_i'}, T_{G_j'}) > Sim_F(T_{P_i}, T_{G_i}) then
        foundChanges := true;
        S_{A_P,A_G} = (S_{A_P,A_G} ∪ T_{P_i'} ↔ T_{G_j'}) \ T_{P_i} ↔ T_{G_j};
        break;
    end
end
return S_{A_P,A_G};
```

4.4. Parameters

There are two parameters that can be tuned to adjust the behavior of the proposed algorithm to make it better fit user preferences:

— **Similarity threshold (th):** A threshold value th that determines if a text passage pair is regarded as being similar. A lower threshold enables more flexibility in the merging, but can also affect the precision of the alignment.
— **Structure freedom (sf):** The likelihood of a topic drift shall be higher when reaching a new section or paragraph in the original text. Thus, to enhance the topical coherence of the aligned text passages, we can disallow the algorithm to merge sentences from different sections or paragraphs. We introduce three structure freedom levels: *max* (i.e. no limits), *mid* (i.e. never exceed a given section) and *min* (i.e. always stay within one paragraph).

We discuss parameter tuning and their influence on the overall effectiveness of the method in the evaluation described in Section 7.

5. SIMILARITY FUNCTION TUNING AND PERFORMANCE

In order to facilitate fine-tuning and evaluation of the interlingual text passage similarity function presented in Section 3, we created a benchmark $Sim-B$. This benchmark defines the similarity scores for the interlingual text passage pairs in the German and the English languages based on shared semantic information. We use parts of this benchmark for the fine-tuning of the similarity function as well as for the evaluation as discussed in the following.

5.1. The *Sim – B* Benchmark for Interlingual Similarity Computation

An important question for the benchmark creation is the selection of the text passage pairs to be annotated. On the one hand, the annotation of all possible text passage

pairs in the partner articles does not appear feasible due to their large number. On the other hand, the majority of text passage pairs that can be built in any partner article pair is rather dissimilar. Therefore, random selection of text passages would not lead to a sufficient number of similar pairs to train the similarity function. Hence, in order to fine-tune the feature weights β_{f_i} in Equation 1, we apply an iterative bootstrapping approach. This approach incrementally collects relevant text passage pairs and systematically refines the weights using supervised machine leaning and user feedback. For simplicity of the annotation, in $Sim - B$ we focus on short text passages, each consisting of a single sentence. In particular, we create three datasets:

The dataset TD_1: We first pre-select sentence pairs from a set of controversial Wikipedia articles [Yasseri et al. 2014] in the German and the English languages aligned via the language links to build the training dataset TD_1. This is performed by using the text passage similarity function following Equation 1 with a set of initial manually defined feature weights. In order to include the sentences that do not contain all features, we varied the feature weights, including 0-weights for the features based on the entity and time annotations. These sentence pairs are judged by users, such that we can learn feature weights for the similarity function using supervised machine learning (in particular, we utilize linear regression for this task).

The dataset TD_2: Then we iteratively refine the feature weights and incrementally collect sentence pairs for the second training dataset TD_2. This dataset contains randomly selected partner articles. We collect sentence pairs having similarity above a manually defined threshold (0.25) and further refine the feature weights. When no substantial changes in the feature weights are observed in the next iterations, we consider the feature weights to be optimal.

The dataset VD: Finally, we create a validation dataset VD. We use this dataset to evaluate the similarity function. This dataset contains sentence pairs extracted from randomly selected partner articles using pooling – a standard evaluation method in Information Retrieval [Manning et al. 2008]. To this extent, we retrieve the ranked list of the most similar sentence pairs generated by different similarity functions (the functions are described in Section 5.3). We ensure that the top-k results of each function are user-annotated. Other sentence pairs, i.e. those ranked below $k = 200$ by all similarity functions are considered to be dissimilar. Pooling method enables a fair comparison of the precision and recall values across the functions considered in the evaluation, even though the absolute recall values can be overestimated.

Table II provides an overview of the article selection method and the size of the datasets.

Dataset	Source	Articles	Possible Sentence Pairs	Annotated Sentence Pairs
TD_1	Controversial Articles	14	2016568	229
TD_2	Random Articles	20	260867	1233
VD	Random Articles	33	20358	300

Table II: Datasets consisting of the German and the English Wikipedia partner articles. During the function tuning and evaluation process, a subset of highly ranked sentence pairs aligned by different methods in each of these datasets has been annotated.

Page	Avg. Rating	Article ID$_1$	Article ID$_2$	Text$_1$	Text$_2$
European Union	1.0	en-635761078	de-136109478	In 2012, the EU was awarded the Nobel Peace Prize.	2012 wurde der Europischen Union der Friedensnobelpreis zuerkannt.
Nicolaus Copernicus	0.4375	en-634443003	de-134393612	He died about 1483.	Als sein Vater 1483 starb, war Nikolaus zehn Jahre alt.

Table III: Example text passages from the $Sim - B$ benchmark. In addition to the sentence pairs and similarity user ratings illustrated in this table, the benchmark also contains additional (dataset) IDs, the single user ratings, the articles, their sentences, semantic annotations and machine translations.

5.2. Similarity Annotations with Users

During the benchmark creation process described in Section 5.1, we annotated the initial training dataset TD_1 with 258 pre-selected sentence pairs in a user study. In total, 11 users (graduate CS students with good knowledge of both languages) participated in the user study. Each user performed at least 50 tasks (an average evaluation time of a set of 50 tasks was 30 minutes). In each task, the user was presented the English sentence and a list of one or more alignment candidates in German. The users were asked to classify each candidate as one of: "same content (i.e. facts)", "partly same content" or "different content" categories. In addition we made the options "don't know" and "corrupted sentence" available to the users. The last option helped to exclude sentences containing occasional errors introduced by the pre-processing from the evaluation. Each sentence pair in this dataset was evaluated by at least 8 users. In addition, we created a set of 12 manually selected parallel sentence pairs as well as 12 randomly selected mismatched sentence pairs to verify the user's input.

To compute the overall user-defined similarity scores for each sentence pair in this training dataset, we assigned the scores of 1.0 to the "same content", 0.5 to the "partly same content" and 0.0 to the "different content" judgements and computed the similarity of a sentence pair as an average of the user scores. As a result of the user study, we obtained an initial training dataset containing 229 aligned sentence pairs (29 corrupted sentences and the sentence pairs added for verification of users' input are ignored): 18 pairs with an average user score in the interval $[0.75, 1]$ (*parallel sentence pairs*), 102 pairs in $(0.25, 0.75)$ (*partially similar*) and 109 pairs in $[0, 0.25]$ (*different*). According to these numbers, there is at least a partial overlap in more than the half of the evaluated sentence pairs.

The datasets TD_2 and VD have been annotated using the same procedure, while employing a smaller number of annotators. Table III shows two examples of the sentence pairs with their ratings in the $Sim - B$ benchmark.

