TempEL: Linking Dynamically Evolving and Newly Emerging Entities

Klim Zaporojets ♮ **Lucie-Aimée Kaffee** ♯ **Johannes Deleu** ♮
Thomas Demeester ♮ **Chris Develder** ♮ **Isabelle Augenstein** ♯
♮ Ghent University – imec, IDLab, Ghent, Belgium
♯ Dept. of Computer Science, University of Copenhagen, Denmark
{klim.zaporojets,johannes.deleu,thomas.demeester,chris.develder}@ugent.be
{kaffee,augenstein}@di.ku.dk

A Supplementary material

A.1 Dataset and code distribution

Link to the dataset The reviewers can access the dataset using the following link: https://cloud.ilabt.imec.be/index.php/s/RinXy8NgqdW58RW. The dataset and the baseline code will be made publicly available in a dedicated GitHub repository upon acceptance.

License TempEL is distributed under Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0).[1]

Maintenance The maintenance and extension to further temporal snapshots of TempEL will be carried out by the authors of the paper. Additionally, we will make the code public to create potential new variations and extensions of TempEL using a number of hyperparameters (see Sections A.4 and A.5 for further details).

A.2 Datasheet for TempEL

In this section we provide a more detailed documentation of the dataset with the intended uses. We base ourselves on the datasheet proposed by [1].

A.2.1 Motivation

For what purpose was the dataset created? The TempEL dataset was created to evaluate how the temporal change of anchor mentions and that of target Knowledge Base (KB; i.e., modification or creation of new entities) affects the *entity linking* (EL) task. This contrasts with the currently existing datasets [9, 7, 8, 6], which are associated with a single version of the target KB such as the Wikipedia 2010 for the widely adopted CoNLL-AIDA[2] dataset. We expect that TempEL will encourage research in devising new models and architectures that are robust to temporal changes both in mentions as well as in the target KBs.

Who created the dataset and on behalf of which entity? The dataset is the result of joint effort involving researchers from the University of Copenhagen and Ghent University.

Who funded the creation of the dataset? The creation of TempEL was funded by the following grants:

[1] https://creativecommons.org/licenses/by-sa/4.0/

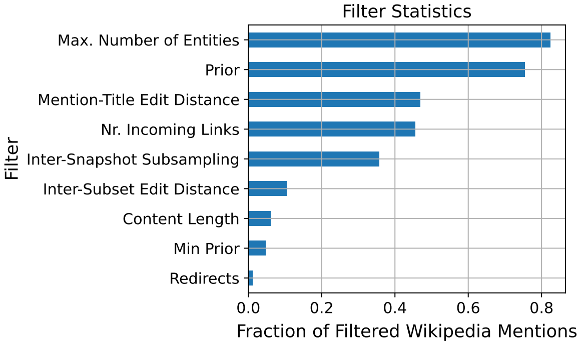

Figure 1: Figure showcasing the fraction of filtered Wikipedia mentions by each of the filters executed during TempEL generation.

1. FWO (Fonds voor Wetenschappelijk Onderzoek) long-stay abroad grant V412922N.
2. The Flemish Government fund under the programme "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen".

A.2.2 Composition

What do the instances that comprise the dataset represent? Each of the instances consists of a mention in Wikipedia linked to target entity, i.e., a Wikipedia page, with a set of attributes. The dataset is organized in 10 yearly temporal snapshots starting from January 1, 2013 until January 1, 2022. See Section A.6 for further details on the attributes associated with each of the instances of our TempEL dataset.

How many instances are there in total? Table 1 of the main manuscript summarizes the number of instances (# Anchor Mentions) of each of the entity categories (*continual* and *new*) in TempEL. See Section A.3 for additional statistics on mention per entity distribution.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? TempEL contains a sample of all the possible anchor mentions linked to target entities from Wikipedia. The following are the filters applied to obtain the instances in the final TempEL dataset whose effect is also summarized in Fig. 1:

1. **Prior-based filtering**: we exclude all the mentions for which the correct entity it refers to has the highest *prior* [12] as calculated in Eq. (1) of the manuscript. This filtering is done with the goal of creating a more challenging dataset.

   Value to create TempEL: mentions with mention prior rank >1 among other mentions referring to the same entity.

   Percentage of filtered out instances: between 74.20% and 76.28%, depending on the temporal snapshot.

   Hyperparameter name: `min_men_prior_rank` (see Table 1 in Section A.4).

2. **Entity relevance filtering**: we impose the restriction for target entity of having at least 10 incoming links (i.e., at least 10 mentions linking to it) in order to be included in TempEL. Additionally, we filter out target entities whose description contains less than 10 tokens. This is done in order to avoid introducing potentially noisy and irrelevant entities that have not been sufficiently established by the Wikipedia community.

   Value to create TempEL: 10 for minimum number of incoming links and 10 for minimum content length (in number of tokens) of target entity.

   Percentage of filtered out instances:

   - Minimum number of incoming links: between 42.66% and 48.32%, depending on the temporal snapshot.
   - Minimum content length: between 0.06% and 0.95% depending on the temporal snapshot.

Hyperparameter names: `min_nr_inlinks` for minimum number of incoming links and `min_len_target_ent` for minimum number of content length tokens (see Table 1 in Section A.4).

3. **Min prior subsampling**: the mentions with very low mention prior are filtered out from TempEL. This way, we avoid introducing too infrequent and potentially erroneous mentions to refer to a particular entity.

   Value to create TempEL: 0.0001

   Percentage of filtered out instances: between 0.37% and 0.61%, depending on the snapshot.

   Hyperparameter name: `min_men_prior` (see Table 1 in Section A.4).

4. **Minimum mentions per entity**: has similar effect as previously explained *min prior subsampling* (see above) filter. We do not use it in the creation of TempEL, relying completely on the *min prior subsampling* filter.

   Value to create TempEL: 1

   Percentage of filtered out instances: 0%

   Hyperparameter name: `min_mens_per_ent` (see Table 1 in Section A.4).

5. **Edit distance mention title**: filters out the anchor mentions that are very similar to target entity page. This way, we expect to reduce the trivial cases where the entity linking can be simply predicted by mapping the mention to the title of the target entity.

   Value to create TempEL: 0.2 (normalized edit distance).

   Percentage of filtered out instances: between 44.85% and 48.99%, depending on the snapshot.

