A Large Test Collection for Entity Aspect Linking

Jordan Ramsdell
jsc57@wildcats.unh.edu
University of New Hampshire
Durham, NH, USA

Laura Dietz
dietz@cs.unh.edu
University of New Hampshire
Durham, NH, USA

ABSTRACT

Given a text with entity links, the task of entity aspect linking is to identify which aspect of an entity is referred to in the context. For example, if a text passage mentions the entity "USA", is USA mentioned in the context of the 2008 financial crisis, American cuisine, or else? Complementing efforts of Nanni et al (2018), we provide a large-scale test collection which is derived from Wikipedia hyperlinks in a dump from 01/01/2020. Furthermore, we offer strong baselines with results and broken-out feature sets to stimulate more research in this area.

Data, code, feature sets, runfiles and results are released under a CC-SA license and offered on our aspect linking resource web page **http://www.cs.unh.edu/~dietz/eal-dataset-2020/**.

ACM Reference Format:
Jordan Ramsdell and Laura Dietz. 2020. A Large Test Collection for Entity Aspect Linking. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3340531.3412875

1 INTRODUCTION

We provide a large-scale dataset for training and evaluating methods for entity aspect linking (EAL), a fine-grained variation on entity linking that discerns which particular aspect of an entity is mentioned in the context.

In contrast to entity aspect linking (EAL), the task of entity linking is to identify and link mentions of entities to their knowledge graph entries. Entity linking provides machine with access to the meaning of text. The availability of entity linking tools is central in the advent of knowledge graphs, semantic annotations, information extraction, information retrieval, and knowledge management. Entity linking is especially useful whenever users need to identify entity-centric information in large quantities of text, such as important people or events pertaining to a topic of interest. However, entity links do not necessarily inform users that they are relevant to a particular topic. While entities are treated as atomic units, and the meaning of an entity can depend on the context in which it occurs. By annotating an entity with an "aspect" that describes which facet of an entity is referred to in its surrounding text, we can provide downstream algorithms. To assist in this task, we provide a large test collection consisting of 1 million entity aspect link instances, with strong baselines and example feature sets.

Many entity linking systems use Wikipedia articles as a general-purpose definition of entities. Analogously, we use Wikipedia to derive aspects for each entity.

Entity Aspect Linking (EAL) Task. Building on the definition of Nanni et al. [12], we formalize the task as a refinement of entity linking as follows:

- Given a paragraph-sized text passage t with entity links to entities $e_1, e_2, ..e_n$.
- For each entity e_i, a catalog of candidate aspects a_{i1}, a_{i2}, $\ldots a_{im}$ is available with name, content, and entity links.
- The task is to predict for each entity e_i the correct aspect a_{ij} that is mentioned in the context t.

Our definition centers on a catalog of candidate aspects to be available for each entity. Here we follow Nanni et al. and use the top-level sections of an entity's Wikipedia page to automatically derive a catalog of aspects, where each section represents one aspect. Administrative sections without topical nature such as "References" or "See Also" are excluded. We use a Wikipedia dump that is offered by the TREC Complex Answer Retrieval track, which exposes section and hyperlink information in a machine-readable format [8].[1] If necessary, entity links for a given text passage can be readily created with an entity linking tool.

Previously, Nanni et al. provided a dataset of 201 manually verified entity aspect links from a Wikipedia dump in 2016. In this work, we provide a much larger entity aspect linking dataset which is derived from a English Wikipedia dump of 2020.

Worked example. Consider the example passage depicted on the right of Figure 1 which mentions the entity "Oyster". In entity aspect linking, we wish to automatically annotate this mention of the target entity with one of its aspects—preferably the aspect that is most representative for the mention in context. In this example, the expression "shucking" indicates that oysters here are referred to in the context of being prepared for food. Often similar clues are found in contextual words and entities. The success of an aspect link prediction depends on 1) methods for understanding language semantics, such as idioms and synonyms, and how they relate to the content of an entity 2) the ability to distinguish which part of the context, *if any*, should match against the name and content of an aspect. The latter is particularly important as target entities are not always salient in the context they occur. For example, here the context is primarily about the Iceworm festival in Alaska, and knowing that shucking oysters relates to food is not particularly important.

[1] Wiki dump used here is available at trec-car.cs.unh.edu/datareleases/v2.4-release.html

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CIKM '20, October 19–23, 2020, Virtual Event, Ireland
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412875

Source Page: Cordova, Alaska
True Aspect: Oyster/As food

The Cordova **Iceworm** Festival takes place each February and is an effective way to thwart the **winter blues**. *Activities include a parade, talent show, royal crowning ceremony, and various competitions such as an* ***oyster shucking*** *contest,* ***ping pong*** *tournament, and a* ***survival suit*** *race.*

Target Entity: Oyster
Aspect: Oyster/As food

Jonathan Swift is quoted as having said, "He was a bold man that first ate an oyster". [...] Opening oysters, referred to as "oyster-shucking", requires skill. The preferred method is to use a special knife (called an **oyster knife**, a variant of a shucking knife), with a short and thick blade about 5 cm long.