Difficulty of the similarity annotation task: To obtain a better understanding of the task difficulty for the users, we computed the Fleiss' κ [Gwet 2014], a statistical measure of agreement between individuals for qualitative ratings, as a measure of the reliability of the user agreement. Note that according to Fleiss' definition, $\kappa < 0$ corresponds to no agreement, $\kappa = 0$ to agreement by chance, and $0 < \kappa \leq 1$ to agreement beyond chance. Here, we considered the seven users with the most ratings. Each of these users evaluated the majority of the 229 sentence pairs in TD_1. If we do not differentiate between the partially overlapping and the parallel sentence pairs, κ value reaches 0.571, which is close to the "substantial agreement". For the three classes ("same content", "partly same content" and "different content"), agreement values are lower and correspond to a "moderate agreement" by using the intervals presented in [Landis

and Koch 1977] ($\kappa \approx 0.474$). In other terms, we could find that for 131 sentence pairs (57.21%), there has been at most 1 user disagreeing with the other users. According to these values, it is presumably easier for the users to decide whether two sentences are at least partially overlapping, than to differentiate between partially overlapping and parallel sentences in this corpus.

Sources of disagreement by similarity annotations: To better understand the reasons for the users' disagreement, we looked at those sentence pairs that were put into different classes by the users. We found that user disagreement can be typically observed in cases with differences in the author perspective and with missing context:

- Difference in the author perspective, generalization:
  - English: "Berlin is known for its numerous cultural institutions, many of which enjoy international reputation".[12]
  - German: "Die Sportereignisse, Universitäten, Forschungseinrichtungen und Museen Berlins genießen internationalen Ruf".[13] (Translated: "The sport events, universities, research institutions and museums of Berlin enjoy international reputation".)
- Missing context: References to other sentences, where the user has to consider the context of the sentence in the Wikipedia article to disambiguate the reference:
  - English: "The church was destroyed in the Second World War and left in ruins".
  - German: "Sie war durch Bombenangriffe im Zweiten Weltkrieg schwer beschädigt worden". (Translated: "It was heavily damaged by bombings in the Second World War".)

Although both sentence pairs in these examples contain similar information and could be viewed as parallel, some users classified them as partially overlapping or even different due to the language-specific differences in the information presentation. In order to increase the user agreement in the second case, missing context could be provided by presenting larger text passage context (e.g. paragraphs) to the users.

Usage of the user annotations for function training and evaluation: Although the absolute scores provided by the individual users can vary for some text passage pairs, the scores aggregated over multiple user judgments provide a comprehensive picture of the relative similarity across the text passage pairs. Such aggregated scores can be effectively used for training and evaluation of the similarity function.

5.3. Similarity Function Evaluation

In this work we establish the baseline for the alignment of text passages containing overlapping information nuggets across languages. Closest to our work, Duh et al. [2013] aimed at the identification of new information in Wikipedia articles and applied cosine similarity measure. This method uses terms obtained after machine translation, stop word removal and stemming. We use cosine similarity in isolation as a baseline to assess the relevance of the other semantic features we proposed. We use the following notations for the similarity function configurations:

- **Co**: Cosine similarity. This method uses terms obtained using machine translation, stop word removal and stemming.
- **CoE**: This function combines Cosine similarity with *Entity Annotations*.
- **CoT**: This function combines Cosine similarity with *Time Annotations*.

[12] http://en.wikipedia.org/wiki/Berlin?oldid=635429067
[13] http://de.wikipedia.org/wiki/Berlin?oldid=136234983

Similarity Function	Average Precision (AP)	
	All Sentences	Without Parallel Sentences
Co	84.35%	74.99%
CoE	85.73%	76.79%
CoT	89.77%	83.32%
CoET	**90.98%**	**84.63%**

Table IV: Average precision values for text passage similarity functions in the alignment task. For each of the functions, the top-200 sentence pairs were collected. Based on the resulting set of 267 sentence pairs (257 for non-parallel sentence pairs), the average precision values were computed.

— **CoET**: This function combines Cosine similarity with both, *Entity Annotations* and *Time Annotations*.

For the validation dataset VD, we collected the top-200 sentence pairs per similarity function under consideration. Given the resulting set of sentence pairs and the average user ratings per sentence pair, we compute the average precision (AP) values of the rankings achieved by the different alignment functions. For each sentence pair, its user similarity score is computed as an average over the scores given by the individual annotators. To determine the relevance of the retrieved sentence pairs for the average precision computation, we apply a threshold of 0.25 on the average user ratings (i.e. we assume that the sentence pair is relevant if an average user rating is ≥ 0.25).

As we can observe in Table IV, **Co** that only uses *Cosine Similarity* achieves an average precision of 84.35%. Additional semantic features we proposed in this article such as *Entity Annotations* in **CoE** ($AP = 85.73\%$) and *Time Annotations* in **CoT** ($AP = 89.77\%$) and in particular the combination of these annotations in **CoET** ($AP = 90.98\%$) enable us to further improve the average precision. This result confirms the high effectiveness of the proposed semantic features for the text passage alignment.

When only considering partially overlapping sentences, several differences can be observed: The absolute average precision values of all similarity functions drop, confirming that it is easier to align the parallel sentences than the partially overlapping ones. This decrease varies dependent on the similarity function: Compared to the case with the parallel sentences, **CoET** shows the lowest decrease, which emphasizes the value of the semantic annotations for the alignment of partially overlapping sentences: In this case **CoET** achieves up to 9.64% improvement in the average precision compared to the purely syntactic-based function **Co**.

An improvement when using the *Time Annotations* is higher than for the *Entity Annotations*, because although very selective *Entity Annotations* contribute to the precise text passage alignment, we cannot use less selective annotations to precisely differentiate the sentences with semantically meaningful information nugget overlap from the rest.

Overall, our evaluation results confirm that the use of semantic features leads to a more precise text passage alignment, in particular with respect to the partially overlapping sentences and indicates that *Time Annotations* and selective *Entity Annotations* are effective features towards identifying semantically similar text passages containing common information nuggets.

Page	User ID	Article ID$_1$	Article ID$_2$	Passage$_1$	Passage$_2$	Title
Winger (sports)	43	en-664310306	de-664310306	10-11 -12-13	7-8-9	Football
Johann Hugo von Orsbeck	43	en-668382450	de-144384256	4-5	8-9	Birth and parents
Johann Hugo von Orsbeck	47	en-668382450	de-144384256	29	61	The end

Table V: Example text passages from the $Align - B$ benchmark. In addition to the user-aligned text passage pairs illustrated in this table, the benchmark also contains the actual sentences.

6. TEXT PASSAGE ALIGNMENT BENCHMARK

In order to better understand the problem of the interlingual text passage extraction and alignment from the user perspective we collect the benchmark $Align - B$ for the method tuning and evaluation in a user study. The aims of the user study were to: 1) Better understand the difficulty of the manual text passage alignment task for the users; 2) Observe, analyze and learn from the user decisions regarding the alignment of text passages; and to 3) Create a benchmark to fine-tune and evaluate automatic methods for this task.

6.1. The $Align - B$ Benchmark for Text Passage Alignment

To collect the user annotations for the interlingual text passage alignment, we randomly selected a set of partner articles from the English and the German Wikipedia. As we focus on the text passages, article pairs where one of the articles mainly consisted of tables and lists were filtered out manually. The resulting dataset contains 55 article pairs coming from several domains and includes, for example, “General Post Office”, “Commuter Rail” and “George William Gray”. With regard to the split obtained by a sentence splitting algorithm, the English articles contain 21.32 sentences on average and the German ones 17.27, which makes a total of 2,123 sentences. Based on the 55 article pairs in this dataset, we created another dataset from the Russian and English Wikipedia. This dataset consists of the 21 article pairs whose articles can be found in the Russian Wikipedia as well.