   Hyperparameter name: `ed_men_title` (see Table 1 in Section A.4).

6. **Redirect filtering**: we filter out anchor mentions that point to redirect pages (pages without content redirecting to other pages in Wikipedia).

   Percentage of filtered out instances: between 1.02% and 1.47%, depending on the snapshot.

7. **Inter-subset filtering**: we enforce normalized edit distance between the mentions in different subsets referring to the same target entity to be higher than 0.2. This entails that the entities in TempEL are linked to at least by 3 mentions with different surface form. The main goal of this filter is to avoid mention-entity tuple memorization by the models [5].

   Value to create TempEL: 0.2 normalized edit distance between mentions in different subsets.

   Percentage of filtered out instances: 10%.

   Hyperparameter name: `ed_men_subsets` (see Table 1 in Section A.4).

8. **Maximum number of entities**: we restrict the number of target entities to 10,000 for *continual* instances. The reason behind this is to build a dataset of manageable size with a reasonable number of target entities to experiment with.

   Value to create TempEL: 10,000 for *continual* entities.

   Percentage of filtered out instances: 82%.

   Hyperparameter name: `nr_ct_ents_per_cut` (see Table 1 in Section A.4)

9. **Maximum number of mentions per entity**: this filtering limits the number of mentions per entity in order for the dataset to not be dominated by most popular entities. Particularly, for test and evaluation subsets we limit the number of mentions per entity to 10. This way, we expect the accuracy scores to not be dominated by links to popular target entities (i.e., entities with a big number of incoming links). The limit for training set is higher (500), since we want it to be representative of the real mention per entity distribution in Wikipedia. The effect of imposing this limits can be observed in Fig. 2 for both *continual* as well as *new* entities represented by a significant leap in the mentions-per-entity curve, particularly noticeable for validation and test subsets.

   Value to create TempEL: 10 for validation and test subsets, 500 for the train subset.

   Percentage of filtered out instances: for *continual* instances, 84% for validation and test subsets and 28% for the train subset. For *new* instances, 45% for validation and test subsets and 0.3% for the train subset.

   Hyperparameter name: `max_mens_per_ent` (see Table 1 in Section A.4).

10. **Inter-snapshot subsampling**: finally, we enforce that the number of continual and new entities as well as the number of mentions stays the same across the temporal snapshots (see Table 1). We achieve this by performing a random mention subsampling in snapshots with higher number of mentions, weighted by the difference in the number of mentions-per-entity. This produces a very similar mention-entity distribution across the temporal snapshots (see Section A.3 for further details).

    Percentage of filtered out instances: between 5% and 35%, it increases for more recent temporal snapshots as they have more instances in Wikipedia.

We do not filter on any attribute that could potentially produce evident biases in TempEL (e.g., gender, geographic location of the entities, etc.).

What data does each instance consist of? Each instance of a snapshot consists of:

1. Cleaned contextual text surrounding the anchor mention from the Wikipedia snapshot. Furthermore, we include the bert-tokenized version of the text used in our baseline.
2. Cleaned textual description of the target entity taken from the Wikipedia snapshot. Furthermore, we include the bert-tokenized version of the text used in our baseline.
3. A set of additional attributes defining the anchor mention and target entity.

For more details about the attributes, see Section A.6. Furthermore, concrete examples of TempEL's instances are showcased in Section A.10.

Is there a label or target associated with each instance? Yes, the target entity is represented by the Wikipedia page id. Furthermore, we also pair it with Wikidata QID of the corresponding Wikidata entity. These targets correspond to the attributes `target_page_id` and `target_qid` described in Table 2 (see Section A.6 for further details).

Is any information missing from individual instances? No, all the instances should have a complete information corresponding to the content as well as to the attributes.

Are relationships between individual instances made explicit? Yes, the relations between each of the instances and the target entity are made explicit by means of `target_page_id` and `target_qid` attributes (see Section A.6 for further details), which uniquely identify the id of the Wikipedia page describing a particular entity and the Wikidata entity respectively.

Are there recommended data splits (e.g., training, development/validation, testing)? Yes, the dataset is divided in train, validation and test subsets (see Table 1 for the distribution).

Are there any errors, sources of noise, or redundancies in the dataset? We have taken multiple measures to build a high quality dataset, minimizing the number of noise or other errors (see Section 3.2 of the main manuscript). Yet, TempEL is not 100% error free, and contains a few errors mostly due to erroneous Wikitext edits by the Wikipedia users.

Is the dataset self-contained, or does it link to or otherwise rely on external resources? Yes, the dataset is self contained and consists of:

1. Instances divided in train, validation and test subsets (see Table 1).
2. A description of all the entities of each of the Wikipedia snapshots. These entities form the complete candidate pool used by the models to predict the correct target entity. Figure 4c of the main manuscript illustrates the temporal evolution in size of the number of candidate entities.

Does the dataset contain data that might be considered confidential? No, Wikipedia is a public resource.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No, we haven't detected instances of such characteristics in TempEL.

Does the dataset identify any subpopulations (e.g., by age, gender)? While there are articles on different subpopulations on Wikipedia, there is no emphasis of the dataset on identifying or annotating those.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? Only based on their Wikipedia article, no editor information is retained.

Does the dataset contain data that might be considered sensitive in any way? Wikipedia is overall a resource aiming to be factual, therefore we can exclude this concern for most instances of TempEL.

A.2.3 Collection process

How was the data associated with each instance acquired? The textual data of the context of anchor mention and that of the description of the target entity is directly taken from the Wikipedia snapshots. Conversely, the attributes associated with each of the instances are calculated (see Section A.6 for further details).

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? The dataset was collected using the Wikipedia dumps from February of 2022. We detail further on the aspects related to the preprocessing, cleaning and labeling of TempEL instances in Section A.2.4 of the datasheet.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? The dataset was automatically generated based on existing Wikipedia articles. Therefore, no human intervention was needed for the dataset generation.

Over what timeframe was the data collected? The TempEL dataset was collected from 10 yearly snapshots of Wikipedia starting from January 1, 2013 until January 1, 2022.