Figure 1: Entity aspect linking example for target entity "Oyster". Left: Context paragraph with entity links from Wikipedia editors; context sentence denoted in italics. Right: True aspect for this link with name, content, and entity links. Other aspects in the catalog are Types, Habitat and Behavior, Ecosystem services, and Diseases. More examples are available online .

Potential Impact. We envision the results of entity aspect linking to be immediately useful for users who track entities in information streams, such as social media, news, and reports. While entity linking is helpful for this use case, the granularity of (whole) entities is often too coarse. Furthermore, some dominant aspects of an entity are much more commonly referred to than others. If a rare aspect is sought, relevant results are drowned out by references to the dominant aspects. Hence, much like rare words that occur in text, the less dominant aspects of an entity might be very informative in particular situations, but are easily overlooked, and their identification currently requires additional work.

One could train individual text classifiers for each entity and its aspects, however users would like to avoid the manual labor it requires. Entity aspect linking provides an alternative by training a universal model across a wide range of target entities and contexts, ready to make aspect link predictions for unseen target entities and unseen contexts. We provide a large test collection for training, that includes aspect catalogs derived from sections of the target entity's Wikipedia page.

In this paper we provide an evaluation paradigm, along with the results of a strong reference baseline, to stimulate research on entity aspect linking. Successful entity aspect linking methods can give rise to improvements on a range of important downstream applications:

- In reputation management, a company would like to know which aspects of its products are being widely discussed on social media. If there are only a few products, one can train a lexicalized classifier to solve this task. However, to track many different products with many newly emerging facets, entity aspect linking offers a flexible alternative.
- For information retrieval, entity-centric query expansion augments the original search query with words and entities of a relevance feedback run [5, 10, 17]. Unfortunately, entities that are mentioned in a large variety of contexts lower the retrieval quality, when spurious contexts are being matched. For example, while the entity USA is relevant for a query about the financial crisis of 2008, it will also match text that mentions USA in the context of fast food. With entity aspect linking, such spurious context matches can be discerned, which is likely to improve the retrieval quality.
- In weakly-supervised relation extraction approaches, a first step is to match training entity pairs to sentences [11, 15]. A common source of errors arises from the assumption that the majority of sentences that mention two related entities express the expected relation type. This source of error particularly affects the extraction of rare and specialized relation types, where this assumption is heavily violated. However, whenever relation types can be associated with entity aspects many false sentence matches can be avoided.
- Many useful knowledge graphs are derived from Wikipedia, where each page represents one node in the graph and links between pages represent edges. With entity aspect linking, an alternative fine-grained knowledge graph can be constructed where each aspect of an entity could represent a node in the graph, while links represent semantic links between aspects of different entities.

Outline. In Section 2 we give an overview over related work. In Section 3 we detail how the provided test collection is derived from a recent Wikipedia dump. Statistics about the obtained datasets are detailed in Section 3.3. In Section 4 we give details on a strong baseline. Section 5 provides evaluation results on train/test combination of our test collections to set a reference point for future work.

2 RELATED WORK AND DATASETS

The work that is most closely related is from Nanni et al. [12], whose method we adopt as a strong baseline. Their work is accompanied by a manually verified aspect linking dataset of only 201 instances semi-automatically derived from Wikipedia sections. Follow-up work [13] includes an online demo. We complement their work with a large dataset and reproducible evaluation protocol to support more research on entity aspect linking.

Several other work uses sections on Wikipedia pages to derive meaningful information. Banerjee et al. [4] and Sauper and Barzilay [16] focus on predicting content suitable to populate sections of a new Wikipedia page using a combined method of retrieval and abstractive summarization. Similarly, the goal of the TREC Complex Answer Retrieval track [7] is to retrieve content for comprehensive summaries for open-domain queries. Fetahu et al. [9] extend Wikipedia articles on news events with up-to-date information from the web. Arnold et al. [1, 2] use sections on Wikipedia articles to (1) learn how to segment articles into different topics and (2) identify answer passages to biomedical questions.

Reinanda et al. [14] accumulate content of sections that share the same heading, to compile a catalog of aspects pertaining to entity types Person and Location. These aspects of entity types are

used to classify the context of entity mentions. Our work differs in that we use an aspect catalog for each entity instance (e.g., Oyster) where Reinanda uses an aspect catalog of entity types (e.g., Person).

Balasubramanian and Cucerzan [3] build topic pages for popular person entities. They use query logs to derive aspects for entities. Three kinds of aspects are differentiated: referring only to the target entity (self), to related entities, or general. Our work is different in that we use Wikipedia sections instead of query logs as a source of aspects.

3 ENTITY-ASPECT LINKING DATASET

3.1 Construction Approach

We automatically create a large-scale test collection for entity aspect linking, using the process below.

