# Multilingual Event Linking to Wikidata

**Anonymous ACL submission**

## Abstract

We present a task of multilingual linking of events to a knowledge base. We automatically compile a large-scale dataset for this task, comprising of 1.8M mentions across 44 languages referring to over 10.9K events from Wikidata. We propose two variants of the event linking task: 1) multilingual, where event descriptions are from the same language as the mention, and 2) crosslingual, where all event descriptions are in English. On the two proposed tasks, we compare multiple event linking systems including BM25+ (Lv and Zhai, 2011a) and multilingual adaptations of the biencoder and crossencoder architectures from BLINK (Wu et al., 2020). In our experiments on the two task variants, we find both biencoder and crossencoder models significantly outperform the BM25+ baseline. Our results also indicate that the crosslingual task is in general more challenging than the multilingual task. We also present a qualitative analysis highlighting various aspects captured by the proposed dataset, including the need for temporal reasoning over context and tackling diverse event descriptions across languages.[1]

## 1 Introduction

Grounding references of entities and events to a knowledge base (KB) is an important component of information extraction systems. This is a well-studied task, especially for linking entity references to KBs like Wikipedia (Ji and Grishman, 2011). In this work, we present a new multilingual task that involves linking *event* references to Wikidata KB.[2]

Event linking differs from entity's as it involves taking into account the event participants as well as its temporal and spatial attributes. Nothman et al. (2012) defines event linking as connecting event references from news articles to a news archive consisting of first reports of the events. Similar to entities, event linking is typically restricted to prominent or report-worthy events. In this work, we use a subset of Wikidata as our event KB and link mentions from Wikipedia articles to these events.[3] Figure 1 illustrates our event linking methodology.

Event linking is closely related to the more commonly studied task of cross-document event coreference (CDEC). In CDEC, the goal is to understand the identity relationship between event mentions. This identity is often complicated by subevent and membership relations among events (Pratapa et al., 2021). Nothman et al. (2012) proposed event linking as an alternative to coreference that helps ground report-worthy events to a KB. They showed that the linking tasks help avoid the traditional bottlenecks seen with the event coreference task. We postulate *linking to be a complementary task to coreference*, where the first mention of an event in a document is typically linked or grounded to the KB and its relationship with the rest of the mentions from the document is captured via coreference.

For the event linking task, we present a new multilingual dataset that grounds mentions from multilingual Wikipedia articles to the corresponding event in Wikidata. Figure 1 presents an example from our dataset that links mentions from three languages to the same Wikidata item. To construct this dataset, we make use of the hyperlinks in Wikipedia articles. These links connect anchor texts (like '2010 European Championships' or "Championnats d'Europe") in context to the corresponding event Wikipedia page ('2010 European Aquatics Championships' or "Championnats d'Europe de natation 2010"). We further connect the event Wikipedia page to its Wikidata item ('Q830917'), facilitating multilingual grounding of mentions to KB events. We use the title and first paragraph from the language Wikipedia pages as our event descriptions (column 2 in Figure 1).

Prior works made use of these types of

[1] Code and data will be released upon publication.

[2] `www.wikidata.org`

[3] We define *mention* as the textual expression that refers to an *event* from the KB.

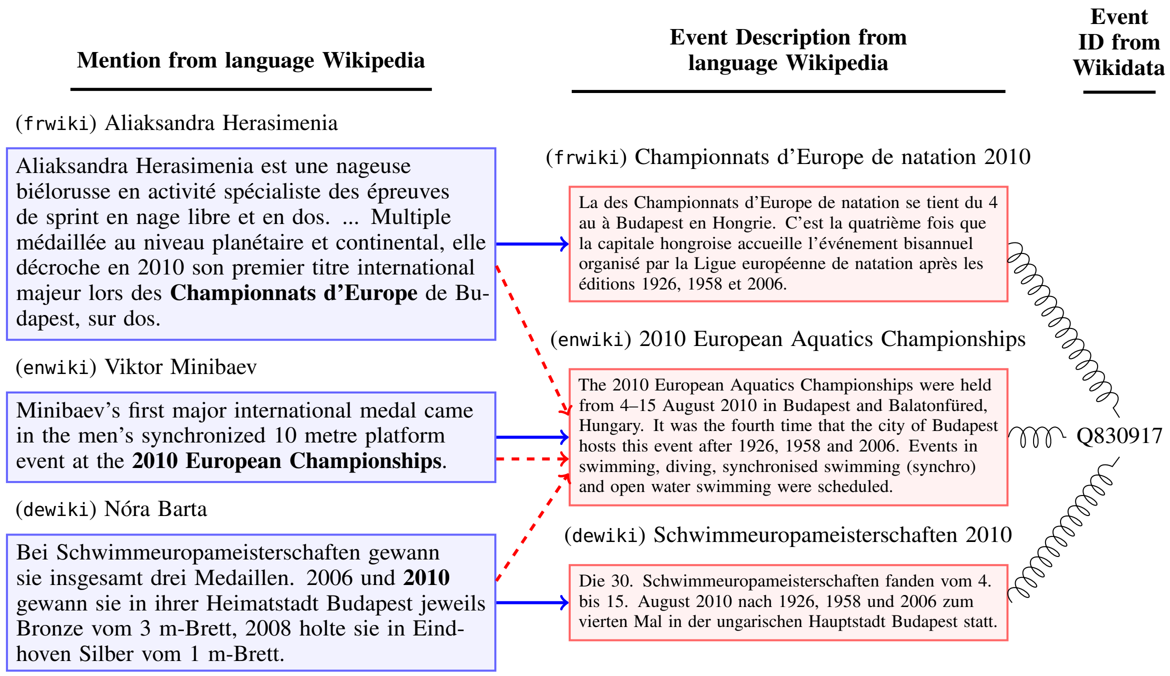


Figure 1: An illustration of multilingual event linking with Wikidata as our interlingua. Mentions from French, English and German Wikipedia (column 1) are linked to the same event from Wikidata (column 3). The title and descriptions for the event Q830917 are compiled from the corresponding language Wikipedias (column 2). The solid blue arrows (⟶) presents our multilingual task, to link `lgwiki` mention to event using `lgwiki` description. The dashed red arrows (- - ›) showcases the crosslingual task, to link `lgwiki` mention to event using `enwiki` description.

Wikipedia hyperlinks for named entity disambiguation (Eshel et al., 2017), entity linking (Logan et al., 2019) and cross-document coreference of events (Eirew et al., 2021) and entities (Singh et al., 2012). Our work is closely related to the English CDEC work of Eirew et al. (2021), but we view the task as linking instead of coreference. This is primarily due to the fact that most hyperlinked event mentions are prominent and typically cover a broad range of subevents, conflicting directly with the notion of coreference. Additionally, our dataset is multilingual, covering 44 languages with Wikidata serving as our *interlingua*. Botha et al. (2020) is another related work from entity linking literature that covers entity references from multilingual Wikipedia articles to Wikidata.

We use the proposed dataset to develop multilingual event linking systems. We present two variants to the linking task, multilingual and crosslingual. In the multilingual setup, mentions from Wikipedia of any language are linked to the events from Wikidata with descriptions taken from the same language (see solid blue arrows (⟶) in Figure 1). In contrast, the crosslingual setup requires that the systems make use of the English event description irrespective of the source mention language (see dashed red arrows (- - ›) in Figure 1). In both tasks, the end goal is to identify the Wikidata ID (e.g. Q830917). Following prior work on entity linking (Logeswaran et al., 2019), we adopt a *zero-shot event linking* approach in all of our experiments. We present results using a retrieve+rank approach based on Wu et al. (2020) that utilizes BERT-based biencoder and crossencoder for our multilingual event linking task. We experiment with two multilingual encoders, mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). We find that both biencoder and crossencoder significantly outperform a tf-idf-based baseline, BM25+ (Lv and Zhai, 2011a). Our results indicate the crosslingual task is more challenging than the multilingual task, possibly due to differences in typology of source and target languages. Our key contributions are,

- We propose a new multilingual NLP task that involves linking multilingual text mentions to a knowledge base of events.
- We release a large-scale dataset for the zero-shot multilingual event linking task by compiling mentions from Wikipedia and their grounding to Wikidata. Our dataset captures 1.8M mentions across 44 languages refering to over 10K events.

- We present two evaluation setups, multilingual and crosslingual event linking. We show competitive results across languages using a retrieve and rank methodology.

In section 2, we posit our work in the existing literature on event/entity linking and coreference. We then describe our dataset collection process (section 3), followed by the proposed multilingual event linking systems in section 4. Finally, we present an evaluation and analysis of the linking systems in section 5 and potential directions for future work in section 6.