In the user study, the user’s task was, given a pair of German-English or Russian-English partner articles, to extract and align similar text passages. In total, 12 users (graduate CS or mathematics students with good knowledge of both languages) participated in the user study for the German-English article pairs. Each user annotated 15 article pairs on average. All 55 article pairs were annotated by at least three users each, 14 of them by four different users. This makes a total set of 179 distinct user annotations of article pairs. In each of these annotations, at least one text passage pair was identified. The average number of text passage pairs annotated per article pair is 3.34. The 21 articles in the Russian-English dataset were manually annotated by at least one user. Three example text passage pairs in $Align - B$ are shown in Table V.

6.2. Task Description and User Interface

The user interface of the study is similar to the interface shown in Fig. 1: At the beginning, the user sees both articles without any marked text passage pairs. By clicking on the sentences, the user can incrementally create text passages in both articles simultaneously, expand the text passages by adding further sentences on each language’s side and finally confirm the alignment of the created text passage pair. To ensure topical coherence of the created text passage pair, the last step requires an input of a brief user-defined English title.

The instructions for the user included the following steps:

Step 1 Read both language versions of the article.
Step 2 Find and align a pair of similar text passages in the two articles. If several alignment candidates are available, select only the best matching pair. If similarity and topics are the same, prefer longer text passages.
Step 3 Give the aligned text passage pair an English title.
Step 4 Continue with Step 2 until all similar text passage pairs across both language editions of the article are aligned.

With these instructions we let the decision if any intermediate sentences should be included to the user, as long as the aligned text passages fulfill the conditions specified in Step 2.

6.3. Difficulty of the Text Passage Alignment Task

In order to better understand the difficulty of the task for the users, we measured the time spent by the users on the task and the user agreement. On average, a user spent approx. 6 minutes to annotate one article pair. The annotation time depends on the article length: If one of the articles is very short, users spent less than a minute; On longer articles, some users spent more than 15 minutes. Overall, we can observe that the task of text passage alignment can be very time consuming especially for longer articles, even for users with good knowledge of both languages.

To capture the overlap across text passages aligned by different users, we consider the intra- and interlingual sentence pairs within the annotated articles and check how different users assign these sentence pairs to text passages. In particular, the intra-lingual measurement estimates the agreement of the users on the extraction step (i.e. for each article we create the set of all possible sentence pairs and check for each sentence pair and rater whether the two sentences were put in the same text passage); The interlingual alignment step illustrates the agreement of the users on the alignment step (i.e. the agreement that two sentences, one from each partner article, belong to an aligned interlingual text passage pair).

We again compute the inter-rater agreement using Fleiss' κ-measure. When considering short articles (less than 2500 characters in both articles together), κ-values for both the extraction task and the interlingual alignment task approach 0.6 indicating substantial agreement. Considering all article pairs independent of their length, we measured $\kappa \approx 0.467$ for the extraction task, and $\kappa \approx 0.473$ for the interlingual alignment task. These κ values illustrate that the extraction and alignment tasks are similar with respect to their difficulty and moderate agreement is possible even in case of longer articles. This also confirms our initial observation performed by the time measurement, that in case of longer articles the task becomes more difficult for the users.

6.4. Observations

By analyzing the user annotations we made important observations with respect to the annotation structure that can help us to further fine-tune the text passage alignment model.

Alignment of the lead sentences: The lead sentence, i.e. the very first sentence of a Wikipedia article, and the lead sections of Wikipedia articles typically have a uniform style: Regarding the English Wikipedia manual of style, an article lead section "should be able to stand alone as a concise overview" and the first sentence "should tell the non-specialist reader what (or who) the subject is."[14] This subject description is rather

[14] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

independent of the specific language and thus it is very likely that there is an aligned text passage pair containing at least the lead sentence of both partner articles. In our study, this was the case for 92.74% of all the article pairs.

Comparison of the text passage length: As we enable users to include additional information in the aligned text passage pairs, although it only occurs in one of the text passages as long as it is related to a common topic, the length of the aligned text passages can differ. The results of our user study show that only 51.59% of the user-aligned text passage pairs have exactly the same number of sentences in both languages. On average, the length of the aligned text passages differs by 0.89 sentences; Overall, the interlingual difference in the length of the user-aligned text passages follows a power law distribution and can exceed three or more sentences in 8.21% of the cases.

Text passages vs. text paragraphs: We assume that each text paragraph and section in the original Wikipedia text hierarchy can form a topically coherent text passage. To estimate the usefulness of the Wikipedia text structure in the context of text passage alignment, we measured how the user-extracted text passages correlate with the article structure. We observed that out of 642 text passages extracted by the users (excluding those consisting of one sentence only), 481 (74.92%) are entirely contained within a single Wikipedia text paragraph, with 241 (37.54%) of them even being equivalent to the Wikipedia text paragraph. 626 (97.51%) text passages are placed within the same Wikipedia section. Although few text passages span across Wikipedia sections, such extraction is less typical.

Text passage titles: For each aligned text passage pair, the users provided titles. These titles can be roughly assigned into three categories as follows:

— Lead section of the article: “Summary”, “Description”, “Definition”, “Name”, “Short biography”.
— Typical sections: “History”, “Family”, “Career”, “Early life”, “Demographics”, “Awards”.
— Article-specific: “Prime minister”, “Ice hockey”, “The accident”, “Bishop of Speyer and Trier”.

We manually categorized the titles given by the users and found that 130 titles (22%) belong to the first, 143 titles (24%) to the second and 324 titles (54%) to the last category. The first category’s titles reoccur very often (“Summary”, ”Description” and ”Definition” are the three most frequent titles) and illustrate the aforementioned observation that the lead paragraphs in each article are likely to be aligned. The titles in the first two categories come from a rather small set of titles and depend on the entity type (e.g. the articles about countries and cities often contain information about demographics). In the last category, there are very specific titles that may be unique for an article and often are more detailed than Wikipedia section titles. In this article we do not perform any automatic labeling of the aligned text passages to annotate the topics. These observations can help to perform automatic labeling in future research.

Disagreement sources: Typical sources of annotator disagreement in our dataset include the differences in the granularity of extracted topics and the varying level of details across the partner articles.

Granularity of extracted topics: The annotators can disagree on the granularity of the topics and the corresponding text passages to be extracted. For example, there is a section about the political career of the Lithuanian politician Algirdas Butkevičius with about ten sentences in each the English and the German articles. One of the annotators performed an alignment of the complete section as a single text passage and entitled it as “Political carrier”. Another annotator performed a higher granularity alignment by splitting it into two text passages entitled “Begin of political carrier”

and "Prime minister"; the third annotator created three text passages: "SDLP[15] Membership", "Minister of Finance" and "Prime Minister".

Level of description details: Another source of disagreement is the case of the interlingual differences between the articles where a very specific description in one article corresponds to a more generic and less detailed description in the partner article. For example, in the English article about the European pine vole, there is only a list of the countries where the animal lives, while the German article has a whole section with a well-phrased text about the animal's habitat. Only one out of three users performed the alignment between the country list and the detailed description.