Dataset Source. The EAL dataset is derived from an English Wikipedia dump from 01/01/2020 [8] from organizers of the TREC Complex Answer Retrieval track [2] (TREC CAR). We follow conventions of Wikipedia-derived knowledge graphs, and take each page as a representation for one entity, i.e., one node in the knowledge graph.

We use the TREC CAR schema for entity ids, paragraph ids, and section ids, e.g. "enwiki:Page%20title/Heading". Location information in the Wikipedia dump is preserved so that additional features can be derived, for example, from metadata of Wikipedia articles and long-range contexts of the page.

Aspect Catalog. The Wikipedia dump exposes information about the page content, such as the hierarchy of sections. The content is further divided into paragraphs, images, lists, and infoboxes. We use top-level sections to define a catalog of aspects, where aspect names are derived from headings, while paragraphs, lists, and subsections are preserved as content. Wikipedia editors include hyperlinks to other Wikipedia pages. We interpret such hyperlinks as entity links, where the target page represents the entity that is being mentioned on the source page.

Aspect Link Ground Truth. In some cases these hyperlinks point to a particular section of a page. We use this section information to derive the ground truth for our EAL dataset as depicted in Figure 2. We convert the following structural elements into our entity aspect link definition in Section 1.

Source of the hyperlink: context t for the aspect link.
Anchor text of the hyperlink: mention of the entity e_i to be linked.
Target page of the hyperlink: target entity e_i.
Target section of the hyperlink: true aspect a_{ij} (ground truth).
Other entity links in text: entities $e_1, e_2, \ldots e_n$ (with e_i being the target entity).
Top-level sections on target page: aspect candidates a_{i1}, a_{i2}, $\ldots$, a_{im}. Only the ground truth a_{ij} is the correct aspect.

We do not derive aspects from subsections, since these are often very specialized and closely related to other sections on a page. Furthermore, the depth and detail of the hierarchy varies greatly across pages.

In some cases the hyperlink refers to a section that is not a top-level section (e.g., a subsection within another section). Since we only use top-levels to define the aspect catalog, we use the top-level parent section as the true aspect a_{ij}.

Because of their rarity, paragraphs typically do not contain more than one hyperlink to a section. As a result, we usually have access to the ground truth aspect for one entity per paragraph (the target entity). Ultimately the task is to provide aspect links for all entities that are mentioned in text t. However this derived dataset only provides training and test data for one entity per context. In the rare cases where multiple entity mentions have a hyperlink to a section, these are broken into multiple training examples, each having one target entity with true aspect.

3.2 Resource

We apply our construction approach to offer a resource with the following filter criteria and dataset splits.

Catalog quality criteria. For each target entity we provide a catalog of candidate aspects derived from top level sections. We omit sections whenever one of the following criteria applies:

- the heading refers to a non-topical section such as "References", "See Also", "Gallery", cf. Table 1.
- the content has fewer than 50 characters.
- the name has no visible characters

In a few cases mistakes and parsing issues of the Wikipedia page led to duplicate section headings, in rare cases empty headings. Those pages were removed during processing.

EAL instance quality criteria. To obtain a high quality dataset we omit EAL examples, whenever any of the following criteria applies:

- the context sentence has fewer than 50 characters.
- the content of the true aspect has fewer than 50 characters (since such aspects are filtered from the catalog).
- the hyperlink refers to a non-topical section such as "References", "See Also", "Gallery", etc as in Table 1.
- target entities that are not regular pages, such as those containing the case-insensitive substring patterns in Table 1, e.g., "List of *".
- target entities, where the aspect catalog has either empty names or multiple aspects with the same name.[3]
- the aspect catalog as fewer than three aspects.

Training and Test splits. The goal of this resource is to offer a large test collection to support reproducible research on entity aspect linking. The choice of training data will likely affect the prediction quality achievable. Hence we provide splits of our dataset that are dedicated for training, validation, and testing.

Neural methods, such as BERT or bi-LSTMs, require vast amounts of training data, which are unwieldy for methods that use elaborate hand-crafted features. Hence we offer both a small and large subset of training data, and ask users of this resource to indicate which training set was used.

Since the task is to train the prediction of aspect links for unseen target entities, we split the dataset so that all aspect links of one target entity are either in train or validation or test.

[2] http://trec-car.cs.unh.edu/datareleases/v2.4-release.html

[3] This happens in rare cases when headings only differ by HTML symbol tags.