## 2 Related Work

Our focus task of multilingual event linking shares resemblance with entity/event linking, entity/event coreference and other multilingual NLP tasks. We present a summary of these connections below.

**Entity Linking:** Our work utilizes hyperlinks between Wikipedia pages to identify event references. This idea was previously explored in multiple entity related works, both for dataset creation (Mihalcea and Csomai, 2007; Botha et al., 2020) and data augmentation during training (Bunescu and Paşca, 2006; Nothman et al., 2008). Another related line of work utilized hyperlinks from general web pages to Wikipedia articles for the tasks of cross-document entity coreference (Singh et al., 2012) and named entity disambiguation (Eshel et al., 2017). Prior work on entity linking has highlighted the need for zero-shot evaluation of systems (Sil et al., 2012; Logeswaran et al., 2019). We adopt this standard in our experiments by using a disjoint sets of events for training and evaluation (see subsection 3.2).

**Event Linking and Coreference:** Linking event expressions to a KB is important for downstream tasks like narrative understanding. For instance, consider a prominent event like '2020 Summer Olympics'. This event has had a large influx of articles in multiple languages. Its often useful to ground the references to specific prominents subevents in KB. Some examples of such events from Wikidata are "Swimming at the 2020 Summer Olympics – Women's 100 metre freestyle" (Q64513990) and "Swimming at the 2020 Summer Olympics – Men's 100 metre backstroke" (Q64514005). Event linking task while important is albeit less explored. Nothman et al. (2012) linked event-referring expressions from news articles to a news archive. These links are made to the first-reported news article regarding the event. In contrast, we focus on events that are prominent enough to warrant a Wikipedia article, a Wikidata item.

Event coreference resolution is closely related to event grounding but assumes a stricter notion of identity between mentions (Nothman et al., 2012). Multiple cross-document coreference resolution works made use of Wikipedia (Eirew et al., 2021) and Wikinews (Minard et al., 2016; Pratapa et al., 2021) for dataset collection. Minard et al. (2016) obtained human translations of English Wikinews articles to create a crosslingual event coreference dataset. In contrast, our dataset uses the original multilingual event descriptions written by language Wikipedia contributors (column 2 in Figure 1).

**Multilingual Tasks:** A majority of the existing NLP datasets (and systems) cater to a small set of world languages (Joshi et al., 2020). There is a growing effort on creating more multilingual datasets and benchmarks. This includes efforts for tasks like natural language inference (XNLI; Conneau et al. (2018)), question answering (TyDi-QA; Clark et al. (2020), XOR QA; Asai et al. (2021)), linking (Mewsli-9; Botha et al. (2020)) as well as comprehensive benchmarks (XTREME-R; Ruder et al. (2021)). To the best of our knowledge, our work presents the first benchmark for the multilingual event linking task.

## 3 Multilingual Event Linking Dataset

Our data collection methodology is closely related to the zero shot entity linking work of Botha et al. (2020) but we take a top-down approach starting from Wikidata. Eirew et al. (2021) identified event pages from English Wikipedia by processing the infobox elements. However, we found relying on Wikidata for event identification to be more robust. Additionally, Wikidata serves as our *interlingua* that connects mentions from numerous languages.

### 3.1 Dataset Compilation

To compile our dataset, we follow a simple three-stage pipeline. First, we identify Wikidata items that correspond to events. Second, for these Wikidata events, we collect links to corresponding language Wikipedia articles. Third, we iterate through all the language Wikipedia dumps to collect mentions that refer to these events.

| | Train | Dev | Test | Total |
|---|---|---|---|---|
| Events | 8653 | 1090 | 1204 | 10947 |
| Event Sequences | 6758 | 844 | 846 | 8448 |
| Mentions | 1.44M | 165K | 190K | 1.8M |
| Languages | 44 | 44 | 44 | 44 |

Table 1: Dataset Summary

2016 Summer Olympics (Q8613)

$\uparrow$ part-of

athletics at the 2016 Summer Olympics (Q18193712)

$\uparrow$ part-of

athletics at the 2016 Summer Olympics–men's 100 metres (Q25397537)

Figure 2: An illustration of event hierarchy in Wikidata.

⋮

2008 European Aquatics Championships (Q643842)

$\downarrow$ followed-by

2010 European Aquatics Championships (Q830917)

$\downarrow$ followed-by

2012 European Aquatics Championships (Q941108)

⋮

Figure 3: An illustration of event sequences in Wikidata.

**Wikidata Event Identification:** Events typically have specific time, location and participants associated with them, distinguishing them from entities. To identify events from Wikidata (WD), we make use of the properties listed for each WD item.[4] Specifically, we consider a WD item to be a candidate event if it contains the following two properties, temporal (duration OR point-in-time OR (start-time AND end-time)) and spatial (location OR coordinate-location). We perform additional post-processing on these candidate event set to remove non-events like empires (Roman Empire: Q2277), missions (Surveyor 7: Q774594), TV series (Deception: Q30180283) and historic places (French North Africa: Q352061).[5] The final set of Wikidata events are prominent or Wiki-worthy events and typically constitute nominal events. The notion of Wiki-worthy events is similar to the concept of news-worthy events from newswire (Nothman et al., 2012; Upadhyay et al., 2016).

**Tackling WD Hierarchy:** WD is a rich structured KB and we observed many instances of hierarchical relationship between our candidate events. See Figure 2 for an example. While this hierarchy adds an interesting challenge to the event grounding task, we observed multiple instances of inconsistency in links. Specifically, we observed references to parent item (Q18193712) even though the child item (Q25397537) was the most appropriate link in context. Therefore, in our dataset, we only include *leaf nodes* as our candidate event set (e.g. Q25397537). This allows us to focus on most atomic events from Wikidata. Expanding the label set to include the hierarchy is an interesting direction for future work.

**Wikidata — Wikipedia:** WD items have pointers to the corresponding language Wikipedia articles.[6] We make use of these pointers to identify Wikipedia articles describing our candidate WD events. Figure 1 illustrates this through the coiled pointers ( ) for the three languages. We make use of the event's Wikipedia article title and its first paragraph as the description for the WD event. Each language version of a Wikipedia article is typically written by independent contributors, so the event descriptions vary across languages.

**Mention Identification:** As mentioned earlier, Wikipedia articles are often connected through hyperlinks. We iterate through each language Wikipedia and collect anchor texts of hyperlinks to the event Wikipedia pages (column 1 in Figure 1). In addition to the anchor text, we also keep the surrounding paragraph as the context. Notably, in multiple occasions, the anchor text can be a temporal expression or location relevant to the event. In the German mention from Figure 1, the anchor text '2010' links to the event Q830917 (2010 European Aquatics Championships). This link can be infered by using the neigboring context ('Schwimmeuropameisterschaften': European Aquatics Championships). In fact, the other span '2006' in the same example links to a different event from Wikidata, Q612454 (2006 European Aquatics Championships). We use the 2021-09-01 XML dumps of language Wikipedias and the 2021-10-04 JSON dump of Wikidata. We use `wikiextractor` tool (Attardi, 2015) to extract text content from the XML Wikipedia dumps. We retain the hyperlinks from the articles for use in mention identification.

**Postprocessing:** To link a mention to its event,

[4] wikidata.org/wiki/Wikidata:List_of_properties

[5] see Table 5 in subsection A.2 of Appendix for the full list of exclusion properties.

[6] https://meta.wikimedia.org/wiki/List_of_Wikipedias

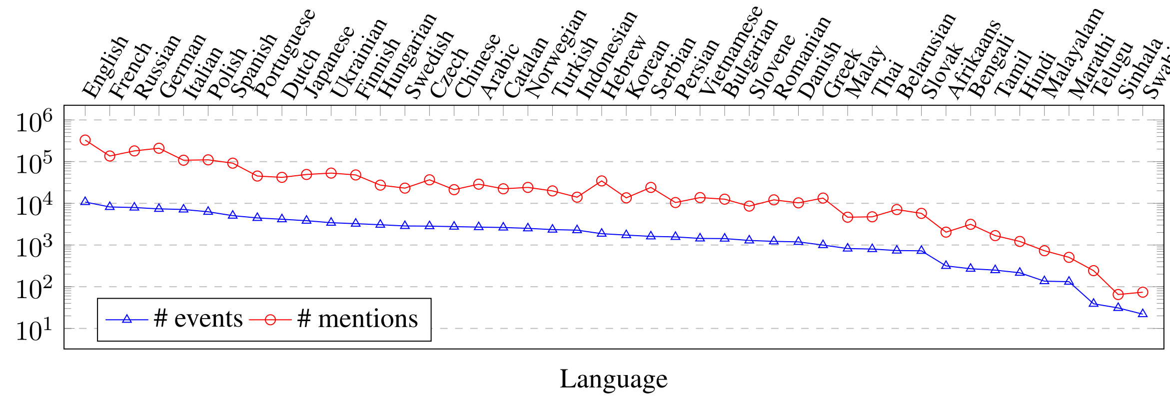


Figure 4: Statistics of events and mentions per language in the proposed dataset. The languages are sorted in the decreasing order of # events. The counts on y-axis are presented in log scale.