As certain disagreement is expected in such cases, we treat all user annotations as correct alternatives during the evaluation.

7. EVALUATION OF INTERLINGUAL TEXT PASSAGE ALIGNMENT

To the best of our knowledge, the problem of interlingual text passage alignment presented in this article has not been addressed by any existing approach. To enable an evaluation of the proposed method, we use state-of-the-art methods for text segmentation and alignment that take different approaches to the extraction and alignment aspects as baselines. We evaluate our approach using the user annotated $Align - B$ benchmark presented in Section 6 and experimentally demonstrate the effectiveness of our method and its superiority with respect to the baselines.

7.1. Methods and Baselines

To enable effective interlingual text passage alignment, *MultiWiki* relies on two main components: 1) Extraction of text passages taking their interlingual context into account; and 2) Interlingual alignment of the extracted text passages. In order to evaluate *MultiWiki*, we use baseline methods, each taking a different approach on the extraction and alignment steps:

— **Sentence alignment baseline (*SA Baseline*):** The *SA Baseline* aligns interlingual sentence pairs using a state-of-the-art sentence alignment function defined in [Duh et al. 2013]. In contrast to our method, the *SA Baseline* does not merge aligned sentences into longer text passages. Therefore, a comparison between our method and this baseline can highlight the impact of the text passage extraction on the evaluation metrics. As the sentences pairs aligned by this baseline are syntactically similar, we expect *SA Baseline* to achieve high precision, but at the price of a significant granularity increase compared to our method.

— **Plagiarism detection baseline (*PD Baseline*):** The goal of the plagiarism detection is to identify contiguous maximal-length passages containing reused text [Sanchez-Perez et al. 2015]. While our problem is more general and includes a broader range of semantically similar text passages, topically coherent plagiarism text passages can constitute valid alignments according to our definition. To facilitate a comparison, we use a state-of-the-art plagiarism detection method as a baseline [Sanchez-Perez et al. 2015]. As plagiarism detection can be viewed as a special case of text passage alignment, we expect this baseline to achieve lower recall compared to our method.

— **Alignment of Wikipedia paragraphs (*WikiParagraphs*):** The *WikiParagraphs* method takes the text paragraphs as specified in the original structure of the Wikipedia articles. We perform the alignment of such pre-defined paragraphs equivalently to *MultiWiki* using the same similarity function and thresholds. Compared to the method proposed in this article, in *WikiParagraphs* the boundaries of the text

[15] Social Democratic Party of Lithuania

passage are defined a-priori by the Wikipedia structure and do not take into account any interlingual context. As such paragraphs are user-defined, we expect *WikiParagraphs* to perform well with respect to the granularity of the alignment, as such paragraphs should be intuitive for human readers. However, as *WikiParagraphs* misses text passages deviating from the Wikipedia text paragraphs, we expect to obtain lower recall values. In addition, the interlingual differences in the paragraph structure can affect precision of the alignment.

— **Alignment of TextTiling segments (*TextTiling*):** This baseline represents the TextTiling algorithm [Hearst 1997] that subdivides texts into topically coherent segments using term distributions irrespective of the original text structure. We perform the interlingual alignment of such segments equivalently to the *MultiWiki* and *WikiParagraphs* methods. Similarly to *WikiParagraphs*, the segmentation performed by *TextTiling* does not take the interlingual context of the article into account. Given their comparable approaches, we expect the results of *TextTiling* and *WikiParagraphs* to be similar.

As discussed in the problem statement in Section 2.1, the most important criterion for an effective text passage alignment is to achieve an optimal combination of the precision, recall and granularity of the aligned text passage pairs. Whereas the individual baselines are naturally optimized for one of these metrics, we expect our method to achieve the best performance with respect to their combination.

7.2. Dataset for Interlingual Text Passage Alignment

To facilitate the evaluation of the effectiveness of *MultiWiki* as well as the parameter tuning, we randomly split the German-English part of the $Align{-}B$ benchmark defined in Section 6 into two disjoint parts. The first part, used as a training dataset, contains 20 article pairs. We refer to this part as $Align - T$ and use it for the parameter tuning of *MultiWiki* presented in Section 7.4. The other part, a validation dataset $Align - V$, contains the remaining 35 article pairs and is used to evaluate our approach. The Russian-English subset of the benchmark is named $Align - R$.

Our $Align - B$ benchmark includes article pairs annotated by up to four different users. These user annotations can indicate some differences with respect to the text passage extraction and alignment. It does not appear feasible to build a single alignment incorporating all possible alignments of different users on the same article: even small deviations in the user annotations (e.g. one extra sentence added to a text passage) would lead to an overall different annotation of the article that cannot be directly merged into a single representation without modifying the user-defined alignment. Therefore, in the evaluation, we first compute the scores for each user and each metric separately and then aggregate them to build average scores over the users per article, normalized by the article lengths.

7.3. Evaluation Metrics

The goal of our evaluation is to compare the quality of the text passage alignment using the methods discussed before with respect to their precision, recall and granularity. Although we expect different methods to optimize some of these metrics in isolation, we are particularly interested in measuring their overall impact. According to the problem statement defined in Section 2.1, such evaluation measure must reward high similarity and penalize high granularity of the alignment. In addition, high recall is important to ensure that as many user-aligned text passage pairs are found as possible.

These requirements are to a large extent addressed by the *plagdet* metric as defined in [Potthast et al. 2010]. In the following we discuss this metric and the adjustments

we made to fit this metric (originally defined in the context of the plagiarism detection) our problem, in particular with respect to the granularity computation.

The *plagdet* metric is based on the character-based precision and recall scores as well as on an additional score for granularity. Put together, the three measures form the *plagdet* score that is used as an overall measure of the alignment effectiveness. *Plagdet* requires a set S of the user-defined text passage alignment cases (ground truth) and a set R of their algorithmic detections. In our case, the set S is obtained from the $Align-B$ benchmark. Each method to be evaluated returns its own set R of text passage detections.

In the context of the interlingual text passage alignment, an adjustment is required for the granularity computation. Plagiarism detection is a directed problem: Given a suspicious document and a source document, the goal is to detect reused text parts in the suspicious document. Therefore, in the original *plagdet* metric, the granularity is defined as a measure of "whether a plagiarism case $s \in S$ is detected as a whole or in several pieces" [Potthast et al. 2010]. That means, it would be sufficient if the alignment algorithm would return longer text passages to perform well with respect to this measure. In our case we aim at matching text passages in both articles simultaneously, as close as possible to the user-defined extraction. Therefore, the measure should reflect the granularity on both sides of the alignment. Thus, we define the symmetric granularity measure $gran_{symm}(S, R)$ to be used in our computation of the *plagdet* score:

$$gran_{symm}(S, R) = \frac{gran(S, R) + gran(R, S)}{2}. \tag{6}$$

In more detail, the individual components of *plagdet* are defined as follows:

— **Precision:** The fraction of the characters in R that are among the user-defined text passage pairs. Precision computed at the character level is further normalized using the text passage lengths.
— **Recall:** The fraction of characters in S that are determined by the algorithm.
— **Plagdet:** A combination of the previous scores that rewards a high F value (harmonic mean of precision and recall) and low granularity: $plagdet(S, R) = \frac{F_1}{log_2(1+gran_{symm}(S,R))} \in [0, 1]$.