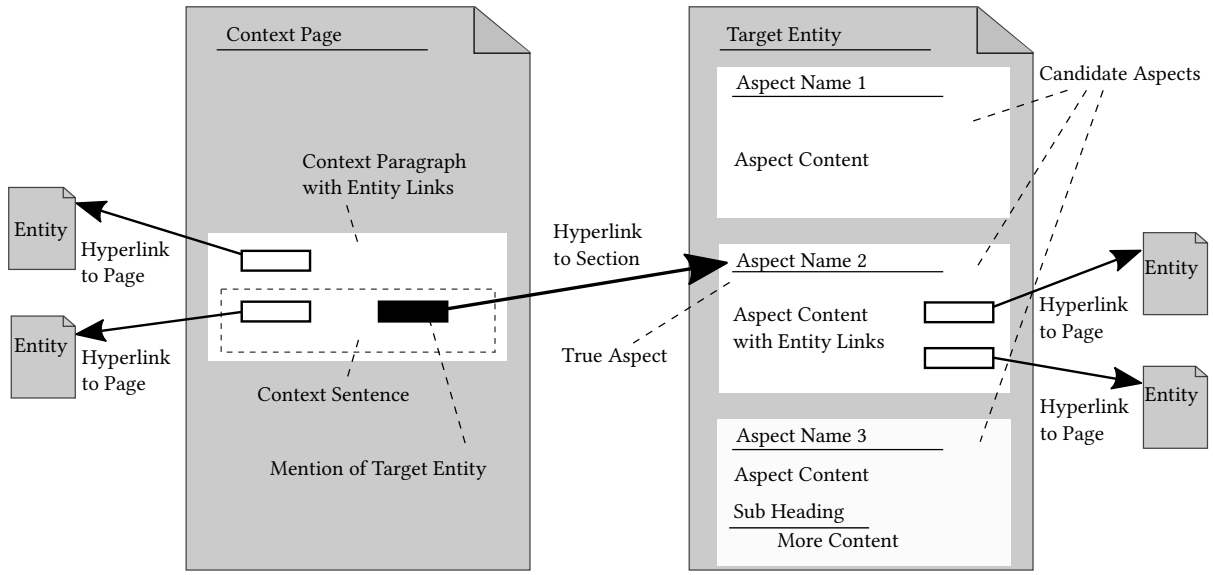

Figure 2: The entity aspect linking dataset is derived from Wikipedia hyperlinks (denoted in solid black) that refer to a section on a target page. Left: the text surrounding the link is used to derive the context and the target entity. Right: Sections on the page of the target entity are used to derive the aspect catalog, using the heading as an aspect name. The content often includes entity links, which are derived from hyperlinks to Wikipedia pages. Nested sections are in-lined into the content field.

Table 1: Patterns to filter Wikipedia pages for target entities and sections for aspects. Patterns are based on filters of TREC CAR: trec-car.cs.unh.edu/process/dataselection.html.

Page Filters		Section Filter	
lists	years of the	see also	awards
years in	lists of	reference	track listing
reference	list of	further reading	tours
glossary of	further reading	external links	sources
discographies	external links	notes	cast
by country	awards	bibliography	discography
by year	discography	gallery	filmography
bibliography	filmography	publications	other

Table 2: Top ten most frequent target entities and the number of EAL instances they participate in. 80 entities have more than 1000 EALs each. The 1000 most frequent target entities are offered separately from the train/val/test sets.

Album	24074
Village	17703
Midfielder	13212
Forward (association football)	12103
Defender (association football)	11944
Professional wrestling match types	5925
Subdivisions of Russia	4494
Administrative divisions of New York (state)	4373
Listed building	4229
Watt	3733

The distribution of target entities in the dataset follows a power-law distribution, where a small number of target entities give rise to a hundred thousand entity aspect links. Table 2 lists the ten most frequent target entities for which we derived EAL instances. To avoid that these overly-frequent entities dominate training and testing, we offer them, and other frequent entities, as a separate dataset partition.

Our dataset is partitioned by successively identifying a set of target entities and splitting off corresponding instances for entity aspect links. All instances associated with a target entity will only be contained in one of the following partitions:

Nanni-Test: Entity aspects associated with target entities used in Nanni's dataset of 201 examples. In contrast, we provide all 18,289 EAL instances associated with these 162 target entities. See Section 3.4 for a detailed discussion.

Overly-Frequent: Entity aspect links (EALs) associated with the 1000 most frequent entities, where frequency is measured as the number of EAL instances for this target entity.

Test: EALs for 1000 random target entities. (Not including target entities in Nanni-Test or Overly-frequent.)

Validation: EALs for additional 1000 target entities.

Train-Small: EALs for additional 1000 random target entities

Train-Remaining: All remaining entity aspect links.

This partitioning allows to train an entity aspect linker on Train-Remaining and/or Train-Small without accidentally leaking test data from Test, Nanni-Test nor Nanni's 201. We offer the Validation set for hyperparameter tuning.

We believe that Overly-Frequent should not be used for training, since many machine learning methods are not resilient towards unbalanced datasets.

Table 3: Number of entities with aspect catalogs, by number of aspects.

Aspects	Number of Entities
none	4,554,227
1	1,131,855
2-3	1,489,612
4-10	699,276
more than 10	18,305

Table 4: Dataset statistics of provided test collection partitions.