the context should contain the necessary temporal information. For instance, its important to be able to differentiate between links to '2010 European Aquatics Championships' vs '2012 European Aquatics Championships'. Therefore, we heuristically remove mention (+context) if it completely misses the temporal expressions from the corresponding language Wikipedia title and description. Additionally, we also remove mentions if their contexts are either too short or too long (<100, >2000 characters). We also prune WD events under the following conditions: 1) only contains mentions from a single language, 2) >50% of the mentions match their corresponding language Wikipedia title (i.e., low diversity), 3) very few mentions (<30). Table 1 presents the overall statistics of our dataset. The full list of languages with their event and mention counts are presented in Figure 4. Each WD event on average has mention references from 9 languages indicating the highly multilingual nature of our dataset. See Table 6 in Appendix for details on the geneological information for the chosen languages. We chose our final set of languages by maximizing for the diversity in language typology, language resources (in event-related tasks and general) and the availability of content on Wikipedia. Wikipedia texts and Wikidata KB are available under CC BY-SA 3.0 and CC0 1.0 license respectively. We will release our dataset under CC BY-SA 3.0.

### 3.2 Task Definition

Given a event mention in context, the task is to link the mention to the corresponding event from a candidate pool of events from Wikidata. For instance, the three mentions from column 1 in Figure 1 are to be linked to the Wikidata event, Q830917. Following Logeswaran et al. (2019), we adopt an in-KB evaluation approach. Therefore, every mention in our dataset refers to a valid event from the KB (Wikidata). We collect descriptions for the Wikidata events from all the corresponding language Wikipedias. Specifically, we obtain the article title and the first paragraph to compile the description. This way we obtain multilingual descriptions for all of our events (column 2 in Figure 1). We propose two variants of the event linking task, *multilingual* and *crosslingual*, depending on the source and target languages. We define the language of input mention as the source language, and the one of event description as the target language. Note that the event label itself (e.g. Q830917) is language-agnostic, facilitating crosslingual learning.

**Multilingual Event Linking:** Given a mention from language X, the linker searches through the event candidates from the same language X to identify the correct link. The source and target language are the same in the multilingual task. Additionally, as noted in Figure 4, the pool of event candidates varies across languages, thereby affecting the difficulty of the individual language linking tasks.

**Crosslingual Event Linking:** Given a mention from any language X, the linker searches the entire pool of event candidates to identify the link. We restrict the target language to English, thereby requiring the linker to only make use of the English descriptions for candidate events. Note that, all the events in our dataset have English descriptions.

**Creating Splits:** The train, dev and test distributions are presented in Table 1. The two tasks, multilingual and crosslingual share the same splits except for the difference in target language descriptions. Following previous works in entity linking

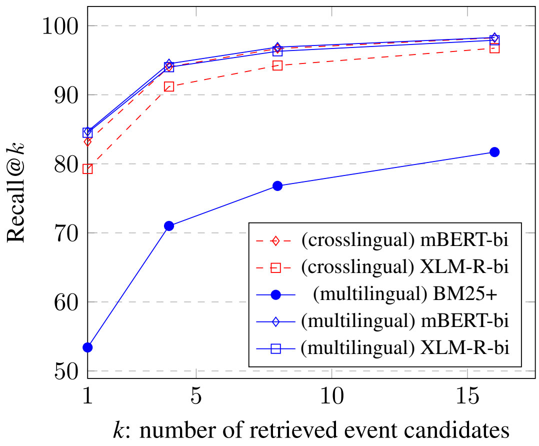


Figure 5: Event candidate retrieval performance on development split.

| Model | Multilingual | | Crosslingual | |
|---|---|---|---|---|
| | Dev | Test | Dev | Test |
| BM25+ | 53.4 | 50.1 | – | – |
| mBERT-bi | 84.7 | 84.6 | 83.2 | **83.9** |
| XLM-R-bi | 84.5 | 84.3 | 79.3 | 79.1 |
| mBERT-cross | 89.8 | **89.3** | 81.3 | 73.9 |
| XLM-R-cross | 88.8 | 87.3 | 81.0 | 75.6 |

Table 2: Event Linking Accuracy. For biencoder models, we report Recall@1.

literature, we focus on the zero-shot linking task. This requires the evaluation events to be fully disjoint from the train events. However, in the case of Wikidata events, a simple random split of events is still not sufficient. We observed a number of cases of event sequences in Wikidata (see Figure 3 for an example). Since events in the above sequence are closely related to each other, we add a constraint that events from a single sequence should not be shared between train, dev and test. Therefore, the event sequences are also disjoint between the splits. Additionally, the presence of these event sequences makes the linking task more challenging. The models need to perform temporal and spatial reasoning to distinguish between events from a sequence.

## 4 Modeling

In this section, we present our systems for multilingual and crosslingual event linking to Wikidata. We follow the entity linking system BLINK (Wu et al., 2020) to adapt a retrieve and rank approach. Given a mention, we first use a BERT-based biencoder to retrieve top-k events from the candidate pool. Then, we use a crossencoder to rerank these top-k candidates and identify the best event label. Additionally, following the baselines from entity linking literature, we also experiment with BM25 as a candidate retrieval method.

**BM25:** is a commonly used tf-idf based ranking function in information retrieval systems. Its found to be a competitive baseline for entity linking. We explore three variants of BM25, the original BM25Okapi (Robertson et al., 1994), BM25+ (Lv and Zhai, 2011a) and BM25L (Lv and Zhai, 2011b). In our experiments, we use the open-source implementation of Brown (2020). We use the mention as our query and the event descriptions as our documents. Since BM25 is a bag-of-words method, we only use in the multilingual task with same source and target languages. To create the documents, we use the concatenation of title and description of events. For the query, we experiment with increasing context window sizes of 8, 16, 32, 64 and 128 along with a mention-only baseline.

**Multilingual-BLINK:** We adapt the entity linking system BLINK (Wu et al., 2020) to event linking. BLINK utilizes a two-stage pipeline consisting of a candidate retrieval model (biencoder) and a candidate re-ranking model (crossencoder).

**Biencoder:** Using two multilingual transformers, we independently encode the context and event candidates. The context input is constructed as "`[CLS]` left context `[MENTION_START]` mention `[MENTION_END]` right context `[SEP]`". For event candidates, we use a concatenation of event's title and description, "`[CLS]` title `[EVT]` description `[SEP]`". In both cases, we use the final layer `[CLS]` token representation as our embedding. For each context, we score the event candidates by taking a dot product between the two embeddings. For training, we follow prior work (Lerer et al., 2019; Wu et al., 2020) to make use of in-batch random negatives. During inference, we run a nearest neighbour search to find the top-k candidates.

**Crossencoder:** In our crossencoder, the input constitutes a concatenation of the context and a given event candidate. Specifically, "`[CLS]` left context `[MENTION_START]` mention `[MENTION_END]` right context `[SEP]` title `[EVT]` description `[SEP]`." We encode the input using a transformer and utilize the last layer `[CLS]` token representation as the output embedding. We add a linear classification layer on top the encoder. Since our event dictionary is large, we run crossencoder training and inference

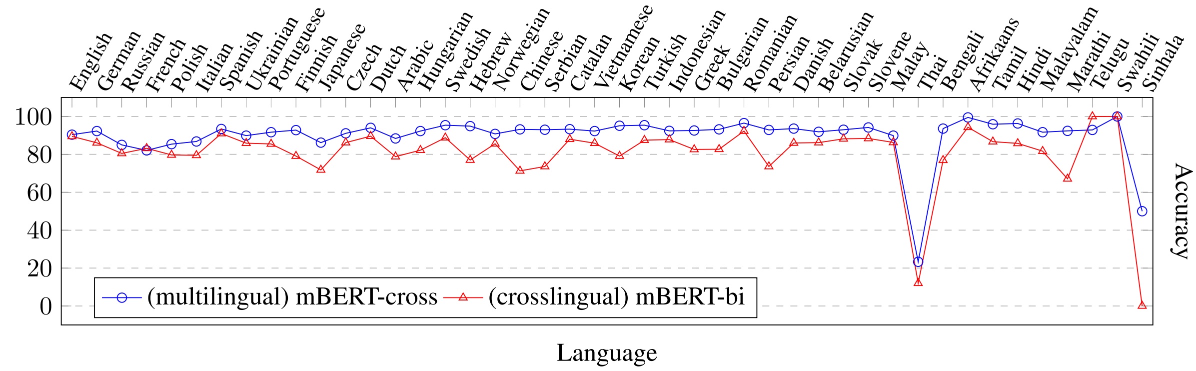


Figure 6: Test accuracy of mBERT-bi and mBERT-cross in multilingual and crosslingual tasks. The languages on the x-axis are sorted in the increasing order of mentions.