Granularity and Inverse Granularity: The $gran_{symm}$ measure defined above is anti-correlated with the effectiveness of the method (with $gran_{symm} = 1$ being the best and $gran_{symm} = |R|$ being the worst). To simplify the presentation of the results we will also use $I - Gran(S, R)$ – i.e. its inverse value: $I - Gran(S, R) = gran_{symm}(S, R)^{-1} \in [\frac{1}{|R|}, 1]$. The $I - Gran$ measure is positively correlated with the method effectiveness with respect to the granularity aspect.

7.4. Parameter Tuning

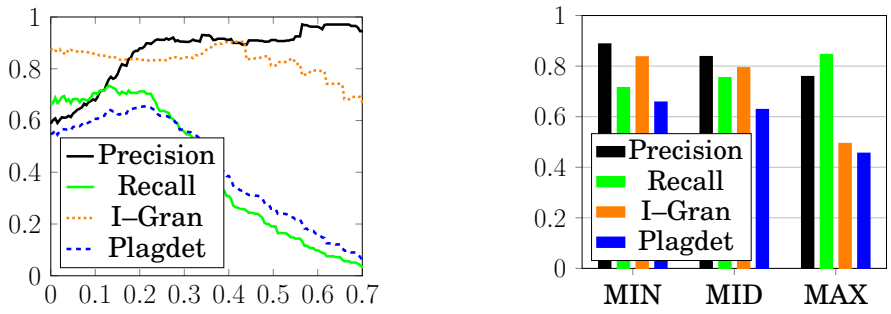
To identify the optimal values for the threshold th and the structure freedom sf parameters, we used the $Align - T$ dataset described in Section 7.2. The best performing values of the parameters on the $Align-T$ are: Similarity threshold $th = 0.21$ and structure freedom $sf = min$ (i.e. the text passage should stay within one Wikipedia text paragraph). Moreover, as we observed that the lead sentences of both articles were aligned in 92.74% of the user aligned articles, we apply a lower similarity threshold (i.e. $th/2$) for the seed text passage pair consisting of the lead sentence per article.

Fig. 3 gives a more detailed overview of how *MultiWiki* behaves for different values of the individual parameters. Obviously, the similarity threshold (Fig. 3a) has a strong

impact on the alignment results. If this threshold is too high, recall decreases as not enough sentence pairs are identified at the beginning of the alignment procedure. The best *plagdet* score for the comparison with the user corpus is achieved when setting $th \approx 0.21$. For very low values of th, recall decreases again: This is because the alignment function initially aligns a large number of sentence pairs, which decreases the effectiveness of the further extraction steps.

With respect to the structure freedom parameter sf (Fig. 3b), we can observe that the best results with respect to the precision, granularity and *plagdet* scores are achieved with the $sf = min$ settings, meaning that the extracted text passages should be contained within one text paragraph. That is consistent with our observation that users do not tend to extract text passages exceeding Wikipedia text paragraphs or sections. The increase in recall value for $sf = max$ accounts for the other cases, where users selected longer text passages.

Overall, we observe that the similarity threshold and the structure freedom parameters are effective to control the alignment results. These parameters allow to put an emphasis on the selected measures such as recall and granularity of the alignment.



(a) Similarity threshold ($sf = min$)

(b) Structure freedom ($th = 0.21$)

Fig. 3: Evaluation scores for different parameters used in our algorithm. In both diagrams, these scores are computed for varying values of the parameter on the X-axis while the other parameter is kept constant.

7.5. Evaluation Results

In this section we first compare the length of text passages extracted by different methods to get insights into their granularity. Then we present the evaluation results of the text passage alignment effectiveness achieved by our method and the baselines presented in Section 7.1.

Text passage length comparison: Table VI provides an overview of the average number of text passage pairs per article alignment and the number of sentences per text passage pair using $Align - T$ and $Align - V$. An optimal alignment should be equivalent to the text passages in the user alignment. As we can observe, the *SA Baseline* that aligns individual sentences contains just two sentences per text passage pair, one in each language. In contrast, the *PD Baseline* that comes from the plagiarism detection domain forms very long text passage pairs sometimes spanning across several Wikipedia sections, containing over 19 sentences on average. Our method *MultiWiki* comes closest to the user alignment (4.43 sentences per text passage pair, vs. 5.00 sentences in the user-defined text passages), which is a good indicator of an appropriate granularity.

Table VI: Number of text passage pairs and sentences aligned by different methods.

	# Text Passage Pairs per Article Pair	# Sentences per Text Passage Pair
User Average	**3.34**	**5.00**
PD Baseline	1.32	19.55
TextTiling	2.16	7.04
WikiParagraphs	2.96	6.64
MultiWiki	4.50	**4.43**
SA Baseline	6.13	2.00

Effectiveness of the alignment methods: The effectiveness results achieved by different methods applied on $Align - V$ and $Align - R$ with respect to the precision, recall, granularity and *plagdet* metrics are shown in Fig. 4.

Fig. 4a presents the precision scores for the five alignment methods applied on the German-English article pairs in $Align - V$. *SA Baseline* and *PD Baseline* achieve precision values of 83.48% and 91.26%, respectively. This is expected, as both of these baselines specifically focus on the syntactic similarity, either by selecting individual sentences (*SA Baseline*), or by extracting plagiarism cases (*PD Baseline*). *TextTiling*, *WikiParagraphs* and *MultiWiki* allow for additional content within the text passages and thus show lower precision values. *WikiParagraphs* that uses predefined Wikipedia paragraphs shows 70.4% precision, which is 8% above *TextTiling*. This number can be significantly improved by enabling flexible extraction in *MultiWiki*, increasing precision to 82.41%.

This flexibility in the text passage extraction enables *MultiWiki* to outperform other methods with respect to the recall metric. The comparison of the recall values is presented in Fig. 4b. Our algorithm *MultiWiki* achieves over 58% recall and outperforms the second best method (*WikiParagraphs*) by 2.6 points for that measure. The *TextTiling* and *PD Baseline* are least flexible, resulting in low recall values of 50.46% and 41.77%. In particular, *PD Baseline* detects too few text passage pairs, whereas *SA Baseline* does not include enough sentences.

Regarding the $I - Gran$ values depicted in Fig. 4c, *MultiWiki* ($I - Gran = 0.86$), *TextTiling* ($I - Gran = 0.83$) and *WikiParagraphs* ($I - Gran = 0.90$) significantly outperform the baseline methods *SA Baseline* ($I - Gran = 0.73$) and *PD Baseline* ($I - Gran = 0.59$). These results confirm the observations made in Table VI where these methods came closest to the user alignment. As *WikiParagraphs* constitutes longer text passages, its $I - Gran$ scores are higher.

The *plagdet* metric in Fig. 4d aggregates the results of precision, recall and granularity. According to this metric, our *MultiWiki* method performs best and achieves the highest *plagdet* score of 0.56, that is 0.03 points better than *WikiParagraphs*. *MultiWiki* outperforms the *SA Baseline* by 0.09 and *PD Baseline* by 0.26 points. The results of the paired t-test confirm statistical significance of this result for the confidence level of 95%.

