Abstract

Topic-specific knowledge graphs can help to support the work of subject matter experts. Through the topical context, many ambiguities can be easily resolved and spurious relationships can be removed. A crucial step in the construction of topic-specific knowledge graphs is to identify topically relevant entities from a larger background knowledge graph and a text corpus. To achieve high-recall coverage, it is not sufficient to only select entities whose name or abstract matches the topic. Our approach demonstrates the efficacy of using relevant passages to identify relevant entities and relations, outperforming relational information from knowledge graph links and bibliographic coupling. Evaluating different approaches on a benchmark of open-domain topics, we find that the co-occurrence of entities in relevant text is one of the strongest indicators (0.2534 MAP), and hence should be included when building topic-specific knowledge bases.

1. Introduction

When constructing general-purpose knowledge bases like DBpedia, WikiData, Freebase, it is important to cover as many entities and relations as possible. However, when using such knowledge bases one often faces issues of disambiguation—for example “AOC” can refer to Aeronautical operational control or Australian Olympic Committee or the human gene AOC3. Such ambiguities need to be resolved for every text query to the knowledge base. Unfortunately, popularity indicators often lead to wrong decisions. However, if we would know the general topic of interest of the querying user, such ambiguities are often rather straight-forward to resolve. For example, if the user is researching information about aeronautics when issuing this query, it is certain that the Olympic committee is not relevant.

Topic-specific knowledge bases are knowledge bases centered on a particular topic, such as aeronautics. The topical context helps with the disambiguation of text queries—or avoids disambiguation by excluding unrelated entities from the knowledge base. Topic-specific knowledge bases are a natural choice for many subject matter experts whose daily work centers around one topic, such as aeronautics, law, or bio-medicine. The advantages go beyond the issue of disambiguation: More detailed relation types can be maintained, topically
spurious connections are excluded, better coverage of specific entities can be offered—which would be regarded as idiosyncratic in general-purpose knowledge bases. Overall a cleaner and more focused knowledge base is more helpful in downstream tasks [Pujara et al., 2017].

In this work, we focus on a crucial step in the construction of topic-specific knowledge bases: The identification of a set of topically relevant entities from a general-purpose knowledge base, with the goal of constructing a topic-specific knowledge base from this set of relevant entities. As a byproduct of our approach, we also provide sets of relevant text passages from which relations and additional entities can be extracted. Another byproduct is an assessment which pairs of entities have a connection that is relevant to the topic. Finally, all selected entities are endowed with a relevance score, which should be used in place of popularity indicators. While the exploitation of these byproducts is worth a detailed study, we leave it for future work and focus on the following task.

We envision that the topic of the (to-be-constructed) knowledge base is specified through a list of short texts such as ‘aeronautics early ideas’, ‘balloon flight’, ‘branches of aviation’, ‘aeronautical engineering’, ‘aerodynamics’, ‘rockets and spacecrafts’. Such texts can be gleaned from the search history of an expert and hence do not provide an additional burden on the user or created interactively [Wolfe et al., 2015].

**Task (Topic-specific Entity Retrieval):** Assume access to a general-purpose knowledge base and large text corpus. Given a topic as a list of text queries, (1) identify which entities in the knowledge base are relevant (emphasizing recall), and (2) associate each entity with a score of relative relevance.

The formulation of our task is related to the task of entity retrieval in information retrieval [Zhiltsov et al., 2015], semantic search [Boston et al., 2014], and question answering [Berant et al., 2013]. However, to construct topic-specific knowledge bases, it is not sufficient to retrieve a small high-precision set of entities. Instead, the topic specification is to be interpreted as a set of prototypes for an overarching topic and we must emphasize to find a high-recall set of relevant entities. The main challenge in high-recall retrieval that it is not sufficient to rely on exact keyword matches as in TF-IDF. Instead, we need to combine several weak indicators of relevance to construct a useful topical knowledge base.

**Contributions.** To find a high-recall set of topically relevant entities, most related work either retrieves entities through matches in the knowledge base entry and/or by examining link patterns in the knowledge base. We demonstrate, that a much more effective approach is to retrieve relevant text passages and use the occurrence of entities as well as the co-occurrence of two entities as an indicator for topical relevance of entities and relation.