Were any ethical review processes conducted (e.g., by an institutional review board)? N/A

A.2.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? The Wikipedia history logs content is available exclusively in Wikitext markup format.[2] In order to obtain cleaned text we proceed as follows:

1. We use MediaWiki API to process the templates which can not be parsed using regular expressions. For example, this is the case of the Wikitext template `Convert`, where the markup like "`{{convert|37|mm|in|abbr=on}}`" is converted to "`1.5 in`".
2. We use regular expressions to extract mentions and links. While this can also be done using online Wikitext parsing tools, we found that these did not account for all the corner cases of mention parsing such as the ones involving the *pipe trick*.[3]
3. Finally, we use `mwparserfromhell`[4] tool for parsing the rest of the Wikitext content.

Furthermore, our dataset files also contain BERT tokenization of the context around the mentions as well as the textual content of entities.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Yes, the raw data containing the Wikipedia history logs was saved on our cloud server in the following link: `https://cloud.ilabt.imec.be/index.php/s/BF9SkmQG2Tdjw8o`.

[2] `https://en.wikipedia.org/wiki/Help:Wikitext`

[3] `https://en.wikipedia.org/wiki/Help:Pipe_trick`

[4] `https://github.com/earwig/mwparserfromhell`

Is the software that was used to preprocess/clean/label the data available? Yes, the software will be made public upon acceptance.

A.2.5 Uses

Has the dataset been used for any tasks already? Yes, in our submitted manuscript we describe a retriever bi-encoder baseline [11] (see Section 4.2).

Is there a repository that links to any or all papers or systems that use the dataset? N/A

What (other) tasks could the dataset be used for? The covered task is temporally evolving entity linking.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? N/A

Are there tasks for which the dataset should not be used? N/A

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes, the dataset is of public access.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? The TempEL dataset will be made public on a GitHub repository together with the code to generate it. The baseline code and models will also be made public on the same repository. Due to the size, the dataset files will be hosted on the cloud server that belongs to Internet Technology and Data Science Lab (IDLab) at Ghent University (`https://cloud.ilabt.imec.be/index.php/s/RinXy8NgqdW58RW`).

When will the dataset be distributed? The dataset will be publicly distributed upon the submission of the camera ready version of our manuscript.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The TempEL dataset will be distributed under Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0).

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? N/A

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? N/A

A.2.6 Maintenance

Who will be supporting/hosting/maintaining the dataset? The maintenance and extension of TempEL will be carried out by the authors of the paper. Additionally, we will make the code publicly available to create potential new variations of TempEL using a number of hyperparameters (see Section A.4 and Section A.5 for further details).

The dataset files will be hosted on the cloud server that belongs to Internet Technology and Data Science Lab (IDLab) at Ghent University (`https://cloud.ilabt.imec.be/index.php/s/RinXy8NgqdW58RW`).

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The owners of the dataset can be contacted at the following e-mail address: `klim.zaporojets@ugent.be`.

Is there an erratum? No, there is no erratum yet.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? The TempEL will be regularly updated with newer snapshots (see Section A.5). In circumstances such as labeling errors, we will release the fixed version of the dataset with the respective version number. The introduction of the new version will be communicated using the TempEL GitHub repository.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? N/A

Will older versions of the dataset continue to be supported/hosted/maintained? Yes, the older version of the dataset will continue to be supported and hosted. All the versions will be numbered and we will provide the link to access each of these versions on our cloud storage server.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, we provide the code and functionality to re-generate and extend the dataset with new temporal snapshots (see Sections A.4 and A.5). Yet, it is the responsibility of the users to provide hosting and maintenance to the newly generated dataset variations.

A.3 Mentions per entity distribution

Figure 2 illustrates the similarity of mention per entity distribution across the temporal snapshots. This is achieved using weighted random subsampling so all the snapshots have equal number of instances (see *Data Distributor* component description in Section 3.1). By enforcing this similarity between temporal snapshots, we ensure that the potential difference in the results is independent of cross-snapshot dataset distributional variations and only influenced by the dynamic temporal evolution of the content in TempEL.

A.4 Dataset creation hyperparameters

Table 1 summarizes the hyperparameters that can be tuned in order to automatically create the TempEL dataset. This way, it is possible for the user to create different variation of the TempEL. The most relevant hyperparameter is `snapshots` that is used to specify the temporal intervals to create the snapshots. Below we detail two possible options we provide to specify such intervals.

Option 1 - explicit snapshot specification The user is expected to provide a list of timestamps in the format of `YYYY-MM-DDTHH:MM:SSZ`, each one defining a different snapshot.

Option 2 - time span and interval This option enables the user to define start and end dates of the time span from which the snapshots should be extracted. Furthermore, the interval value (i.e., by using keywords such as “weekly” or specifying the interval in seconds) has to also be specified.

A.5 Dataset extension

Additionally, we provide the option to extend the already existing dataset with new snapshots. Similarly as in the creation of new dataset (see Section A.4 above), the `snapshots` hyperparameter is used to specify new snapshots which are then added to already existing TempEL dataset.

A.6 Mention and entity attributes

[5] For train, validation and test sets respectively.

Table 1: Hyperparameters that can be tuned during TempEL dataset creation.

Hyperparamter	Description	TempEL
`snapshots`	Details (e.g., timestamps) of the temporal snapshots to be generated.	10 years
`nr_ct_ents_per_cut`	Number of *continual* entities per snapshot.	10,000
`min_mens_per_ent`	Minimum number of links a particular mention needs to have to target entity in order to be considered to be added in TempEL.	1
`min_men_prior`	Minimum mention prior (see Eq. (1) in the main manuscript).	0.0001
`max_men_prior`	Maximum mention prior.	0.5
`min_men_prior_rank`	Minimum rank of mention prior among all the mentions pointing to a specific entity.	2
`min_ent_prior`	Minimum entity prior as defined in [12]: the ratio of links to the entity with respect to all of the links in the Wikipedia snapshot.	0.0
`max_ent_prior`	Maximum entity prior.	1.0
`min_nr_inlinks`	Minimum number of incoming links per entity.	10
`min_len_target_ent`	Minimum length of target entity page (in tokens).	10
`max_mens_per_ent`	Maximum number of mentions per entity.	500/10/10[5]
`ed_men_title`	Minimum normalized edit distance between the mentions and the title of the target page they are linked to.	0.2
`ed_men_subsets`	Minimum normalized edit distance between the mentions in different subsets linked to the same target entity.	0.2
`stable_interval`	In seconds, the interval of time before the end of each snapshot from which the most stable version of Wikipedia has to be taken (see Section 3.2 for further details).	2,592,000 (30 days)
`equal_snapshots`	Whether the number of instances and the number of mentions per entity distribution is the same across the snapshots (see Section 3.2 for further details). Equal cross-snapshot mention per entity distribution in Fig. 2 is the result of setting this hyperparameter in True.	True