Overall, our evaluation results confirm the high effectiveness of our *MultiWiki* method. *MultiWiki* achieves the highest *plagdet* and recall scores and outperforms the baselines with respect to granularity due to its flexibility in the extraction process. This result also demonstrates that existing approaches like *SA Baseline* and *PD Baseline* that optimize for syntactic similarity cannot be effectively applied to the problem of interlingual text passage alignment presented in this article. When using the predefined paragraphs for the alignment, *WikiParagraphs* outperforms *TextTiling* in every

aspect. Hence, the text paragraphs in Wikipedia represent a more intuitive division of the article into its subtopics from the user perspective than the *TextTiling* method.

English-Russian dataset: To confirm the generalizability of our approach on other language pairs, we evaluated the alignment methods on the Russian-English article pairs in $Align - R$, using the same values for the similarity functions weights and passage alignment parameters as in the English-German case. The results follow a similar distribution compared to the English-German evaluation. *MultiWiki* achieves a *plagdet* score of 0.63, outperforming the other methods as seen in Fig. 4h. As we did not perform any language-specific training on the Russian-English data, these evaluation results suggest that the training results obtained in one language can be effectively applicable to the sentence similarity computation and text passage alignment in other language pairs.

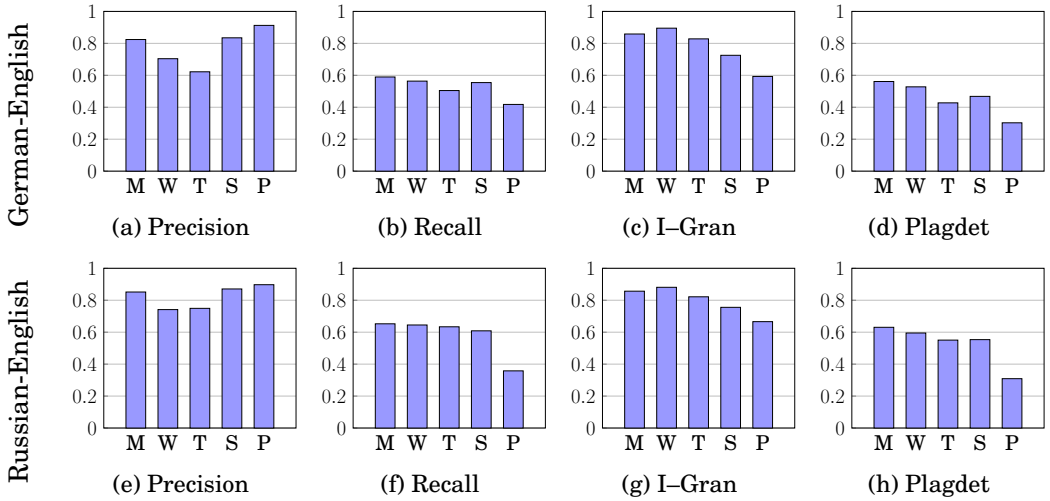


Fig. 4: Evaluation metric scores for two language pairs, different metrics, alignment methods and baselines. M: *MultiWiki*, W: *WikiParagraphs*, T: *TextTiling*, S: *SA Baseline*, P: *PD Baseline*.

8. RELATED RESEARCH

The problem of the interlingual text passage alignment in partner articles discussed in this paper has not been addressed by any existing approach. In the following we discuss related applications to analyze interlingual differences in multilingual Wikipedia as well as related methods for interlingual text alignment. Finally, we discuss available benchmarks.

Analyzing interlingual differences in Wikipedia: The problem of identifying information missing in a particular language edition using other Wikipedia language editions has been considered at different levels of granularity, including suggestion of the articles missing in a particular language edition to the Wikipedia editors [Wulczyn et al. 2016], finding complementary sentences within a partner article [Duh et al. 2013] and detection of missing infobox information [Adar et al. 2009]. Information propagation across languages has been considered by Hale [2014], who studied the behavior of the editors simultaneously working on multiple Wikipedia language editions. Interlingual information propagation has also been considered in our recent demonstration

paper, where we proposed a graphical user interface to observe changes in the interlingual article similarity over time [Gottschalk and Demidova 2016]. All these approaches target discovery of interlingual similarities in Wikipedia, while targeting aspects different from MultiWiki.

Further approaches attempt to automatically compare partner articles to identify their overall similarity. In Barrón-Cedeño et al. [2014], the authors compare different metrics to compute an overall similarity of the articles in different languages. Manypedia [Massa and Scrinzi 2012] provides an automatic translation to English, and points out article statistics and concept similarity metrics computed based on the article interlinking. The Omnipedia interface [Bao et al. 2012] visualizes the information summarized from multiple language editions using topic extraction methods. However, none of the existing approaches enables the detailed interlingual comparison of the partner articles at the text passage level as facilitated by MultiWiki.

Interlingual text passage alignment: Text passage alignment has been considered in the context of machine translation applications, where the goal is to create parallel corpora to train translation models. In this context, existing approaches aim to extract parallel text passages from a bilingual parallel document corpus (e.g. [Rasooli et al. 2011], [Gupta and Pala 2012]). Existing approaches in this area typically assume that the paragraphs are parallel and contain translated text, such that adjustments of the paragraph boundaries are not required. Rasooli et al. [2011] use predefined paragraph boundaries and apply similarity measures, similar to the method *WikiParagraphs* used as a baseline in this paper. Gupta et al. [2012] allow for many-to-many paragraph alignments (i.e. they merge neighbored paragraphs) based on the assumption of a common text flow in both documents. These assumptions are not applicable to the partner articles in Wikipedia. In contrast, the MultiWiki method does not require any parallel corpora and facilitates an alignment of similar text passages irrespective of the differences in the paragraph structure.

At a higher granularity level, several works have also considered an alignment of individual parallel sentences in the context of machine translation [Adafre and De Rijke 2006], [Mohammadi and Ghasem-Aghaee 2010] and identification of complementary sentences in partner articles [Duh et al. 2013]. Sentence alignment alone fails to provide an overview of the overlapping article parts due to its high granularity. In addition, in MultiWiki we face the problem of the alignment of partially overlapping text, that can contain intermediate unrelated sentence parts or entire sentences. The use of semantic features enables MultiWiki to overcome the limitations related to the syntactic alignment of parallel sentences and achieve better recall and granularity of the alignment, as demonstrated by our experimental evaluation.

Interlingual text reuse and plagiarism detection: Text reuse occurs for various reasons and can be of different granularity, including reuse of entire documents as well as extracts thereof, such as sentences, facts or text passages (also known as local text reuse [Seo and Croft 2008]). In this context, plagiarism detection is a special form of local text reuse detection. Interlingual plagiarism detection focuses on identification of reused text passages across a suspicious document and possible source documents. In the first step, plagiarism detection methods try to identify source documents for a given suspicious input document. Then, they search for text fragments that are found both in the suspicious document and the source documents [Alzahrani et al. 2010].

There are two different approaches to identify plagiarized text passages: (a) Subdividing the text into sections to build a tree structure of the document that is used to reduce the number of comparisons [Chow and Rahman 2009]; and (b) Bottom-up combination of sentences into text passages [Alzahrani et al. 2010]. While (a) relies on similar passage partitioning between the texts, (b) highly relies on correctly aligned sentences: In [Alzahrani et al. 2010], they merge aligned sentences that have a dis-

tance of at most 10 characters. This is similar to the merging of neighbored text passage pairs in our method, but lacks the inclusion of the intermediate sentences in an aligned text passage and there is no similarity re-computation between text passages consisting of more than one sentence.