only on the top-k event candidates retrieved using the above biencoder model. During training, we optimize a softmax loss to predict the gold event candidate within the retrieved top-k. For inference, we predict the highest scoring context-candidate tuple from the top-k candidates. We experiment with two multilingual transformer-based encoders, mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). Henceforth, we refer to them as mBERT-bi, XLM-RoBERTa-bi and mBERT-cross, XLM-RoBERTa-cross for biencoder and crossencoder configurations respectively. For crossencoder training and inference, we use the retrieval results from the same BERT-based biencoder, i.e., mBERT-cross uses outputs from mBERT-bi. In all of our experiments, we optimize all the encoder layers. For biencoder training, we use AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-05 and a linear warmup schedule. We restrict the context and candidate lengths to 128 sub-tokens and select the best epoch (of 5) on the development set. For crossencoder training, we also use AdamW optimizer with a learning rate of 2e-05 and a linear warmup schedule. We restrict the overall sequence length to 256 sub-tokens and select the best epoch (of 5) on the development set.

## 5 Evaluation

We present our results on the development and test splits of the proposed dataset. In our experiments, we use bert-base-multilingual-uncased and xlm-roberta-base from Huggingface transformers (Wolf et al., 2020). For the multilingual task, even though the candidate set is partly different between languages, we share the model weights across languages. We believe this weight sharing helps in improving the performance on low-resource languages (Arivazhagan et al., 2019). We follow the standard metrics from prior work on entity linking, both for retrieval and reranking.

**Recall@$k$:** fraction of contexts where the gold event is contained in the top-k retrieved candidates.

**Accuracy:** fraction of contexts where the predicted event candidate matches the gold candidate. Logeswaran et al. (2019) defined two scores, normalized and unnormalized accuracies. Normalized accuracy only includes data instances where the gold candidate is found in the top-k retrieved candidates. Its primarily to compare different ranking methods with the same underlying retrieval model. Unnormalized accuracy reports the score on all the data instances, giving a better indicator of the overall end-to-end performance (retrieve+rank). In this work, we only report the unnormalized accuracy.

**Results:** In Figure 5, we present the retrieval results for both multilingual and crosslingual tasks. As expected, the Recall@$k$ scores increase with $k$ (1,4,8,16), but the performance is mostly similar for $k$=8 and $k$=16 for both biencoder models. The biencoder models significantly outperform the best BM25 configuration, BM25+ (with a context window of 16). For a detailed comparison of various configurations of BM25 baseline, refer to Figure 7 in Appendix. Based on the above results on the development set, we select $k$=8 for our crossencoder experiments.[7] As we notice, biencoders perform very well on the development set for both multilingual and crosslingual tasks.[8] Table 2 presents the accuracy scores for all of our experiment configurations. In addition to the crossencoder results, we report the Recall@1 scores from the retrieval meth-

[7] see Table 4 in Appendix for Recall@8 scores for all the configurations.

[8] We also report the results on test set for completeness.

| | Mention (+Context) | Predicted (P), Gold (G) Label |
|---|---|---|
| 1 | At the 2000 Summer Olympics in Sydney, Sitnikov competed only in two swimming events. ... Three days later, in the **100 m freestyle**, Sitnikov placed fifty-third on the morning prelims. ... | (**P**) Swimming at the 2008 Summer Olympics – Men's 100 metre freestyle, (**G**) Swimming at the 2000 Summer Olympics – Men's 100 metre freestyle |
| 2 | ... war er bei der Oscarverleihung 1935 erstmals für einen Oscar für den besten animierten Kurzfilm nominiert. Eine weitere Nominierung in dieser Kategorie erhielt er **1938** für "The Little Match Girl" (1937). | (**P**) The 9th Academy Awards were held on March 4, 1937, ... (**G**) The 10th Academy Awards were originally scheduled ... but due to ... were held on March 10, 1938, .. |
| 3 | Ivanova won the silver medal at the 1978 World Junior Championships. She made her senior World debut at the **1979 World Championships**, finishing 18th. Ivanova was 16th at the 1980 Winter Olympics. | (**P**) FIBT World Championships 1979, (**G**) 1979 World Figure Skating Championships |
| 4 | ... 攝津號與其姐妹艦河內號於1914年10月至11月間參與了青島戰役 的最後階段 ... | (**P**) Battle of the Yellow Sea, (**G**, en) Siege of Tsingtao: The siege of Tsingtao (or Tsingtau) was the attack on the German port of Tsingtao (now Qingdao) ... (**G**, zh) 青島戰役（，）是第一次世界大戰初期日本進攻德國膠州灣殖民地及其首府青島的一場戰役，也是唯一的一場戰役。 |

Table 3: Examples of errors by the event linking system.

ods for comparison. In the multilingual task, we find the mBERT crossencoder model to perform the best and significantly better than the corresponding biencoder model. However, in the crosslingual task, mBERT biencoder performs the best. Additionally, as expected, the crosslingual task is more challenging than the multilingual task. Due to the large number of model parameters, all of our reported results were based on a single training run.

**Performance by Language:** Multilingual and crosslingual tasks have three major differences: 1) source & target language, 2) language-specific descriptions can be more informative than English descriptions, and 3) candidate pool varies language (see Figure 4). Figure 6 plots the per-language test accuracy for our two best configurations from Table 2. As expected, we see *slightly lower* crosslingual performance, especially for medium and low-resource languages.[9] Thai and Sinhala are supported by XLM-R but not mBERT. We also perform qualitative analysis of errors made by our mBERT-based biencoder models on multilingual and crosslingual tasks. We summarize our observations from this analysis below,

**Temporal Reasoning:** The event linker occasionally performs insufficient temporal reasoning in the context (see example 1 in Table 3). Since our dataset contains numerous event sequences, such temporal reasoning is often important.

**Temporal and Spatial expressions:** In cases where the anchor text is a temporal or spatial expression, we found the system sometimes struggle to link to the event even if the link can be infered given the context information (see example 2 in Table 3). We believe these examples will also serve as interesting challenge for future work on our dataset.

**Event Descriptions:** Crosslingual system occasionally struggles with the English description. In example 4 from Table 3, we notice the mention matches exactly with the language Wikipedia title but it struggles with English description. Therefore, depending on the event, we hypothesize that language-specific event descriptions can sometimes be more informative than the English description.

**Dataset Errors:** We found system errors caused due to issues with the dataset. Specifically, we found instances where the context doesn't provide sufficient information needed for grounding (see example 3 in Table 3). Albeit uncommon, we found a few cases where the human annotated hyperlinks in Wikipedia can sometimes be incorrect.[10]

## 6 Conclusion & Future Work

In this work, we present a new task for multilingual linking of events to Wikidata. We compile a large-scale dataset for this task from Wikipedia and Wikidata, and propose two variants of the task, multilingual and crosslingual. Our results using a BERT-based biencoder and crossencoder models indicate that the crosslingual task is more challenging than the multilingual task. As highlighted earlier, one direction for future work is to expand the Wikidata event set to include hierarchical relations. Another direction would be expand the scope of the dataset by including references from newswire articles of Wikinews. Event linking systems could be improved further with better temporal reasoning and handling of temporal and spatial expressions.

[9]see Figure 10 in Appendix for per-language comparisons in other model configurations.

[10]For more detailed examples, refer to Table 7, Table 9 and Table 10 in Appendix.

## References


N. Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Z. Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Giusepppe Attardi. 2015. WikiExtractor.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Dorian Brown. 2020. Rank-BM25: A Collection of BM25 Algorithms in Python.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2498–2510, Online. Association for Computational Linguistics.

Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. 2017. Named entity disambiguation for noisy text. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. 2019. Pytorch-biggraph: A large scale graph embedding system. In *Proceedings of Machine Learning and Systems*, volume 1, pages 120–131.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yuanhua Lv and ChengXiang Zhai. 2011a. Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, page 7–16, New York, NY, USA. Association for Computing Machinery.

Yuanhua Lv and ChengXiang Zhai. 2011b. When documents are very long, bm25 fails! In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 1103–1104, New York, NY, USA. Association for Computing Machinery.

Rada Mihalcea and Andras Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 233–242, New York, NY, USA. Association for Computing Machinery.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA).