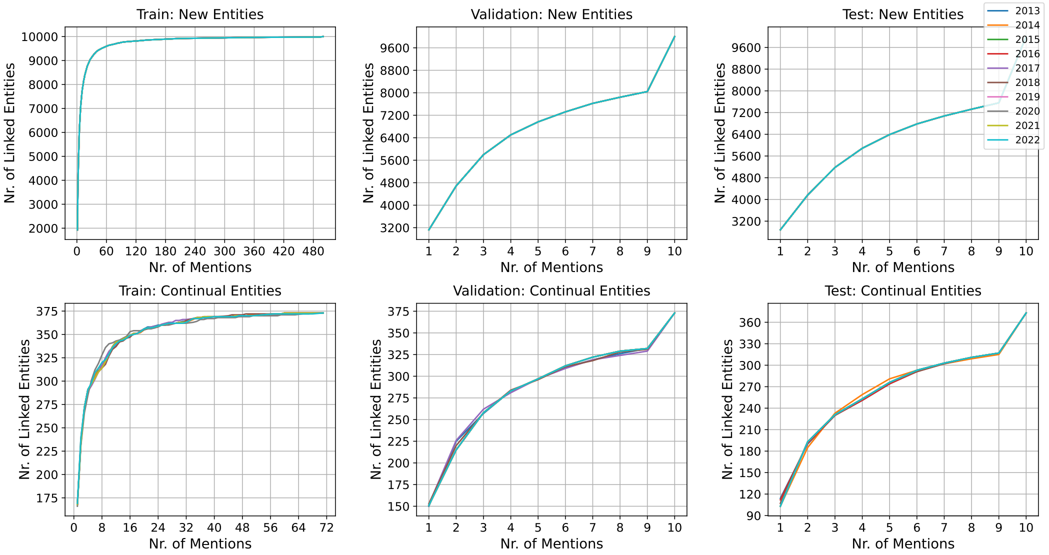

Figure 2: Similar distribution of the data across the temporal snapshots (number of mentions per entity). This structurally unbiased setting enable to study exclusively the temporal effect on the performance of the models for each of the different time periods.

Table 2: Attributes associated to each of the mention-entity pairs for each of the temporal snapshots in TempEL.

Attribute	Description
`subset`	The name of current subset (i.e., train, validation or test).
`target_page_id`	The unique Wikipedia page id of the target entity.
`target_qid`	The unique Wikidata QID of the target entity.
`snapshot`	The timestamp of the temporal snapshot from which the anchor mention and target entity attributes were extracted.
`target`	The textual content of the target entity Wikipedia page.
`target_len`	The length in tokens of target Wikipedia page.
`target_title`	The title of target entity Wikipedia page.
`category`	Category of the target entity (*new* or *continual*).
`mention`	The text of the mention.
`context_left`	The textual context to the left of the mention.
`context_right`	The textual context to the right of the mention.
`anchor_len`	The length in tokens of the Wikipedia page where the anchor mention is located.
`ed_men_title`	Normalized edit distance between the anchor mention and the title of the target Wikipedia page.
`overlap_type`	Overlap type between the anchor mention and the target title as defined by [4].
`men_prior`	The mention prior (see Eq. (1) of the main manuscript).
`men_prior_rank`	The rank of the current anchor mention compared to other mentions in Wikipedia pointing to target entity.
`avg_men_prior`	The average value of prior of the mentions linked to the target entity in Wikipedia for `snapshot`.
`ent_prior`	Entity prior as defined in [12]: the ratio of links to the entity with respect to all of the links in the Wikipedia snapshot.
`nr_inlinks`	Total number of incoming links to target entity.
`nr_dist_mens`	Number of distinct (i.e., with different surface form) mentions linked to target entity.
`nr_mens_per_ent`	Number of times the current mention appears in Wikipedia linked to target entity.
`nr_mens_extracted`	Number of anchor mentions per current target entity in the `subset`.
`anchor_creation_date`	The creation date (timestamp) of Wikipedia page where the anchor mention is located.
`anchor_revision_date`	The timestamp of when the anchor Wikipedia page was last revised.
`target_creation_date`	The timestamp of when the target Wikipedia entity page was created.
`target_revision_date`	The timestamp of when the target Wikipedia entity page was last modified.

Table 2 describes the anchor mention and target entity related attributes present in TempEL. These attributes can be used to perform more in-depth analysis of the results.

A.7 Baseline implementation details

We base our bi-encoder baseline model on the publicly available BLINK code.[6] We train all the models for 10 epochs with the learning rate of 1e-04 and the batch size of 64. We use AdamW optimizer with 10% of warmup steps. Finally, we rely on `transformers` library [10] to get the pre-trained BERT-large representations. All the experiments were run on NVIDIA V100 GPU with the following execution times:

1. *Training*: 36 hours to train for 10 epochs per single snapshot.
2. *All Wikipedia entity encoding*: 7 days per finetuned model (on all the 10 Wikipedia snapshots) running on a single V100 GPU.
3. *Evaluation*: 30 seconds per finetuned model per snapshot using FAISS [3] library on GPU.

A.8 Total amount of compute and the type of resources used to create TempEL

In this section we provide the details on the computational resources used in each of the processing steps (see Section 3.1 and Fig. 2 for further details) to create the TempEL dataset:

1. *Snapshot Data Extraction*: this processing step is responsible for creating the snapshots from the Wikipedia log files from February 1, 2022. This is a multi-processing step that is executed on a cluster with 80 CPUs and 110 GB of RAM and takes 5 days and 8 hours to complete.
2. *Snapshot Dataset Building*: this is a multi-processing step that is executed on a cluster with 30 CPUs and 250 GB of RAM and takes 5 hours to complete.

A.9 License of the assets

We base the implementation of our baseline bi-encoder model on the publicly available BLINK [11] code. This asset is made available under MIT License (`https://opensource.org/licenses/MIT`).

A.10 Examples

This section presents two illustrative examples of instances in TempEL. The first example contains the anchor mention linked to *continual* entity, while the second one is the example of a link to *new* entity. Both of the examples were taken from the snapshot of January 1, 2021. Furthermore, we trim the content length (e.g., `target` attribute value) to only a few tokens for space reasons.