There are two important differences between plagiarism detection and the problem of finding similar text passages across Wikipedia article pairs: First, plagiarism detection is a directed problem: Given a suspicious document and one or more source documents, the goal is to search for text passages in the suspicious document that are based on the source documents. In the Wikipedia text passage alignment, there is no direction: Due to the independent evolution of Wikipedia articles in different language editions, both partner articles assume the role of a suspicious document and a source document at a time. Second, a plagiarized text passage must be based on a text passage in another document. Therefore, many plagiarism detection methods use syntactic similarity measures like n-grams or the longest common subsequence [Alzahrani et al. 2010] that take the order of words or characters into account. In our case, aligned text passages share common information without necessarily being based on the same source. Because of these differences, our MultiWiki method allows for larger portions of additional information such as unaligned sentences or facts within the aligned text passages.

Benchmarks: Existing parallel corpora (e.g. [Koehn 2005], [Steinberger et al. 2006]) cover parallel sentences from particular domains (e.g. news domain or parliamentary proceedings). Smith et al. [2010] provided a small benchmark with 225 parallel sentences extracted from 20 manually chosen Wikipedia articles. However, existing corpora focus on parallel sentences and do not include the sentences with partially overlapping information nuggets. SemEval workshop on semantic evaluation [Agirre et al. 2016] includes a cross-lingual semantic textual similarity task and provides Spanish-English bilingual sentence pairs. The sentences within this benchmark come from the domains different from Wikipedia, are rather short and rarely contain time information, as opposed to our $Sim - B$ benchmark. In this work we incrementally build a benchmark $Sim - B$ for the alignment of sentences with substantial semantic overlap for the German and the English Wikipedia. Furthermore, we collect a user-annotated benchmark $Align - B$ for text passage extraction and alignment from partner articles. We make both benchmarks available to facilitate further research in this area.

9. DISCUSSION AND FUTURE WORK

In this article we tackled the problem of interlingual alignment of semantically similar text passages across partner articles in Wikipedia. Partner articles evolve independently in different language editions and can therefore reflect community-specific points of view on particular topics, or indicate other differences with respect to the content, structure and quality of the information they contain. MultiWiki is the first method that facilitates direct comparison of the similarities and differences in the interlingual partner article pairs at the text passage level, providing users with a detailed overview.

Contributions of the article: In order to facilitate a comprehensive overview of the interlingual similarities and differences in an article pair, we defined text passage alignment as an optimization problem. This problem maximizes the semantic similarity across the aligned text passages while reducing their granularity. This way, we aim at obtaining possibly long and semantically similar interlingual text passage pairs. Then, we designed a method to address this optimization problem and defined a semantic similarity measure for the interlingual text passage alignment along with a greedy algorithm to perform text passage extraction. A further contribution of this

article are the user-annotated benchmarks containing aligned text passages from the German, Russian and the English Wikipedia language editions. Our evaluation results on German-English and Russian-English article pairs demonstrate that our method achieves a good balance between precision, recall and granularity of the aligned text passages as measured against the user annotations.

Extensions to other language pairs: The MultiWiki system is publicly available[16] and its demonstration currently supports four language pairs: German-English, Dutch-English, Portuguese-English and Russian-English. The set of the language pairs supported by MultiWiki is extendible as long as the minimal requirements on the availability of the language processing tools are satisfied. This includes sentence splitting, tokenization and machine translation for the corresponding language pair. As machine translation is used to obtain the term vector representations of text passages, the only requirement on the translation quality is the correct translation for the majority of the terms. As we observed, entity annotations and time annotations can further increase the precision of the interlingual text passage alignment. Annotation tools such as DBpedia Spotlight and HeidelTime are already available in a number of languages making it possible to further extend the number of languages supported by MultiWiki in the future. Another interesting direction for future research is the reduction of the need for machine translation while aligning multilingual text by the development of further text similarity features, e.g. by utilizing multilingual word embeddings [Vulić and Moens 2015] or language-independent word sense disambiguation [Pilehvar et al. 2013], [Moro et al. 2014].

Domain adaptation: In this article we focused on the interlingual text passage alignment in Wikipedia. Using this corpus, we can utilize its specific features such as the interlingual links between partner articles, the comparable text styles and the encyclopedic nature of the articles which enables to extract a relatively high number of semantic annotations. In our future research we would like to consider an adaptation of this approach to other domains, such as multilingual news and social media, where these features may not be available to the same extent. Adaptation to these domains may require establishing interlingual links at the article level, as well as adaptation of the similarity function and alignment algorithms to better match the features and text structure in these domains.

Exclusiveness in the alignment model: In our problem statement we assume the mutual exclusiveness of text passages within the article as well as with respect to the alignment. The intuition behind this assumption is that such exclusive alignment can facilitate a better overview of an article pair and avoid overlaps across text passages and alignments, as such overlaps would contradict the overview goal. Note that the interlingual alignment also shapes the text passages, i.e. the borders of the aligned text passages are mutually dependent and determined during the alignment process to increase the interlingual similarity. In practice, it is possible that an article contains multiple alternatives for an alignment, from which we select only the best matching one in these settings. For example, the information from the first English sentence in Fig. 1 (the construction of the bridge and the official opening) is spread across two different German sentences. This information could be aligned under a different problem setting where one would focus on a particular predefined text passage and find all possible matching text passages in the other article. Such problem variation would require an adaptation of the alignment algorithms and is an interesting extension for future work.

Consecutiveness in the alignment model: In our problem statement we also require the consecutiveness of sentences in a text passage. This modeling decision is taken in

[16] https://github.com/sgottsch/multiwiki

favor of providing an overview of the overlapping parts and putting complementary information in context. The structure of the resulting text passages is shaped by the fragmentation of similar information in both languages. Our model does not require that each sentence in a text passage has a 1:1 correspondence in the alignment, such that aligned text passages can contain partially overlapping sentences or intermediate sentences with no correspondence. Consequently, when building pairs of consecutive text passages, sentences on one side can contain complementary (or sometimes contradictory information) and deliver its language-specific context. An interesting direction for future research is to develop interlingual Information Extraction methods that would allow precise identification of the corresponding and additional information nuggets within the aligned text passages.

Cultural studies and Applications: The MultiWiki text passage alignment method presented in this article can facilitate and support cross-lingual case studies. In the context of cultural studies like [Rogers 2013] the proposed text passage alignment approach can reduce the amount of information that needs to be manually analyzed by researchers and enables researches to focus on the essentially overlapping article parts. A case study with the digital humanities researchers utilizing the system for cultural analytics is an interesting extension for our future work. The cross-lingual text passage alignment can also enhance a wide range of interlingual applications that use Wikipedia as an information source. Example applications that use interlingual Wikipedia content include multilingual summarization [Baralis et al. 2015] and cross-lingual text classification [Ni et al. 2011]. Our method can provide a more precise context for such applications through the targeted alignment of the most relevant text passages. For another example, in our recent demo paper [Gottschalk and Demidova 2016] we presented a novel graphical user interface to analyze temporal evolution of partner articles. This tool uses the interlingual text passage alignment presented in this article to facilitate a detailed visual article comparison. As these examples illustrate, interlingual text passage alignment methods developed in this article have a great potential to facilitate cultural studies and spawn a variety of novel interlingual applications.