Joel Nothman, James R. Curran, and Tara Murphy. 2008. Transforming Wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 124–132, Hobart, Australia.

Joel Nothman, Matthew Honnibal, Ben Hachey, and James R. Curran. 2012. Event linking: Grounding event reference in a news archive. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232, Jeju Island, Korea. Association for Computational Linguistics.

Adithya Pratapa, Zhengzhong Liu, Kimihiro Hasegawa, Linwei Li, Yukari Yamakawa, Shikun Zhang, and Teruko Mitamura. 2021. Cross-document event identity via dense annotation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 496–517, Online. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *TREC*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avirup Sil, Ernest Cronin, Penghai Nie, Yinfei Yang, Ana-Maria Popescu, and Alexander Yates. 2012. Linking named entities to any database. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 116–127, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to wikipedia.

Shyam Upadhyay, Christos Christodoulopoulos, and Dan Roth. 2016. "making the news": Identifying noteworthy events in news articles. In *Proceedings of the Fourth Workshop on Events*, pages 1–7, San Diego, California. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

# A Appendix

## A.1 Ethical Considerations

In this work, we presented a new dataset compiled automatically from Wikipedia and Wikidata. After the initial collection process, we perform rigorous post-processing steps to reduce potential errors in our dataset. Our dataset is multilingual with texts from 44 languages. In our main paper, we state these languages as well as their individual representation in our dataset. As we highlight in the paper, the proposed linking systems only work for specific class of events (nominal) due to the nature of our dataset.

## A.2 Dataset

After identifying potential events from Wikidata, we perform additional post-processing to remove any non-event items. In Table 5, we present the list of all Wikidata properties we use for removing non-event items from our corpus. In Table 6, we present the list of all languages from our dataset along with their language genealogy and distribution in the dataset.

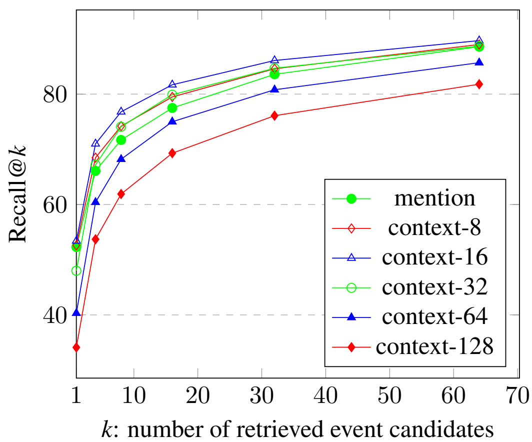


Figure 7: Effect of context window size on BM25+ retrieval performance.

| Retriever | Multilingual | | Crosslingual | |
|---|---|---|---|---|
| | Dev | Test | Dev | Test |
| BM25+ | 76.8 | 70.5 | – | – |
| mBERT-bi | 96.9 | **97.1** | 96.7 | **97.2** |
| XLM-R-bi | 96.3 | 96.7 | 94.2 | 95.3 |

Table 4: Event candidate retrieval results, Recall@8.

## A.3 Modeling

**Experiments:** We use the base versions of mBERT and XLM-RoBERTa in all of our experiments. In the biencoder model, we use two multilingual encoders, one each for context and candidate encoding. In crossencoder, we use just one multilingual encoder and a classification layer. We ran our experiments on a mix of GPUs, TITANX, v100, A6000 and a100. Each training and inference runs were run on a single GPU. Both biencoder and crossencoder were run for 5 epochs and we select the best set of hyperparameters based on the performance on the development set. On the a100 GPU, biencoder takes about 1.5hrs per epoch and the crossencoder takes ~20hrs per epoch (with k=8).

**Results:** In Figure 7, we present results on the development set from all the explored configurations. In Table 4, we show the Recall@8 scores from all the retrieval models. Based on the performance on development set, we selected $k$=8 for our crossencoder training and inference. We also report the test scores for completeness. Figure 8 presents the retrieval recall scores. Figure 9 presents the retrieval recall scores for BM25+ (context length 16) method. Figure 10 presents a detailed comparison of per-language accuracies between multilingual and crosslingual tasks for each biencoder and crossencoder models.

We also present full examples of system errors we identified through a qualitative analysis. Table 7 presents examples of system errors due to insufficient temporal reasoning in the context. Table 8 presents examples of system errors on mentions that are temporal or spatial expressions. Table 9 presents examples of system errors on crosslingual task due to issues related with tackling non-English mentions. Table 10 presents examples of system errors that were caused due to errors in the dataset itself.

| Property | Property_Label | URI | URI_Label |
|---|---|---|---|
| P31 | instance_of | Q48349 | empire |
| P31 | instance_of | Q11514315 | historical_period |
| P31 | instance_of | Q3024240 | historical_country |
| P31 | instance_of | Q11042 | culture |
| P31 | instance_of | Q28171280 | ancient_civilization |
| P31 | instance_of | Q1620908 | historical_region |
| P31 | instance_of | Q3502482 | cultural_region |
| P31 | instance_of | Q465299 | archaeological_culture |
| P31 | instance_of | Q568683 | age |
| P31 | instance_of | Q763288 | lander |
| P31 | instance_of | Q4830453 | business |
| P31 | instance_of | Q24862 | short_film |
| P31 | instance_of | Q1496967 | territorial_entity |
| P31 | instance_of | Q68 | computer |
| P31 | instance_of | Q486972 | human_settlement |
| P31 | instance_of | Q26529 | space_probe |
| P31 | instance_of | Q82794 | geographic_region |
| P31 | instance_of | Q43229 | organization |
| P31 | instance_of | Q15401633 | archaeological_period |
| P31 | instance_of | Q5398426 | television_series |
| P31 | instance_of | Q24869 | feature_film |
| P31 | instance_of | Q11424 | film |
| P31 | instance_of | Q718893 | theater |
| P31 | instance_of | Q1555508 | radio_program |
| P31 | instance_of | Q17343829 | unincorporated_community_in_the_United_States |
| P31 | instance_of | Q254832 | Internationale_Bauausstellung |
| P31 | instance_of | Q214609 | material |
| P31 | instance_of | Q625298 | peace_treaty |
| P31 | instance_of | Q131569 | treaty |
| P31 | instance_of | Q93288 | contract |
| P31 | instance_of | Q15416 | television_program |
| P31 | instance_of | Q1201097 | detachment |
| P31 | instance_of | Q16887380 | group |
| P31 | instance_of | Q57821 | fortification |
| P31 | instance_of | Q15383322 | cultural prize |
| P31 | instance_of | Q515 | city |
| P31 | instance_of | Q537127 | road_bridge |
| P31 | instance_of | Q20097897 | sea_fort |
| P31 | instance_of | Q1785071 | fort |
| P31 | instance_of | Q23413 | castle |
| P31 | instance_of | Q1484988 | project |
| P31 | instance_of | Q149621 | district |
| P31 | instance_of | Q532 | village |
| P31 | instance_of | Q2630741 | community |
| P31 | instance_of | Q3957 | town |
| P31 | instance_of | Q111161 | synod |
| P31 | instance_of | Q1530022 | religious_organization |
| P31 | instance_of | Q51645 | ecumenical_council |
| P31 | instance_of | Q10551516 | church_council |
| P31 | instance_of | Q1076486 | sports_venue |
| P31 | instance_of | Q17350442 | venue |
| P31 | instance_of | Q13226383 | facility |
| P31 | instance_of | Q811979 | architectural_structure |
| P31 | instance_of | Q23764314 | sports_location |
| P31 | instance_of | Q15707521 | fictional_battle |
| P36 | capital | * | |
| P2067 | mass | * | |
| P1082 | population | * | |
| P1376 | captial_of | * | |
| P137 | operator | * | |
| P915 | filming_location | * | |
| P162 | producer | * | |
| P281 | postal_code | * | |
| P176 | manufacturer | * | |
| P2257 | event_interval | * | |
| P527 | has_part | * | |
| P279 | subclass_of | * | |

Table 5: List of properties used for postprocessing Wikidata events. If a candidate event has the property ‘P31’, we prune them depending on the corresponding. For example, we only prune items that are instances of empire, historical period etc., For other properties like P527, P36, we prune items if they contain this property.