A.10.1 Example 1: continual target entity

Table 3 illustrates an example of the link to *continual* target entity *Sacramental_bread*. It is worth noting that the creation date of this entity in Wikipedia (`target_creation_date` attribute) is of January 3, 2005. Yet, the version saved in the snapshot (`target_revision_date` attribute) is from December 30, 2020.

[6] `https://github.com/facebookresearch/BLINK`

Table 3: Example of the instance corresponding to mention link to *continual* entity (Sacramental_bread created in 2005-01-03) in TempEL.

Attribute	Value
subset	train
target_page_id	1359030
target_qid	Q207104
snapshot	2021-01-01T00:00:00Z
target	“Sacramental bread, sometimes called altar bread, Communion ...”
target_len	7,568
target_title	“Sacramental_bread”.
category	continual
mention	“host”
context_left	“... devotional image, portrait or other religious symbol (such as the”
context_right	“). Garland paintings were typically collaborations between a ...”
anchor_len	6,519
ed_men_title	0.9411
overlap_type	LOW_OVERLAP
men_prior	0.0750
men_prior_rank	7
avg_men_prior	0.6864
ent_prior	1.7790e-6
nr_inlinks	225
nr_dist_mens	13
nr_mens_per_ent	79
nr_mens_extracted	58
anchor_creation_date	2009-09-25T21:09:07Z
anchor_revision_date	2020-10-04T16:15:13Z
target_creation_date	2005-01-03T17:41:14Z
target_revision_date	2020-12-30T12:38:50Z

Table 4: Example of the instance corresponding to mention link to *new* entity (COVID-19_pandemic_in_Portland,_Oregon created in 2020-03-23) in TempEL.

Attribute	Value
`subset`	train
`target_page_id`	63449958
`target_qid`	Q88484856
`snapshot`	2021-01-01T00:00:00Z
`target`	“The COVID-19 pandemic was confirmed to have reached ...”
`target_len`	26,432
`target_title`	“COVID-19_pandemic_in_Portland,_Oregon”
`category`	new
`mention`	“COVID-19 pandemic”
`context_left`	“Xico Xico and Xica both offered pickup service during the”
`context_right`	“, as of May 2020. ”
`anchor_len`	2,437
`ed_men_title`	0.5405
`overlap_type`	AMBIGUOUS_SUBSTRING
`men_prior`	0.0009
`men_prior_rank`	4
`avg_men_prior`	0.2548
`ent_prior`	2.9255e-7
`nr_inlinks`	37
`nr_dist_mens`	3
`nr_mens_per_ent`	23
`nr_mens_extracted`	18
`anchor_creation_date`	2020-12-08T00:23:50Z
`anchor_revision_date`	2020-12-09T15:41:18Z
`target_creation_date`	2020-03-23T04:22:55Z
`target_revision_date`	2020-11-16T03:59:06Z

A.10.2 Example 2: new target entity

Table 4 illustrates an example of the link to *new* target entity *COVID-19_pandemic_in_Portland,_Oregon*. It is worth noting that the creation date of this entity in Wikipedia (`target_creation_date` attribute) is of March 23, 2020, which belongs to the interval of the considered snapshot: from January 1, 2020 until January 1, 2021.

A.11 Additional results

Tables 5-11 present the results for different accuracy@K for $K \in \{1, 2, 4, 8, 16, 32, 64\}$. Furthermore, Fig. 3 illustrates the mean in- and out-of-snapshot (see Section 4.2 of the main manuscript) accuracy@K performance across temporal snapshots on the following four target entity categories:

1. *Continual*: all the target *continual* entities (i.e., the entities that exist across all the temporal snapshots in TempEL dataset).

2. *COVID-19*: target *new* entities that have COVID-related (e.g., “COVID”, “coronavirus”, etc.) terms in the target entity title.

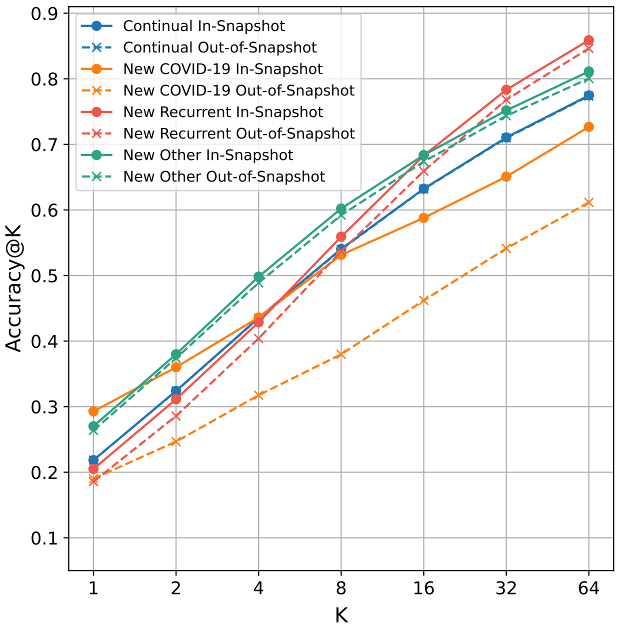

Figure 3: Accuraccy@K for different values of $K \in \{1, 2, 4, 8, 16, 32, 64\}$. The results are grouped in four main categories: (i) mentions linked to *continual* entities that exist in all of the TempEL snapshots, (ii) mentions linked to *COVID-related* new entities (i.e., with keywords such as "COVID" in target entity title), (iii) mentions linked to *recurrent* new entities (i.e., entities representing events occurring periodically such as "2018 BNP Paribas Open"), and (iv) mentions linked to *other* new entities.

3. *Recurrent*: target *new* entities whose titles contain the year and some of the keywords (e.g., "league", "election", "cup", etc.) that indicate that an entity is a repetitive event (e.g., "2018 BNP Paribas Open" which is part of *yearly* BNB Paribas Open competitions).
4. *Other*: all the other target *new* entities.