REFERENCES

Sisay Fissaha Adafre and Maarten De Rijke. 2006. Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL '06)*. 62–69.

Eytan Adar, Michael Skinner, and Daniel S. Weld. 2009. Information Arbitrage Across Multi-lingual Wikipedia. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM '09)*. ACM, New York, NY, USA, 94–103. DOI:http://dx.doi.org/10.1145/1498759.1498813

Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. 497–511.

Salha Alzahrani, Naomie Salim, Chow Kok Kent, Mohammed Salem Binwahlan, and Ladda Suanmali. 2010. The Development of Cross-Language Plagiarism Detection Tool Utilising Fuzzy Swarm-Based Summarisation. In *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA '10)*. Cairo, Egypt, 86–90. DOI:http://dx.doi.org/10.1109/ISDA.2010.5687287

Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. 2012. Omnipedia: Bridging the Wikipedia Language Gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1075–1084. DOI:http://dx.doi.org/10.1145/2207676.2208553

Elena Baralis, Luca Cagliero, Alessandro Fiori, and Paolo Garza. 2015. MWI-Sum: A Multilingual Summarizer Based on Frequent Weighted Itemsets. *ACM Transactions on Information Systems (TOIS)* 34, 1, Article 5 (Sept. 2015), 35 pages. DOI:http://dx.doi.org/10.1145/2809786

Alberto Barrón-Cedeño, Monica Lestari Paramita, Paul Clough, and Paolo Rosso. 2014. A Comparison of Approaches for Measuring Cross-Lingual Similarity of Wikipedia Articles. In *Advances in Information Retrieval*. DOI:http://dx.doi.org/10.1007/978-3-319-06028-6_36

Tommy W. S. Chow and M. K. M. Rahman. 2009. Multilayer SOM With Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection. *IEEE Transactions on Neural Networks* (2009). DOI:http://dx.doi.org/10.1109/TNN.2009.2023394

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2013. Accurate Parallel Fragment Extraction from Quasi-Comparable Corpora using Alignment Model and Translation Lexicon. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP '13)*. Nagoya, Japan, 1144–1150.

Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 659–666. DOI:http://dx.doi.org/10.1145/1390334.1390446

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-SEMANTICS '13)*. ACM, New York, NY, USA, 121–124. DOI:http://dx.doi.org/10.1145/2506182.2506198

Kevin Duh, Ching-Man Au Yeung, Tomoharu Iwata, and Masaaki Nagata. 2013. Managing Information Disparity in Multilingual Document Collections. *ACM Trans. Speech Lang. Process.* 10, 1, Article 1 (March 2013), 28 pages. DOI:http://dx.doi.org/10.1145/2442076.2442077

Manaal Faruqui and Shankar Kumar. 2015. Multilingual Open Relation Extraction Using Cross-lingual Projection.. In *HLT-NAACL*, Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar (Eds.). The Association for Computational Linguistics, 1351–1356.

Elena Filatova. 2009. Directions for Exploiting Asymmetries in Multilingual Wikipedia. In *Proceedings of the 3rd International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3 '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 30–37.

Simon Gottschalk and Elena Demidova. 2016. Analysing Temporal Evolution of Interlingual Wikipedia Article Pairs. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 1089–1092. DOI:http://dx.doi.org/10.1145/2911451.2911472

Ankush Gupta and Kiran Pala. 2012. A Generic and Robust Algorithm for Paragraph Alignment and its Impact on Sentence Alignment in Parallel Corpora. In *Workshop on Indian Language and Data: Resources and Evaluation (WILDRE '12)*. 18–27.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Scott A. Hale. 2014. Multilinguals and Wikipedia Editing. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci '14)*. ACM, New York, NY, USA, 99–108. DOI:http://dx.doi.org/10.1145/2615569.2615684

Marti A. Hearst. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Comput. Linguist.* 23, 1 (March 1997), 33–64.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit (MT Summit '05)*. AAMT, AAMT, Phuket, Thailand, 79–86.

J. R. Landis and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Paolo Massa and Federico Scrinzi. 2012. Manypedia: Comparing Language Points of View of Wikipedia Communities. In *Proceedings of the 8th International Symposium on Wikis and Open Collaboration (WikiSym '12)*. ACM, New York, NY, USA, Article 21, 9 pages. DOI:http://dx.doi.org/10.1145/2462932.2462960

Mehdi Mohammadi and Nasser Ghasem-Aghaee. 2010. Building Bilingual Parallel Corpora Based on Wikipedia. In *Proceedings of the 2010 Second International Conference on Computer Engineering and Applications - Volume 02 (ICCEA '10)*. IEEE Computer Society, Washington, DC, USA, 264–268. DOI:http://dx.doi.org/10.1109/ICCEA.2010.203

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)* 2 (2014), 231–244.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2011. Cross Lingual Text Classification by Mining Multilingual Topics from Wikipedia. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 375–384. DOI:http://dx.doi.org/10.1145/1935826.1935887

Monica Lestari Paramita, Paul D. Clough, Ahmet Aker, and Robert J. Gaizauskas. 2012. Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey, 790–797.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL '13)*. Sofia, Bulgaria, 1341–1351.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 997–1005.

Mohammad Sadegh Rasooli, Omid Kashefi, and Behrouz Minaei-Bidgoli. 2011. Extracting Parallel Paragraphs and Sentences from English-persian Translated Documents. In *Proceedings of the 7th Asia Conference on Information Retrieval Technology (AIRS'11)*. Springer-Verlag, Berlin, Heidelberg, 574–583. DOI:http://dx.doi.org/10.1007/978-3-642-25631-8_52

Richard Rogers. 2013. *Digital Methods*. The MIT Press, Chapter Wikipedia as Cultural Reference.

Miguel A. Sanchez-Perez, Alexander Gelbukh, and Grigori Sidorov. 2015. Adaptive Algorithm for Plagiarism Detection: The Best-Performing Approach at PAN 2014 Text Alignment Competition. In *Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction - Volume 9283 (CLEF'15)*. Springer-Verlag New York, Inc., New York, NY, USA, 402–413. DOI:http://dx.doi.org/10.1007/978-3-319-24027-5_42

Jangwon Seo and W. Bruce Croft. 2008. Local Text Reuse Detection. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 571–578. DOI:http://dx.doi.org/10.1145/1390334.1390432

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora Using Document Level Alignment. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 403–411.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Toma Erjavec, and Dan Tufi. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*. 2142–2147.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation* 47, 2 (2013), 269–298. DOI:http://dx.doi.org/10.1007/s10579-012-9179-y

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 363–372. DOI:http://dx.doi.org/10.1145/2766462.2767752

Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. 2016. Growing Wikipedia Across Languages via Recommendation. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 975–985. DOI:http://dx.doi.org/10.1145/2872427.2883077

Taha Yasseri, Anselm Spoerri, Mark Graham, and Janos Kertesz. 2014. The most controversial topics in Wikipedia: A multilingual and geographical analysis. In *Global Wikipedia:International and cross-cultural issues in online collaboration*. Scarecrow Press.

Received July 2016; revised November 2016; accepted November 2016