| Language | Code | Events | Mentions | Genus |
|---|---|---|---|---|
| Afrikaans | af | 316 | 2036 | Germanic |
| Arabic | ar | 2691 | 28801 | Semitic |
| Belarusian | be | 737 | 7091 | Slavic |
| Bulgarian | bg | 1426 | 12570 | Slavic |
| Bengali | bn | 270 | 3136 | Indic |
| Catalan | ca | 2631 | 22296 | Romance |
| Czech | cs | 2839 | 36658 | Slavic |
| Danish | da | 1189 | 10267 | Germanic |
| German | de | 7371 | 209469 | Germanic |
| Greek | el | 997 | 13361 | Greek |
| English | en | 10747 | 328789 | Germanic |
| Spanish | es | 5064 | 91896 | Romance |
| Persian | fa | 1566 | 10449 | Iranian |
| Finnish | fi | 3253 | 47944 | Finnic |
| French | fr | 8183 | 136482 | Romance |
| Hebrew | he | 1871 | 34470 | Semitic |
| Hindi | hi | 216 | 1219 | Indic |
| Hungarian | hu | 3067 | 27333 | Ugric |
| Indonesian | id | 2274 | 14049 | Malayo-Sumbawan |
| Italian | it | 7116 | 108012 | Romance |
| Japanese | ja | 3832 | 49198 | Japanese |
| Korean | ko | 1732 | 13544 | Korean |
| Malayalam | ml | 136 | 730 | Southern Dravidian |
| Marathi | mr | 132 | 507 | Indic |
| Malay | ms | 824 | 4650 | Malayo-Sumbawan |
| Dutch | nl | 4151 | 41973 | Germanic |
| Norwegian | no | 2514 | 24092 | Germanic |
| Polish | pl | 6270 | 110381 | Slavic |
| Portuguese | pt | 4466 | 45125 | Romance |
| Romanian | ro | 1224 | 12117 | Romance |
| Russian | ru | 7929 | 180891 | Slavic |
| Sinhala | si | 31 | 65 | Indic |
| Slovak | sk | 726 | 5748 | Slavic |
| Slovene | sl | 1288 | 8577 | Slavic |
| Serbian | sr | 1611 | 24093 | Slavic |
| Swedish | sv | 2865 | 23152 | Germanic |
| Swahili | sw | 22 | 74 | Bantoid |
| Tamil | ta | 250 | 1682 | Southern Dravidian |
| Telugu | te | 39 | 243 | South-Central Dravidian |
| Thai | th | 800 | 4749 | Kam-Tai |
| Turkish | tr | 2342 | 19846 | Turkic |
| Ukrainian | uk | 3428 | 53098 | Slavic |
| Vietnamese | vi | 1439 | 13744 | Viet-Muong |
| Chinese | zh | 2759 | 21259 | Chinese |
| **Total** | | **10947** | **1805866** | |

Table 6: Proposed dataset summary (by languages)

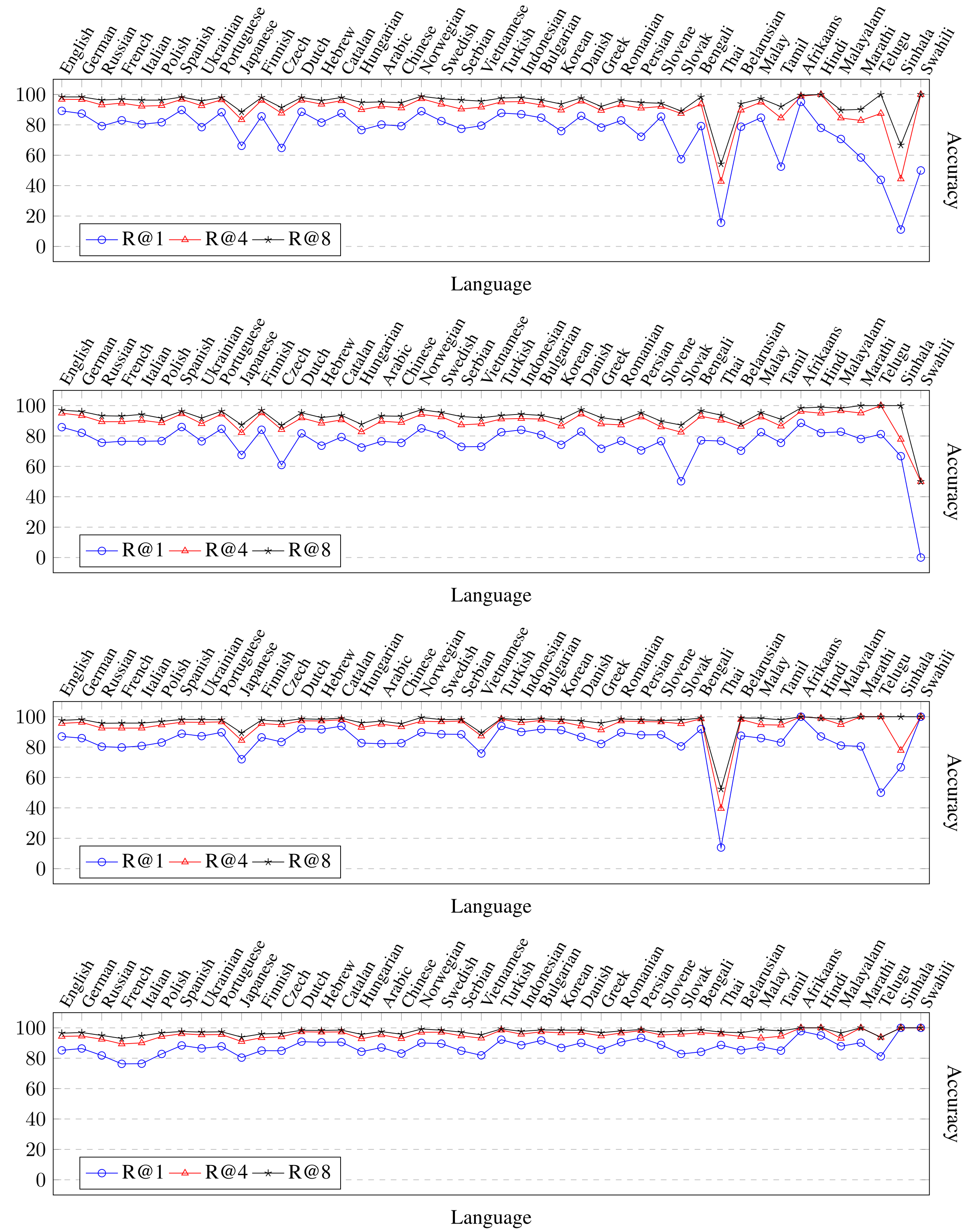


Figure 8: Retrieval recall scores on development set for mBERT and XLM-R in multilingual and crosslingual settings.

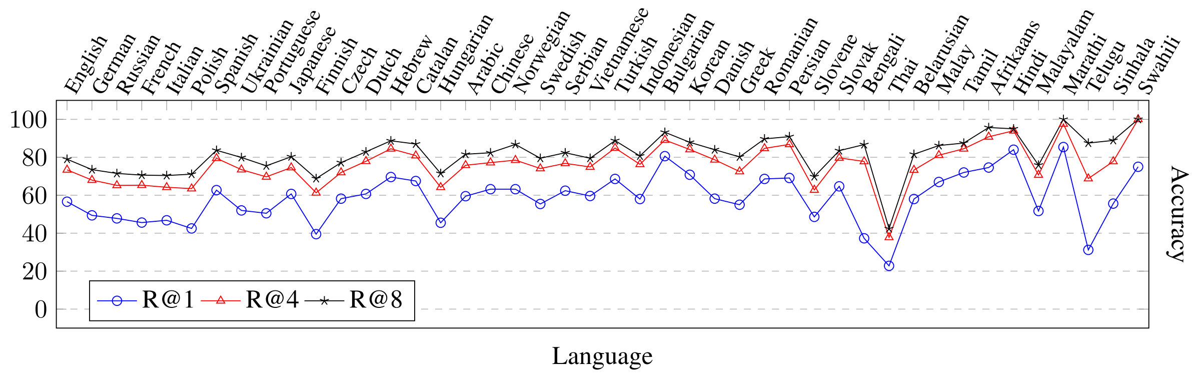


Figure 9: Retrieval recall scores on development set for BM25+ in multilingual setting.