The following are the main conclusions that can be drawn from the graph in Fig. 3 that support or complement the findings described in Section 4.2 of the main manuscript:

1. New entities that require fundamentally new, previously non-existent knowledge to be disambiguated tend to have the lowest out-of-snapshot performance. This is the case of COVID-19 related disambiguation instances. These instances also experience the highest boost in performance when evaluated on in-snapshot setting (i.e., the model is evaluated and finetuned on the same temporal snapshot).
2. The difference between in- and out-of-snapshot performances on *continual* entities is the lowest. This is also supported by Fig. 4b and Figs. 5a–5b in the main manuscript. This suggests that the actual knowledge needed to disambiguate most of the *continual* entities in TempEL changes very little with time.
3. The model has the highest accuracy@64 performance on *recurrent* new entities. Yet, the performance on these entities drops sharply for lower values of K. We hypothesize that predicting the correct recurrent event gets more challenging as K decreases because of the large number of very similar candidates to pick from (e.g., many "BNP Paribas Open" championships that only differ in very few details such as the date).
4. The difference between in- and out-of-snapshot performance for *other* new entities is lower than for *recurrent* and *COVID-19* related ones. This is driven by new entities that are derived from existing entities in Wikipedia (i.e., their content is a copy of already established entities). We hypothesize that the model requires little additional knowledge to disambiguate these entities. Still, it is part of future work to study *other* new entities more in detail in order

to find cases that represent intrinsically new knowledge similar to the identified COVID-19 entity cluster.

Table 5: **Accuracy@1** for *continual* (top) and *new* (bottom) entities. The intensity of colors is set on a row-by-row basis and indicates whether performance is **better** or **worse** compared to the year the model was finetuned on (i.e., the values that form the white diagonal).

Continual Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.225	0.219	0.215	0.217	0.212	0.206	0.203	0.203	0.197	0.192
2014	0.229	0.226	0.220	0.221	0.217	0.212	0.211	0.207	0.203	0.197
2015	0.228	0.223	0.219	0.219	0.216	0.211	0.208	0.206	0.204	0.196
2016	0.230	0.227	0.222	0.221	0.218	0.214	0.211	0.208	0.205	0.199
2017	0.240	0.237	0.229	0.229	0.226	0.221	0.219	0.216	0.211	0.207
2018	0.238	0.236	0.228	0.229	0.226	0.222	0.219	0.217	0.211	0.206
2019	0.237	0.235	0.228	0.228	0.226	0.220	0.217	0.216	0.212	0.208
2020	0.232	0.227	0.223	0.221	0.219	0.214	0.210	0.209	0.205	0.199
2021	0.239	0.235	0.231	0.230	0.228	0.222	0.219	0.217	0.213	0.210
2022	0.238	0.235	0.229	0.229	0.226	0.222	0.218	0.218	0.214	0.206

New Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.280	0.226	0.253	0.203	0.230	0.198	0.226	0.144	0.168	0.212
2014	0.291	0.268	0.258	0.201	0.234	0.217	0.245	0.150	0.159	0.214
2015	0.252	0.206	0.206	0.181	0.194	0.179	0.210	0.139	0.174	0.193
2016	0.277	0.248	0.242	0.214	0.221	0.206	0.226	0.144	0.181	0.206
2017	0.271	0.226	0.223	0.176	0.230	0.201	0.219	0.144	0.173	0.204
2018	0.284	0.255	0.240	0.190	0.228	0.268	0.246	0.157	0.178	0.222
2019	0.278	0.243	0.237	0.177	0.223	0.230	0.230	0.130	0.174	0.203
2020	0.284	0.236	0.225	0.206	0.214	0.201	0.212	0.183	0.177	0.221
2021	0.291	0.236	0.232	0.195	0.219	0.229	0.230	0.183	0.214	0.217
2022	0.294	0.260	0.251	0.188	0.206	0.241	0.240	0.170	0.170	0.219

Table 6: **Accuracy@2** for *continual* (top) and *new* (bottom) entities. The intensity of colors is set on a row-by-row basis and indicates whether performance is **better** or **worse** compared to the year the model was finetuned on (i.e., the values that form the white diagonal).

Continual Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.337	0.330	0.324	0.322	0.317	0.311	0.306	0.302	0.301	0.293
2014	0.339	0.335	0.329	0.328	0.322	0.317	0.314	0.310	0.306	0.299
2015	0.339	0.333	0.327	0.325	0.323	0.317	0.312	0.309	0.305	0.299
2016	0.341	0.334	0.328	0.326	0.322	0.316	0.314	0.310	0.306	0.301
2017	0.351	0.346	0.338	0.338	0.332	0.328	0.324	0.320	0.316	0.309
2018	0.348	0.342	0.336	0.334	0.331	0.327	0.323	0.322	0.315	0.309
2019	0.348	0.345	0.337	0.335	0.332	0.325	0.322	0.320	0.317	0.310
2020	0.341	0.336	0.330	0.327	0.322	0.316	0.312	0.310	0.307	0.300
2021	0.349	0.344	0.338	0.335	0.331	0.325	0.321	0.319	0.315	0.310
2022	0.348	0.343	0.336	0.336	0.331	0.325	0.321	0.320	0.317	0.309

New Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.401	0.322	0.359	0.310	0.327	0.309	0.340	0.266	0.236	0.291
2014	0.397	0.366	0.357	0.318	0.328	0.347	0.357	0.278	0.234	0.306
2015	0.358	0.305	0.319	0.277	0.276	0.294	0.304	0.265	0.249	0.272
2016	0.379	0.351	0.345	0.344	0.308	0.320	0.315	0.270	0.244	0.311
2017	0.372	0.328	0.340	0.290	0.317	0.313	0.339	0.266	0.250	0.294
2018	0.395	0.369	0.346	0.305	0.326	0.380	0.344	0.270	0.250	0.306
2019	0.397	0.363	0.346	0.296	0.303	0.344	0.341	0.250	0.249	0.294
2020	0.385	0.343	0.337	0.321	0.294	0.323	0.319	0.301	0.250	0.315
2021	0.392	0.338	0.346	0.303	0.308	0.334	0.333	0.301	0.286	0.307
2022	0.408	0.372	0.355	0.301	0.294	0.352	0.336	0.289	0.250	0.322

Table 7: **Accuracy@4** for *continual* (top) and *new* (bottom) entities. The intensity of colors is set on a row-by-row basis and indicates whether performance is **better** or **worse** compared to the year the model was finetuned on (i.e., the values that form the white diagonal).