Partition	EAL instances	Target entities
Nanni-Test	18289	162
Overly-Frequent	429160	1000
Test	4967	1000
Validation	4313	1000
Train-Small	5498	1000
Train-Remaining	544892	106392

3.3 Dataset Statistics

Different subsets contain a differing number of EAL instances. Table 4 provides statistics on the number of EAL instances and target entities per dataset partition. Each of Train-Small, Validation and Test contain about 5000 EAL instances with on average five EAL instances per entity. For data-hungry training methods, Train-Remaining offers two orders of magnitude more EAL instances.

Aspect catalogs for all entities on Wikipedia are offered as a separate resource. This includes entities that would fail the page filter test; however the section filter applies to their catalog. This can lead to entities that don't have aspects, especially disambiguation pages and page stubs. The statistics are provided in Table 3.

3.4 Differences to Nanni's 201 Data Set

Since Nanni's 201 dataset was derived from a Wikipedia dump of 2016, the Wikipedia pages of some target entities have changed. Three pages were deleted and hence not included in Nanni-Test partition. Ten pages have changed names which results in a new entity id, which were identified through the convention of including redirects upon renaming. A few target entities were associated with more than one instance in Nanni's 201. Of the remaining 196 target entities, 30 were excluded based on our quality criteria. As a result our version of Nanni-Test includes instances associated with remaining 168 target entities of Nanni's 201.

Where Nanni et al. only include one or two EAL instances per entity, our Nanni-Test dataset includes all EAL instances associated with the target entity. This is because the paragraph ids have changed since v1.5, and it is not possible to uniquely reconstruct the context used by Nanni et al..

There are some differences between our creation process and the process used by Nanni: In some cases section hyperlinks actually refer to a subsection (as opposed to a top-level section). While Nanni used the aspect name of the ancestral top-level section, the given aspect content was that of the subsection. As this can potentially lead to ambiguous definitions of the entity aspect catalog (same name but different content) we deviated from this approach. Instead we offer an entity aspect catalog that can be computed independently of the train/test set, and hence applied to yet unseen entity references as well.

Furthermore, Nanni's 201 was created with a Wikipedia dump of an old version v1.5 of the TREC CAR Wiki parser which did not remove all HTML tags like `<ref>`, invalid Wikipedia templates, and in some cases was omitting parts of the Wikipedia page. This dataset is derived from a dump created with a much improved v2.4 parser. Also, we used a more recent Wikipedia dump that contains topics of recent interest.

Nanni et al. suggest to train/test the entity aspect linking method using 5-fold cross validation with RankLib, but did not publish the folds. We suspect that the choice of training data will affect the observed evaluation result. Hence we suggest (and document) the evaluation paradigm with our dataset and publish results obtained by the baseline using different training data sets.

3.5 Discussion on Automatic Test Collections

Fully automatically created test collections like the one described here, give rise to large publicly available datasets at a low cost. However, they also deserve to be inspected with some suspicion, since there is no guarantee that hyperlinks to sections are informative descriptions of the context. Moreover, it is not certain that derived aspect catalogs are representative of all useful aspects for one entity. For this reason we conduct a quality assessment of a sample from the automatically derived dataset and report results in Section 5. Using a similar technique, we provided a fully-automatic dataset for the TREC Complex Answer Retrieval track, for which the validity is confirmed through manual assessments produced by NIST [6]. In any case, we recommend to use this dataset in combination with the manually verified Nanni's 201 dataset. Our provided training / test splits are designed to support this usage mode, without accidentally leaking test data during training.

To develop advanced neural methods, it is mandatory to have access to very large training datasets. For example 200 instances are not sufficient to re-train a BERT model for a new task. Hence, we provide a large collection of 1 million examples free of charge under a CreativeCommons license.

When using section-hyperlinks, there are a few caveats to be aware of: The authoring tools on Wikipedia do not provide good support for the inclusion of hyperlinks to Wikipedia sections. Hence, such hyperlinks are not very common. We notice that once a section link was included, many related Wikipedia pages also include the same link—an example is the target entity Midfielder. We speculate that commonly pages are created by using a related page as a template. As a result, some target entities have more than a thousand section hyperlinks, such as Villages and Album. The frequency distribution that is very different from the usual distribution of page hyperlinks, for instance United States only has eight examples in our EAL dataset. For this reason, we exclude EAL data from overly frequent target entities in our train/val/test split, but provide them separately.

Table 5: Similarity features produced for combinations of context and aspect part.

context	aspect part	BM25	TFIDF	Overlap	W2Vec
sentence words	name words	X	X		X
paragraph words	name words	X	X		X
sentence words	content words	X	X	X	X
paragraph words	content words	X	X	X	X
sentence entities	content entities	X	X	X	
paragraph entities	content entities	X	X	X	

4 BASELINE

We complement the dataset release with a strong baseline as suggested by Nanni et al., which combines the following similarity features with a list-wise learning-to-rank approach.