| **Mention (+Context)** | **Predicted Label** | **Gold Label** |
|---|---|---|
| At the 2000 Summer Olympics in Sydney, Sitnikov competed only in two swimming events. He eclipsed a FINA B-cut of 51.69 (100 m freestyle) from the Kazakhstan Open Championships in Almaty. On the first day of the Games, Sitnikov placed twenty-first for the Kazakhstan team in the 4 × 100 m freestyle relay. Teaming with Sergey Borisenko, Pavel Sidorov, and Andrey Kvassov in heat three, Sitnikov swam a lead-off leg and recorded a split of 52.56, but the Kazakhs settled only for last place in a final time of 3:28.90. Three days later, in the **100 m freestyle**, Sitnikov placed fifty-third on the morning prelims. Swimming in heat five, he raced to a fifth seed by 0.15 seconds ahead of Chinese Taipei's Wu Nien-pin in 52.57. | *Swimming at the 2008 Summer Olympics – Men's 100 metre freestyle*: The men's 100 metre freestyle event at the 2008 Olympic Games took place on 12–14 August at the Beijing National Aquatics Center in Beijing, China. There were 64 competitors from 55 nations. | *Swimming at the 2000 Summer Olympics – Men's 100 metre freestyle*: The men's 100 metre freestyle event at the 2000 Summer Olympics took place on 19–20 September at the Sydney International Aquatic Centre in Sydney, Australia. There were 73 competitors from 66 nations. Nations have been limited to two swimmers each since the 1984 Games. |
| In 2012, WWE reinstated their No Way Out pay-per-view (PPV), which had previously ran annually from 1999 to 2009. The following year, however, No Way Out was canceled and replaced by Payback, which in turn became an annual PPV for the promotion. The first Payback event was held on June 16, 2013 at the Allstate Arena in Rosemont, Illinois. The 2014 event was also held in June at the same arena and was also the first Payback to air on the WWE Network, which had launched earlier that year. In 2015 and 2016, the event was held in May. The 2016 event was also promoted as the first PPV of the New Era for WWE. In July 2016, WWE reintroduced the brand extension, dividing the roster between the Raw and SmackDown brands where wrestlers are exclusively assigned to perform. The **2017 event** was in turn held exclusively for wrestlers from the Raw brand, and was also moved up to late-April. | *Battleground (2017)*: Battleground was a professional wrestling pay-per-view (PPV) event and WWE Network event produced by WWE for their SmackDown brand division. It took place on July 23, 2017, at the Wells Fargo Center in Philadelphia, Pennsylvania. It was the fifth and final event under the Battleground chronology, as following WrestleMania 34 in April 2018, brand-exclusive PPVs were discontinued, resulting in WWE reducing the amount of yearly PPVs produced. | *Payback (2017)*: Payback was a professional wrestling pay-per-view (PPV) and WWE Network event, produced by WWE for the Raw brand division. It took place on April 30, 2017 at the SAP Center in San Jose, California. It was the fifth event in the Payback chronology. Due to the Superstar Shake-up, the event included two interbrand matches with SmackDown wrestlers. It was the final Payback event until 2020, as following WrestleMania 34 in 2018, WWE discontinued brand-exclusive PPVs, which resulted in the reduction of yearly PPVs produced. |

Table 7: Examples of errors by the event linking system. (temporal reasoning related)

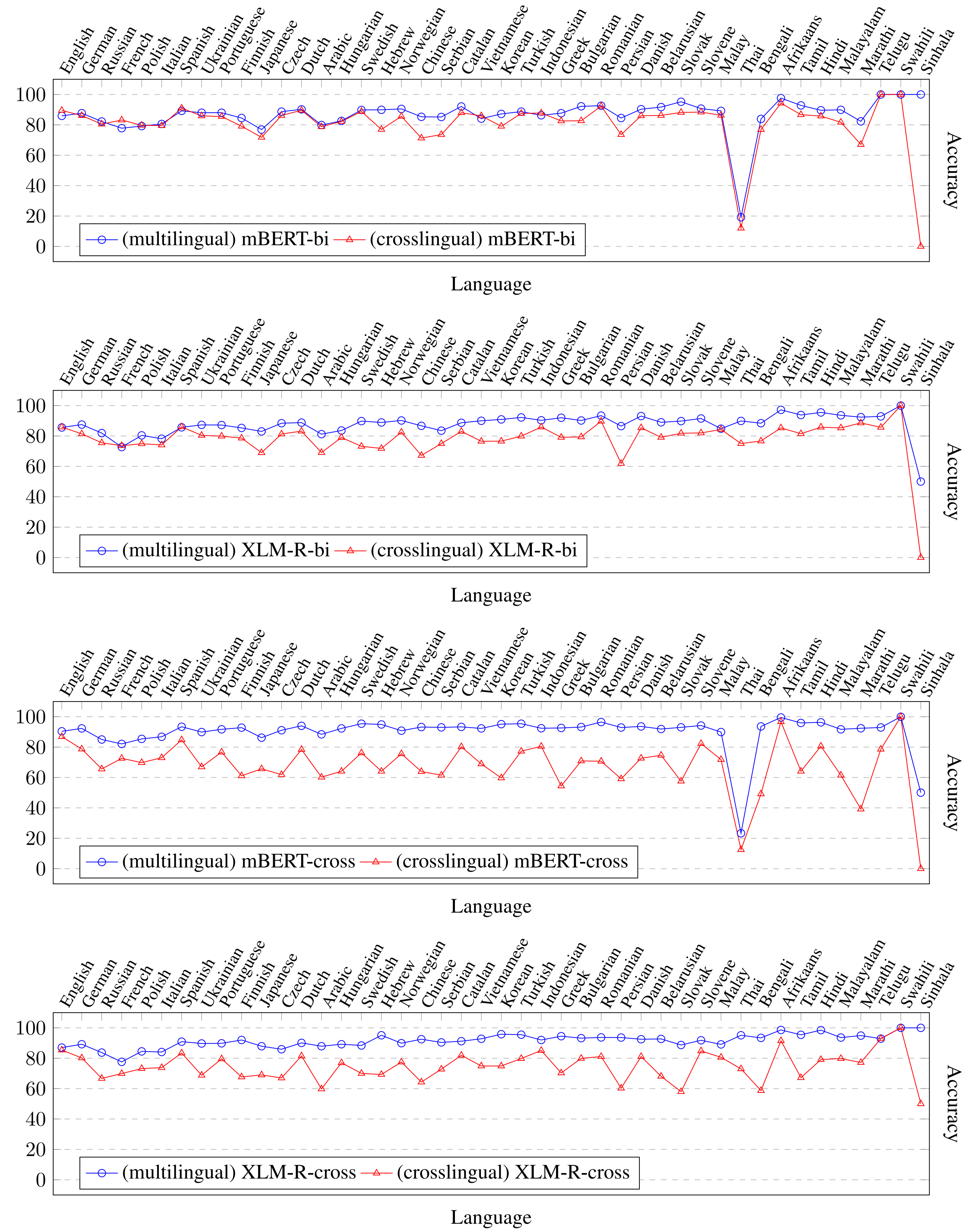


Figure 10: Test accuracy of mBERT-bi, XLM-R-bi, mBERT-cross, XLM-R-cross in multilingual and crosslingual tasks. The languages on the x-axis are sorted in the increasing order of mentions.

| Mention (+Context) | Predicted Label | Gold Label |
|---|---|---|
| Paul Wing (August 14, 1892 – May 29, 1957) was an assistant director at Paramount Pictures. He won the **1935** Best Assistant Director Academy Award for "The Lives of a Bengal Lancer" along with Clem Beauchamp. Wing was the assistant director on only two films owing to his service in the United States Army. During his service, Wing was in a prisoner camp that was portrayed in the film "The Great Raid" (2005). | *8th Academy Awards*: The 8th Academy Awards were held on March 5, 1936, at the Biltmore Hotel in Los Angeles, California. They were hosted by Frank Capra. This was the first year in which the gold statuettes were called "Oscars". | *7th Academy Awards*: The 7th Academy Awards, honoring the best in film for 1934, was held on February 27, 1935, at the Biltmore Hotel in Los Angeles, California. They were hosted by Irvin S. Cobb. |
| Für "Holiday Land" (1934) war er bei der Oscarverleihung 1935 erstmals für einen Oscar für den besten animierten Kurzfilm nominiert. Eine weitere Nominierung in dieser Kategorie erhielt er **1938** für "The Little Match Girl" (1937). | *9th Academy Awards*: The 9th Academy Awards were held on March 4, 1937, at the Biltmore Hotel in Los Angeles, California. They were hosted by George Jessel; music was provided by the Victor Young Orchestra, which at the time featured Spike Jones on drums. This ceremony marked the introduction of the Best Supporting Actor and Best Supporting Actress categories, and was the first year that the awards for directing and acting were fixed at five nominees per category. | *10th Academy Awards*: The 10th Academy Awards were originally scheduled for March 3, 1938, but due to the Los Angeles flood of 1938 were held on March 10, 1938, at the Biltmore Hotel in Los Angeles, California. It was hosted by Bob Burns. |