Continual Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.449	0.442	0.439	0.433	0.428	0.422	0.417	0.417	0.411	0.405
2014	0.455	0.448	0.443	0.439	0.433	0.428	0.424	0.423	0.416	0.410
2015	0.455	0.446	0.444	0.438	0.434	0.427	0.422	0.422	0.415	0.409
2016	0.453	0.446	0.442	0.437	0.432	0.426	0.422	0.422	0.415	0.408
2017	0.464	0.458	0.454	0.448	0.443	0.438	0.434	0.433	0.428	0.423
2018	0.461	0.453	0.449	0.445	0.440	0.437	0.430	0.431	0.425	0.417
2019	0.462	0.455	0.452	0.446	0.443	0.437	0.433	0.434	0.427	0.421
2020	0.455	0.446	0.442	0.438	0.433	0.427	0.422	0.423	0.417	0.411
2021	0.461	0.454	0.450	0.445	0.440	0.434	0.429	0.428	0.423	0.416
2022	0.460	0.453	0.450	0.444	0.440	0.433	0.429	0.430	0.424	0.417

New Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.512	0.442	0.479	0.429	0.426	0.421	0.455	0.392	0.328	0.397
2014	0.526	0.486	0.475	0.434	0.446	0.463	0.489	0.410	0.317	0.406
2015	0.479	0.414	0.452	0.401	0.372	0.403	0.430	0.389	0.337	0.377
2016	0.500	0.464	0.466	0.463	0.418	0.434	0.430	0.408	0.330	0.414
2017	0.507	0.448	0.452	0.401	0.428	0.445	0.474	0.394	0.328	0.408
2018	0.520	0.487	0.477	0.428	0.435	0.496	0.469	0.388	0.340	0.417
2019	0.517	0.486	0.482	0.419	0.415	0.475	0.472	0.398	0.339	0.403
2020	0.506	0.449	0.457	0.418	0.414	0.443	0.443	0.414	0.331	0.428
2021	0.509	0.453	0.457	0.422	0.421	0.446	0.439	0.417	0.383	0.427
2022	0.527	0.491	0.472	0.439	0.397	0.471	0.474	0.422	0.341	0.434

Table 8: **Accuracy@8** for *continual* (top) and *new* (bottom) entities. The intensity of colors is set on a row-by-row basis and indicates whether performance is **better** or **worse** compared to the year the model was finetuned on (i.e., the values that form the white diagonal).

Continual Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.556	0.551	0.546	0.539	0.532	0.526	0.520	0.520	0.513	0.507
2014	0.563	0.559	0.553	0.546	0.540	0.532	0.527	0.526	0.520	0.514
2015	0.561	0.555	0.552	0.543	0.540	0.533	0.526	0.526	0.520	0.514
2016	0.559	0.554	0.550	0.542	0.537	0.531	0.524	0.524	0.518	0.511
2017	0.569	0.565	0.562	0.555	0.549	0.542	0.537	0.537	0.530	0.525
2018	0.567	0.561	0.558	0.550	0.544	0.537	0.532	0.531	0.523	0.519
2019	0.571	0.565	0.562	0.554	0.550	0.541	0.537	0.537	0.529	0.524
2020	0.561	0.555	0.553	0.545	0.539	0.532	0.527	0.528	0.522	0.515
2021	0.565	0.559	0.557	0.548	0.544	0.535	0.530	0.530	0.524	0.519
2022	0.566	0.560	0.556	0.549	0.545	0.537	0.532	0.533	0.527	0.521

New Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.632	0.572	0.585	0.563	0.541	0.539	0.574	0.517	0.425	0.504
2014	0.633	0.624	0.586	0.565	0.569	0.561	0.599	0.538	0.421	0.531
2015	0.603	0.541	0.559	0.524	0.495	0.534	0.562	0.510	0.425	0.497
2016	0.626	0.608	0.600	0.586	0.532	0.572	0.567	0.526	0.428	0.526
2017	0.617	0.570	0.567	0.532	0.534	0.567	0.587	0.528	0.435	0.517
2018	0.634	0.606	0.585	0.566	0.559	0.611	0.594	0.527	0.449	0.526
2019	0.651	0.621	0.601	0.536	0.536	0.590	0.605	0.523	0.459	0.526
2020	0.633	0.582	0.574	0.553	0.533	0.548	0.563	0.540	0.434	0.543
2021	0.637	0.584	0.577	0.555	0.531	0.565	0.571	0.549	0.492	0.546
2022	0.646	0.632	0.593	0.554	0.504	0.581	0.591	0.541	0.433	0.556

Table 9: **Accuracy@16** for *continual* (top) and *new* (bottom) entities. The intensity of colors is set on a row-by-row basis and indicates whether performance is **better** or **worse** compared to the year the model was finetuned on (i.e., the values that form the white diagonal).

Continual Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.648	0.643	0.639	0.632	0.626	0.617	0.613	0.613	0.605	0.600
2014	0.657	0.650	0.647	0.639	0.635	0.627	0.622	0.620	0.613	0.608
2015	0.651	0.645	0.642	0.636	0.633	0.624	0.619	0.619	0.612	0.608
2016	0.652	0.646	0.643	0.637	0.631	0.621	0.616	0.615	0.610	0.605
2017	0.660	0.655	0.652	0.646	0.640	0.633	0.628	0.628	0.621	0.618
2018	0.656	0.651	0.647	0.642	0.636	0.627	0.624	0.622	0.614	0.611
2019	0.662	0.658	0.653	0.646	0.642	0.633	0.630	0.630	0.622	0.618
2020	0.652	0.647	0.644	0.636	0.632	0.622	0.619	0.619	0.612	0.608
2021	0.655	0.650	0.648	0.641	0.635	0.627	0.624	0.622	0.615	0.611
2022	0.657	0.651	0.647	0.641	0.637	0.630	0.625	0.625	0.619	0.614

New Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.748	0.690	0.686	0.690	0.648	0.647	0.676	0.627	0.526	0.612
2014	0.761	0.730	0.691	0.681	0.661	0.670	0.706	0.641	0.522	0.625
2015	0.727	0.661	0.677	0.660	0.606	0.629	0.664	0.610	0.530	0.589
2016	0.746	0.701	0.712	0.701	0.647	0.662	0.670	0.621	0.514	0.629
2017	0.733	0.681	0.686	0.659	0.666	0.662	0.691	0.637	0.539	0.614
2018	0.759	0.701	0.697	0.670	0.665	0.705	0.694	0.643	0.539	0.624
2019	0.761	0.714	0.702	0.673	0.656	0.690	0.696	0.633	0.559	0.634
2020	0.746	0.683	0.678	0.677	0.632	0.654	0.661	0.650	0.538	0.640
2021	0.750	0.689	0.687	0.670	0.636	0.667	0.667	0.650	0.582	0.648
2022	0.760	0.726	0.692	0.676	0.630	0.672	0.690	0.637	0.536	0.649

Table 10: **Accuracy@32** for *continual* (top) and *new* (bottom) entities. The intensity of colors is set on a row-by-row basis and indicates whether performance is **better** or **worse** compared to the year the model was finetuned on (i.e., the values that form the white diagonal).