4.1 Features

All features are based on similarities between context and (parts of) an aspect. Most similarities are based on words, but we also include similarities based on entity links in context and aspect content. The full list of combination is depicted in Table 5.

The following similarities are used. We exclude Nanni's RDF2Vec feature since it is difficult to produce and does not perform well.

BM25: using context as query and aspect part as document, use BM25 with default parameters as a ranking model.

TFIDF: cosine tf-idf score between context and aspect part. We use the tf-idf variant with tf log normalization and smoothed inverse document frequency.

OVERLAP: number of unique words/entities shared between context and aspect part (no normalization).

W2VEC: Word embedding similarity between context and aspect part. Word vectors are weighted by their TF-IDF weight. The pretrained word embeddings were taken from word2vec-slim, a reduced version of Google News word2vec model.[4]

Corpus statistics of word frequencies and entity link frequencies were created from 200,000 random pages of English 2020 Wikipedia.

4.2 Machine Learning

We combine the features using machine learning toolkits.

RankLib: List-wise learning-to-rank toolkit[5], using coordinate ascent to optimize for mean-average precision. Z-score normalization is enabled. We use 20 restarts per fold with 20 iterations each.

Rank-lips: List-wise learning-to-rank toolkit[6] with mini-batched training, using coordinate ascent to optimize for mean-average precision. Mini-batches of 1000 instances are iterated until the training MAP score changes by less than 1%. To avoid local optima, 20 restarts are used per fold or subset. Z-score normalization is activated.

[4] Available at https://github.com/eyaler/word2vec-slim
[5] http://www.lemurproject.org/ranklib.php
[6] http://www.cs.unh.edu/~dietz/rank-lips/

5 REFERENCE RESULTS

We provide reference evaluation results on both Nanni's 201 dataset and the test collection provided in this work.

5.1 Evaluation Paradigm

To ensure that results of new entity aspect linking methods are comparable, we introduce friendly names for different experimental setups.

Small/Test Train on Train-Small and predict on Test.

Small/Nanni-Test Train on Train-Small and predict on Nanni-Test.

Small/Nanni's 201 Train on Train-Small and predict on Nanni's 201.

Nanni's 201-CV 5-fold cross-validation on Nanni's 201. (Original evaluation protocol of Nanni et al.)

We also conduct an experiment with the larger training data set. Following Nanni et al., we provide results using features from either sentence or paragraph contexts.

Suggested evaluation metrics are precision-at-1 (P@1), mean-average Precision (MAP), and normalized discounted cumulative gain (ndcg@20), as implemented in the trec_eval package.[7]

5.2 Results

Table 6 compares the performance of between our new dataset and results of Nanni et al..

5.2.1 Findings. The results on the Small/Test experiment, are in line with Nanni's findings: Smaller contexts, such as sentences, offer better prediction quality than paragraphs. We find that the most informative feature is the similarity of sentence context with the aspect name, followed by sentence context and aspect content.

Each of the features by themselves perform rather poorly when compared to our learned models. For example, when we evaluate each sentence features with respect to MAP on Small/Test, the highest MAP obtained is 0.149, which is far lower than the MAP of 0.766 that we obtain using our rank-lips model. This suggests that the problem of predicting the relevance of entity aspects is complex and requires multiple indicators for relevance.

Table 7 displays results of additional experiments. When both sentence and paragraph features are combined, the results are not improving over sentence features alone. Similarly, for this reference model, training on the larger training set "Train-Remaining" does not significantly improve results.

A potential avenue for future work lies in developing approaches to more effectively use entity links: In some cases spurious entities are matched in the wrong aspect, in other cases a related, but slightly different entity is mentioned in the right aspect.

Regarding the Oyster example in Figure 1, our method ranks the correct aspect "Oyster/As Food" the highest, followed by "Oyster/Human history" on the second rank.

5.2.2 Reproduction of Nanni's method. We are able to reproduce Nanni's results with our re-implementation, described in Section 4. In the Nanni's 201-CV experiment, three variations of our learning-to-rank methods obtain results that are withing standard-error bars

[7] https://github.com/usnistgov/trec_eval

Table 6: Evaluation results using Train-Small and Nanni's 201. Significance is analyzed with a standard error overlap test: ▼ below standard error, ▲ above standard error.