The difference in effectiveness is even more pronounced for highly specialized single-query topics. However, even when building topic-specific knowledge bases for larger and multi-faceted topics, such as Radiocarbon Dating, using contextual relevance is a simple yet effective indicator for importance. While a full exploration of all prior approaches is out of scope for this work, our study suggests that more attention to the textual context in which entities occur is a worthwhile endeavor.

**Outline.** We discuss related work in Section 2, introduce our approach in Section 3, and study its effectiveness in Section 4.
2. Related Work

Related work has been studied in question answering, information retrieval, semantic search, and knowledge base construction, which we review below.

2.1 Question Answering and Semantic Search

Many question answering tasks ask to retrieve one (of few) entities in response to a concise description. The Dr. QA system [Chen et al., 2017] demonstrates that using a combination of information retrieval and a neural network, many open-domain entity-centric questions can be answered from free text. An alternative is semantic parsing where the question is translated into a SPARQL query, which is used to select the relevant entities [Berant et al., 2013]. Many issues evolve around matching terms in the question to properties and types, and how to use the link structure most effectively [Zou et al., 2014]. Combined knowledge graph and corpus approaches have demonstrated potential Savenkov and Agichtein [2016].

2.2 Entity Retrieval

Several information retrieval models are modified to search in structured entries of knowledge graphs, using the information of entity such as name, type, category, outlinks/inlinks [Pound et al., 2010, Raviv et al., 2012]. For example, in the Fielded Sequential Dependence Model model [Zhiltsov et al., 2015] uses unigram and sparse bigram matches of the query in the knowledge graph fields name, category and neighbors. Garigliotti and Balog [Garigliotti and Balog, 2017] explore the use of entity type (i.e. taxonomic) information to exploit type-based similarity between queries and entities. While only rarely used, text-based indicators from the context of an entity was found to be a strong feature for entity ranking tasks [Schuhmacher et al., 2015, Dietz, 2019].

2.3 Entity Set Expansion and Graph Walks

On topically focused knowledge graphs, such as ConceptNet, random walks and other graph-based methods are effective in finding relevant entities [Kotov and Zhai, 2012]. However, in application to large general-purpose knowledge graphs derived from Wikipedia, indicators from the link structure offer only disappointing performance [Boston et al., 2014].

A common approach is to start with a set of seed entities, then try to expand the set to additional entities that are similar to the seed set. For example, Pantel et al. [2009] represent entities through a vector representation, then include the most similar entities. This method can be refined with neural networks [Rastogi et al., 2019].

2.4 Personalized Knowledge Bases

Several works use queries to inform knowledge base construction. For example, different topics can be extracted from query logs and web pages to extract topical sets of entities [Pasca and Van Durme, 2008]. However, many cold start knowledge bases are constructed from a corpus, without a given topic, such as the Tinkerbell or KELVIN systems [Al-Badrashiny et al., 2017, Mayfield et al., 2014]. Wolfe et al. [2017] suggest to build personalized “pocket” knowledge bases through three steps: discovering related entities, identifying mentions, and
extracting relevant relations. In contrast, our work focuses on the first stage, to discover related entities using both a general-purpose knowledge base and a large text corpus.

Most of the related work focuses on identifying a small set of correct entities, which explains the effectiveness of exact keyword matching and semantic parsing. However, we focus on obtaining a high-recall set of entities that are relevant to different degrees, for which we show that a keyword matching approach is not sufficient. In contrast, we find that one of the most useful indicator is to use the context of entity mentions in relevant text passages.

3. Approach: Identifying Topically Relevant Entities

In contrast to previous work which relies on keyword matches in entity names and categories, or graph walks on related entities, we hypothesize that one of the most effective indicators for the relevance of entities lies in the text surrounding the entity mentions. We argue that entities which are co-mentioned in topically relevant text passages, are likely to be in a relation that is topically relevant. For this reason, we demonstrate that co-coupling patterns defined on the co-occurrence graph are more effective than co-coupling patterns on the general-purpose knowledge base. We also explore a hybrid approach, where relations for co-coupling and bibliographic coupling are weighted by topical relevance using text retrieval indicators.