Continual Entities

Test Train	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.723	0.719	0.716	0.710	0.705	0.697	0.693	0.692	0.687	0.682
2014	0.731	0.727	0.723	0.717	0.714	0.706	0.702	0.702	0.695	0.690
2015	0.727	0.723	0.721	0.714	0.710	0.703	0.700	0.699	0.693	0.688
2016	0.726	0.721	0.719	0.713	0.709	0.700	0.696	0.696	0.692	0.687
2017	0.734	0.730	0.726	0.722	0.718	0.710	0.706	0.706	0.701	0.696
2018	0.732	0.727	0.724	0.719	0.714	0.707	0.702	0.701	0.697	0.693
2019	0.736	0.731	0.727	0.723	0.718	0.711	0.708	0.707	0.703	0.698
2020	0.727	0.722	0.719	0.714	0.711	0.703	0.699	0.699	0.693	0.689
2021	0.728	0.724	0.721	0.715	0.712	0.705	0.702	0.701	0.696	0.691
2022	0.730	0.726	0.723	0.717	0.714	0.707	0.704	0.703	0.698	0.695

New Entities

Test Train	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.839	0.763	0.778	0.763	0.752	0.736	0.763	0.718	0.626	0.686
2014	0.852	0.794	0.791	0.767	0.756	0.765	0.788	0.736	0.635	0.701
2015	0.835	0.756	0.774	0.763	0.711	0.727	0.760	0.706	0.632	0.701
2016	0.848	0.771	0.801	0.779	0.756	0.759	0.765	0.722	0.633	0.709
2017	0.845	0.760	0.788	0.754	0.763	0.747	0.779	0.716	0.638	0.710
2018	0.847	0.776	0.785	0.766	0.760	0.788	0.778	0.735	0.645	0.726
2019	0.856	0.786	0.786	0.764	0.765	0.769	0.785	0.740	0.669	0.713
2020	0.850	0.771	0.775	0.771	0.747	0.751	0.763	0.746	0.642	0.734
2021	0.852	0.771	0.774	0.757	0.734	0.749	0.768	0.743	0.676	0.741
2022	0.852	0.797	0.784	0.759	0.752	0.752	0.780	0.739	0.643	0.733

Table 11: **Accuracy@64** for *continual* (top) and *new* (bottom) entities. The intensity of colors is set on a row-by-row basis and indicates whether performance is **better** or **worse** compared to the year the model was finetuned on (i.e., the values that form the white diagonal).

Continual Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.785	0.782	0.778	0.772	0.769	0.762	0.758	0.758	0.754	0.750
2014	0.792	0.790	0.785	0.781	0.777	0.771	0.767	0.767	0.763	0.760
2015	0.786	0.784	0.782	0.777	0.773	0.769	0.765	0.764	0.760	0.757
2016	0.789	0.784	0.781	0.777	0.773	0.768	0.763	0.763	0.758	0.755
2017	0.794	0.791	0.788	0.785	0.781	0.775	0.771	0.772	0.768	0.763
2018	0.791	0.788	0.786	0.782	0.778	0.773	0.769	0.769	0.764	0.760
2019	0.795	0.792	0.789	0.784	0.781	0.776	0.772	0.773	0.767	0.765
2020	0.787	0.783	0.782	0.777	0.774	0.768	0.765	0.765	0.761	0.756
2021	0.788	0.785	0.782	0.777	0.773	0.769	0.764	0.764	0.761	0.757
2022	0.790	0.787	0.783	0.779	0.776	0.771	0.768	0.768	0.764	0.760

New Entities

Test **Train**	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2013	0.910	0.819	0.853	0.826	0.841	0.812	0.819	0.791	0.688	0.774
2014	0.908	0.848	0.862	0.827	0.843	0.832	0.842	0.814	0.704	0.791
2015	0.898	0.823	0.849	0.822	0.808	0.813	0.832	0.788	0.706	0.781
2016	0.897	0.832	0.862	0.832	0.839	0.823	0.823	0.802	0.718	0.791
2017	0.906	0.832	0.857	0.817	0.840	0.824	0.835	0.791	0.714	0.808
2018	0.908	0.835	0.858	0.830	0.846	0.853	0.835	0.806	0.728	0.803
2019	0.910	0.842	0.853	0.821	0.842	0.843	0.841	0.810	0.734	0.799
2020	0.903	0.828	0.844	0.835	0.843	0.819	0.833	0.817	0.728	0.811
2021	0.910	0.825	0.852	0.825	0.837	0.817	0.830	0.814	0.761	0.812
2022	0.905	0.846	0.852	0.820	0.830	0.830	0.832	0.808	0.732	0.823

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[2] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 782–792, 2011.

[3] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

[4] Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 2019 Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 3449–3460, 2019.

[5] Yasumasa Onoe and Greg Durrett. Fine-grained entity typing for domain independent entity linking. In *Proceedings of the 2020 Conference on Artificial Intelligence (AAAI 2020)*, pages 8576–8583, 2020.

[6] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021)*, 2021.

[7] Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. GERBIL–benchmarking named entity recognition and linking consistently. *Semantic Web*, 9(5):605–625, 2018.

[8] Özge Sevgili, Artem Shelmanov, Mikhail Y. Arkhipov, Alexander Panchenko, and Chris Biemann. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570, 2022.

[9] Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, et al. GERBIL: general entity annotator benchmarking framework. In *Proceedings of the 2015 International Conference on World Wide Web (WWW 2015)*, pages 1133–1143, 2015.

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2020)*, pages 38–45, 2020.

[11] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 6397–6407, 2020.

[12] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 2016 SIGNLL Conference on Computational Natural Language Learning (CoNLL 2016)*, pages 250–259, 2016.