	Paragraph Context			Sentence Context		
Small/Test	P@1	MAP	ndcg@20	P@1	MAP	ndcg@20
Rank-lips	0.581±0.007	0.744±0.004	0.808±0.003	0.624±0.007	0.772±0.004	0.829±0.003
RankLib	0.571±0.007	0.737±0.004	0.803±0.003	0.615±0.007	0.766±0.004	0.825±0.003
Small/Nanni-Test						
Rank-lips	0.575±0.004▼	0.737±0.002▼	0.802±0.002▼	0.673±0.004▲	0.807±0.002▲	0.856±0.002▲
RankLib	0.591±0.004▲	0.749±0.002▲	0.812±0.002▲	0.648±0.002▼	0.795±0.002▼	0.848±0.002▼
Small/Nanni's 201						
Rank-lips	0.622±0.034▲	0.770±0.021▲	0.827±0.016▲	0.607±0.034	0.756±0.022	0.815±0.017
RankLib	0.512±0.035▼	0.712±0.022▼	0.785±0.016▼	0.657±0.033	0.787±0.021	0.840±0.016
Nanni's 201-CV						
Rank-lips (no Z-score)	0.622±0.034	0.768±0.022	0.835±0.016	0.622±0.034	0.761±0.022	0.820±0.017
Rank-lips	0.622±0.034	0.763±0.022	0.835±0.017	0.662±0.033	0.782±0.022	0.835±0.017
RankLib	0.617±0.034	0.762±0.022	0.822±0.017	0.632±0.034	0.772±0.022	0.829±0.016
Nanni et al [12]	0.637±0.034	0.777±0.021	0.833±0.016	0.667±0.034	0.790±0.022	0.842±0.016

Table 7: Additional evaluation results. Significance test over other results on the same test set.

	P@1	MAP	ndcg@20
Train-Remaining/Test			
Sentence Context			
Rank-lips	0.628±0.007	0.774±0.004	0.830±0.003
Small/Test			
Paragraph + Sentence Context			
Rank-lips	0.626±0.007	0.772±0.004	0.830±0.004
Small/Nanni's 201			
Paragraph + Sentence Context			
Rank-lips	0.657±0.034	0.784±0.214	0.838±0.016

of results reported by Nanni et al.. The method is resilient with respect to potential differences in corpus statistics, parsing methods, feature implementation, etc. This demonstrates that the feature sets we distribute with our resource constitute an appropriate reference point for future method development.

5.2.3 Quality of Data Sets. Because of the higher number of EAL instances for training and test, results for both RankLib and rank-lips are very similar. Small error bars of ±0.004 indicate that even small performance differences can be revealed in this experimental setup. For comparison, the error bars in Nanni's original setup (Nanni's 201-CV) are an order of magnitude larger.

On the whole, all test subsets seem to be of comparable difficulty, albeit slightly more difficult than the one used by Nanni. Hence, it is important to report the used train/test set with future results.

When predicting on Nanni-Test, we found that small variations in the model's parameters can have large effects the quality, which is visible in the significance analysis. This is likely caused by the degree of imbalance in the dataset, where a few of the 162 target entities give rise to a large proportion of the 18289 EAL test instances. To avoid such issues, we separately offer Overly-Frequent separately. In contrast, Test, Validation, Train-Small, and Train-Remaining are more balanced datasets, suitable for method development and analysis.

5.2.4 Influence of Training data. The choice of training data may have a strong influence on the prediction quality. Nevertheless, we find that when—instead of cross-validating on Nanni's 201—we train on Train-Small and predict on Nanni's 201 as a held-out test set, we obtain similar performance results and similar error bars. This suggests that both training sets are equally appropriate for training competitive machine learning methods.

5.3 Manual Verification

Authors manually inspected 61 automatically derived entity aspect links, before quality filters in Table 1 were applied. The correct aspect was represented in 52 instances. Of the remaining nine, six are removed by the quality filter, one was incorrect, for two the context was not sufficient to verify the aspect. Out of the 55 EAL instances that are included in our test collection, three could not be positively confirmed, resulting in an estimated error rate of 5%.

6 CONCLUSION

In this resource we provide a large data set of 1 million entity aspect linking instances. Instances are harvested automatically from a Wikipedia dump from January 1st, 2020. The test collection is partitioned into training, validation, and test sets which are compatible with the dataset of 201 instances provided by Nanni et al. [12]. We establish an evaluation paradigm on training/testing on different partitions and offer strong reference baselines. In addition to the dataset with creation scripts, we offer downloadable feature sets, runs, and results on the resource website.[8] We will share updated results—please share your results with us.

[8] http://www.cs.unh.edu/~dietz/eal-dataset-2020/

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1846017. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

A EXAMPLES OF TARGET ENTITY "OYSTER"

Source Page: Depuration
True Aspect: Oyster/As food

> [...] One research study attempts to link the benefits of consumer awareness of shellfish depuration and found that surveyed restaurants were reluctant to sell depurated seafood. *Whereas in the same study, consumers surveyed indicated they were prepared to pay a premium for depurated* ***oysters****.* However, the willingness to pay a premium was expressed after the consumer was informed about depuration and depurated seafood indicating the average consumer was unaware about the depuration process.

Source Page: Caprella mutica
True Aspect: Oyster/Habitat and behaviour

> [...] *Along with additional specimens discovered in 1983 in* ***Coos Bay****,* ***Oregon****, these populations are believed to have been introduced to the area as a result of the importation of* ***oyster spat*** *of the Pacific oyster (Crassostrea gigas) from Japan for* ***oyster farming****.* Oysters are usually transported with algae as a packing material, particularly **Sargassum muticum** in which C. mutica are associated with.