Table 8: Examples of errors by the event linking system. (temporal or spatial expression related)

| Mention (+Context) | Predicted Label | Gold Label |
|---|---|---|
| Nel 2018 ha preso parte alle Olimpiadi di Pyeongchang, venendo eliminata nel primo turno della finale e classificandosi diciannovesima nella gara di **gobbe**. | *Snowboarding at the 2018 Winter Olympics – Women's parallel giant slalom*: The women's parallel giant slalom competition of the 2018 Winter Olympics was held on 24 February 2018 Bogwang Phoenix Park in Pyeongchang, South Korea. | *Freestyle skiing at the 2018 Winter Olympics – Women's moguls*: The Women's moguls event in freestyle skiing at the 2018 Winter Olympics took place at the Bogwang Phoenix Park, Pyeongchang, South Korea from 9 to 11 February 2018. It was won by Perrine Laffont, with Justine Dufour-Lapointe taking silver and Yuliya Galysheva taking bronze. For Laffont and Galysheva these were first Olympic medals. Galysheva also won the first ever medal in Kazakhstan in freestyle skiing. |
| تقارب إسرائيل واليابان على أساس القيم الديموقراطية والاشتراكية المشتركة، واستطاعت من خلال عضويتها في الاشتراكية الدولية أن تنشئ صلات وثيقة مع الحزب الاشتراكي الياباني الذي تبنى مهمة التعريف بإسرائيل ومنجزاتها في اليابان. وإبَّان **حرب 1956** انضمت اليابان إلى الدول التي طالبت مصر باحترام المعاهدات الدولية الخاصة بالملاحة في قناة السويس. وأصدرت بيان مقتضب أعلنت فيه أسفها لوصول الأمور إلى حد الصدام المسلح. | *Hungarian Revolution of 1956*: The Hungarian Revolution of 1956 (), or the Hungarian Uprising, was a nationwide revolution against the Hungarian People's Republic and its Soviet-imposed policies, lasting from 23 October until 10 November 1956. Leaderless at the beginning, it was the first major threat to Soviet control since the Red Army drove Nazi Germany from its territory at the end of World War II in Europe. | *Suez Crisis*: The Suez Crisis, or the Second Arab–Israeli war, also called the Tripartite Aggression () in the Arab world and the Sinai War in Israel, |
| 攝津號戰艦於1909年4月1日在橫須賀海軍工廠鋪設龍骨，後於1909年1日18日舉行下水儀式，並於1912年7月1日竣工，總造價為11,010,000日圓。海軍大佐田中盛秀於1912年12月1日出任本艦艦長，並編入第一分遣艦隊。翌年的多數時候，攝津號均巡航於中國外海或是接受戰備操演。當第一次世界大戰於1914年8月間爆發時，本艦正停泊於廣島縣吳市軍港。攝津號與其姐妹艦河內號於1914年10月至11月間參與了**青島戰役**的最後階段，並於外海以艦砲密集轟炸德軍陣地。本艦於1916年12月1日離開第一分遣艦隊，並送往吳市進行升級作業。升級作業於1917年12月1日完成，該艦隨後編入第二分遣艦隊，直至1918年7月23日重新歸入第一分遣艦隊為止。自此時起，攝津號戰艦上所有的QF 12磅3英吋40倍徑艦砲均移除，並以QF 12磅3英吋40倍徑防空砲取代，另亦移除了兩具魚雷發射管。1918年10月28日，攝津號戰艦成為大正天皇於海上校閱時所搭乘的旗艦。 | *Battle of the Yellow Sea*: The Battle of the Yellow Sea (; ) was a major naval battle of the Russo-Japanese War, fought on 10 August 1904. In the Russian Navy, it was referred to as the Battle of 10 August. The battle foiled an attempt by the Russian fleet at Port Arthur to break out and form up with the Vladivostok squadron, forcing them to return to port. Four days later, the Battle off Ulsan similarly ended the Vladivostok group's sortie, forcing both fleets to remain at anchor. | *Siege of Tsingtao*: The siege of Tsingtao (or Tsingtau) was the attack on the German port of Tsingtao (now Qingdao) in China during World War I by Japan and the United Kingdom. The siege was waged against Imperial Germany between 27 August and 7 November 1914. The siege was the first encounter between Japanese and German forces, the first Anglo-Japanese operation of the war, and the only major land battle in the Asian and Pacific theatre during World War I. |

Table 9: Examples of errors by the event linking system. (language-related)

| Mention (+Context) | Predicted Label | Gold Label |
|---|---|---|
| He established his own production company, Emirau Productions, named after the **battle in World War II** in which Warren was injured. | *First Battle of El Alamein*: The First Battle of El Alamein (1–27 July 1942) was a battle of the Western Desert Campaign of the Second World War, fought in Egypt between Axis forces (Germany and Italy) of the Panzer Army Africa () (which included the under Field Marshal () Erwin Rommel) and Allied (British Imperial and Commonwealth) forces (Britain, British India, Australia, South Africa and New Zealand) of the Eighth Army (General Claude Auchinleck). | *Landing on Emirau*: The Landing on Emirau was the last of the series of operations that made up Operation Cartwheel, General Douglas MacArthur's strategy for the encirclement of the major Japanese base at Rabaul. A force of nearly 4,000 United States Marines landed on the island of Emirau on 20 March 1944. The island was not occupied by the Japanese and there was no fighting. It was developed into an airbase which formed the final link in the chain of bases surrounding Rabaul. The isolation of Rabaul permitted MacArthur to turn his attention westward and commence his drive along the north coast of New Guinea toward the Philippines. |
| Ivanova won the silver medal at the 1978 World Junior Championships. She made her senior World debut at the **1979 World Championships**, finishing 18th. Ivanova was 16th at the 1980 Winter Olympics. | *FIBT World Championships 1979*: The FIBT World Championships 1979 took place in Königssee, West Germany. It was the first championships that took place on an artificially refrigerated track. The track also hosted the luge world championships that same year, the first time that had ever happened in both bobsleigh and luge in a non-Winter Olympic year (Igls hosted both events for the 1976 games in neighboring Innsbruck.). | *1979 World Figure Skating Championships*: The 1979 World Figure Skating Championships were held in Vienna, Austria from March 13 to 18. At the event, sanctioned by the International Skating Union, medals were awarded in men's singles, ladies' singles, pair skating, and ice dance. |
| Изначально открытие башни должно было состояться в декабре 2011 года, но после **землетрясения** строительство замедлилось из-за нехватки средств. | *2011 Christchurch earthquake*: A major earthquake occurred in Christchurch, New Zealand, on Tuesday 22 February 2011 at 12:51 p.m. local time (23:51 UTC, 21 February). The () earthquake struck the Canterbury region in the South Island, centred south-east of the centre of Christchurch, the country's second-most populous city. It caused widespread damage across Christchurch, killing 185 people, in the nation's fifth-deadliest disaster. | *2011 Tōhoku earthquake and tsunami*: The occurred at 14:46 JST (05:46 UTC) on 11 March. The magnitude 9.0–9.1 (Mw) undersea megathrust earthquake had an epicenter in the Pacific Ocean, east of the Oshika Peninsula of the Tōhoku region, and lasted approximately six minutes, causing a tsunami. It is sometimes known in Japan as the , among other names. The disaster is often referred to in both Japanese and English as simply 3.11 (read šan ten ichi-ichiïn Japanese). |
| ポワント・デュ・オック (Pointe du Hoc) から向かったアメリカ軍のレンジャー部隊の8個中隊と共に、アメリカ第29歩兵師団は海岸の西側の側面を攻撃した。アメリカ第1歩兵師団は東側からのアプローチを行った。これは、この戦争において、**北アフリカ**、シチリア島に続く3回目の強襲上陸であった。オマハビーチの上陸部隊の主目標は、サン＝ロー (Saint-Lô) の南に進出する前にポール＝アン＝ベッサン (Port-en-Bessin) とヴィル川 (Vire River) 間の橋頭堡を守ることであった。 | *Tunisian campaign*: The Tunisian campaign (also known as the Battle of Tunisia) was a series of battles that took place in Tunisia during the North African campaign of the Second World War, between Axis and Allied forces. The Allies consisted of British Imperial Forces, including a Greek contingent, with American and French corps. The battle opened with initial success by the German and Italian forces but the massive supply interdiction efforts led to the decisive defeat of the Axis. Over 250,000 German and Italian troops were taken as prisoners of war, including most of the Afrika Korps. | *Operation Torch*: Operation Torch (8 November 1942 – 16 November 1942) was an Allied invasion of French North Africa during the Second World War. While the French colonies formally aligned with Germany via Vichy France, the loyalties of the population were mixed. Reports indicated that they might support the Allies. American General Dwight D. Eisenhower, supreme commander of the Allied forces in Mediterranean Theater of Operations, planned a three-pronged attack on Casablanca (Western), Oran (Center) and Algiers (Eastern), then a rapid move on Tunis to catch Axis forces in North Africa from the west in conjunction with Allied advance from east. |

Table 10: Examples of errors by the event linking system. (also errors in the dataset)