Source Page: Harris Creek (Maryland)
True Aspect: Oyster/Habitat and behaviour

> *The* ***Nature Conservancy****, and the Oyster Recovery Partnership,* ***Maryland Department of Natural Resources****, the* ***National Oceanographic and Atmospheric Administration****, and the* ***U.S. Army Corps of Engineers*** *planted* ***oyster spat*** *on 350 underwater acres.* Planting began in 2012. Water quality is measured with a vertical profiler and water quality sondes moored at the bottom. [..]

Source Page: Eurypanopeus depressus
True Aspect: Oyster/Habitat and behaviour

> *This crab has an* ***omnivorous*** *diet which includes* ***algae****,* ***detritus****,* ***oyster spat****s,* ***polychaete*** *worms,* ***sponge****s,* ***amphipod****s and other small* ***crustacean****s.* When fully submerged it moves about on the substrate but when exposed by the retreating tide it conceals itself, being particularly associated with beds of the **eastern oyster** (Crassostrea virginica). [...]

Source Page: Starvegoat Island
True Aspect: Oyster/Human history

> Starvegoat Island (or Starve Goat Island) was a small island in the **Providence River**, **Providence, Rhode Island**. The island also appears as "Sunshine Island" on the 1927 North American datum map produced by the **US Army Corps of Engineers** 30th Battalion. The island was the southeastern most point in the city of Providence. *During the 19th and early 20th centuries, it was known for its* ***oystering****.* [...]

Source Page: Solomons, Maryland
True Aspect: Oyster/Human history

> [...] *In a* ***traffic circle*** *outside the Arts Building stands a landmark* ***bronze*** *fountain-sculpture made for Annmarie Garden which depicts a* ***Chesapeake Bay*** *waterman standing in a boat while holding* ***oyster-harvesting tongs****.* [...]

REFERENCES

[1] Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A Gers, and Alexander Löser. 2019. SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. *Transactions of the Association for Computational Linguistics* 7 (2019), 169–184.

[2] Sebastian Arnold, Betty van Aken, Paul Grundmann, Felix A Gers, and Alexander Löser. 2020. Learning Contextualized Document Representations for Healthcare Answer Retrieval. In *Proceedings of The Web Conference 2020.* 1332–1343.

[3] Niranjan Balasubramanian and Silviu Cucerzan. 2009. Automatic generation of topic pages using query-based aspect models. In *Proceedings of the 18th ACM conference on Information and knowledge management.* 2049–2052.

[4] Siddhartha Banerjee and Prasenjit Mitra. 2015. Wikikreator: Improving Wikipedia Stubs Automatically. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* 867–877.

[5] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity Query Feature Expansion Using Knowledge Base Links. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (Gold Coast, Queensland, Australia) *(SIGIR '14).* Association for Computing Machinery, New York, NY, USA, 365–374. https://doi.org/10.1145/2600428.2609628

[6] Laura Dietz and Jeff Dalton. 2020. Humans Optional? Automatic Large-Scale Test Collections for Entity, Passage, and Entity-Passage Retrieval. *Datenbank-Spektrum* (2020), 1–12.

[7] Laura Dietz and John Foley. 2019. TREC CAR Y3: Complex Answer Retrieval Overview.. In *Proceedings of Text REtrieval Conference (TREC).*

[8] Laura Dietz and Ben Gamari. 2020. TREC CAR 2.4: A Machine-Readable Wikipedia Dump.

[9] Besnik Fetahu, Katja Markert, and Avishek Anand. 2015. Automated News Suggestions for Populating Wikipedia Entity Pages. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15).* Association for Computing Machinery, New York, NY, USA, 323–332. https://doi.org/10.1145/2806416.2806531

[10] Xitong Liu and Hui Fang. 2015. Latent entity space: a novel retrieval approach for entity-bearing queries. *Information Retrieval Journal* 18, 6 (2015), 473–503.

[11] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.* 1003–1011.

[12] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. 2018. Entity-aspect linking: providing fine-grained semantics of entities in context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries.* 49–58.

[13] Federico Nanni, Jingyi Zhang, Ferdinand Betz, and Kiril Gashteovski. 2019. EAL: A Toolkit and Dataset for Entity-Aspect Linking. (2019).

[14] Ridho Reinanda, Edgar Meij, and Maarten de Rijke. 2016. Document Filtering for Long-Tail Entities. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16).* Association for Computing Machinery, New York, NY, USA, 771–780. https://doi.org/10.1145/2983323.2983728

[15] Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 74–84.

[16] Christina Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1.* Association for Computational Linguistics, 208–216.

[17] Chenyan Xiong and Jamie Callan. 2015. Query expansion with freebase. In *Proceedings of the 2015 international conference on the theory of information retrieval.* 111–120.