We are making the following claims:

• Claim 1: Finding entities through relevant passages is more effective than searching in the knowledge base for entity entries (using names and short description).
• Claim 2: To predict the relevance of an entity relation, entity co-occurrences in relevant passages are more effective than link patterns on edges of a general-purpose knowledge base.
• Claim 3: Count-based bibliographic and co-coupling similarities can be improved by weighting edges in the knowledge base by the relevance of the shared entities.

We study these claims through the following approaches, both in isolation and in combination. We define three styles of entity relations, which are based on knowledge base links, co-occurrences, and relevance-based coupling as depicted in Figure 1 and defined in Sections 3.1–3.3.

3.1 Relevant Co-occurrence Graph

We define the graph of relevant co-occurrences as follows: Using a large corpus of text passages, we retrieve a ranking of passages for the topic (for example, top 100 passages using Lucene’s BM25 model). After entity linking the passages, we take note of pairs of co-occurring entities \((e_i, e_j)\).

Co-occ Relevance (Relevance-based co-occurrences): From the passage ranking we derive a topical relevance weight of the edge by accumulating reciprocal ranks of passages in which the entities co-occur as in Equation 1.

\[
f_{ecr}(e_i, e_j) = \sum_{\forall P: e_i, e_j \text{ mentioned in } P} \frac{1}{\text{rank}(P)}, i \neq j
\]
Co-occ Count (Co-occurrence count): We also explore a simpler alternative by simply counting the number of occurrences as in Equation 2. We only count occurrences in passages that were retrieved.

\[ f_{ecc}(e_i, e_j) = \sum_{\forall P: e_i, e_j \text{ mentioned in } P} 1, i \neq j \quad (2) \]

Mention Freq: We count the number of mentions of each entity \( e_i \) in each passage.

The graph of relevant co-occurrences only makes use of the knowledge base through the entity linker—it does not incorporate any relations provided in the given general-purpose knowledge base.

### 3.2 Unweighted Link Patterns

As baselines we compare to several widely-used link patterns which can be derived from any general-purpose knowledge base. In this line of work, we are ignoring the semantics of different relationship predicates. Concretely, we include the following approaches:

**Direct links:** For every pair of entities \((e_i, e_j)\) we determine if a direct link exists between the entities \(e_i\) and \(e_j\). We count the number of different relation predicates or occurrences, which are used as a relevance score. We consider four types of direct links as follows:

1. Outlink: Number of outlinks from the entry of entity \(e_i\) to entity \(e_j\).
2. Inlinks: Number of incoming links to the entry of \(e_i\) from \(e_j\).
3. Bidi: 1 if a bidirectional link exists and 0 otherwise. For a bidirectional link, both entities \(e_i\) and \(e_j\) link to each other.
4. Undirected: Sum of outlinks and inlinks between \(e_i\) and \(e_j\).
**Coupling - baseline:** Bibliographic coupling and co-coupling are similarity measures that score a semantic relationship between two entities based on the number of shared in or outlinks.

1. Biblio Count (Count-based Bibliographic Coupling): Number of outlinks (i.e., citations) shared by entities $e_i$ and $e_j$.
2. Co Count (Count-based Co-coupling): Number of shared inlinks between the entities $e_i$ and $e_j$.

### 3.3 Relevance-weighted Bibliographic Coupling

We further study a hybrid between link-based and content-based approaches. Traditionally, coupling measures focus on counting the number of shared pages that link to or from both entity $e_i$ and $e_j$. We suggest relevance-weighted coupling measures by first retrieving a ranking of knowledge base entries $E$, then accumulating the relevance scores of shared pages.

**Biblio Relevance (Relevance-weighted Bibliographic Coupling):** Accumulated rank scores for entries $E$ that link to both entities $e_i$ and $e_j$, as in Equation 3.

$$ f_{eb}r(e_i, e_j) = \sum_{\forall E: e_i, e_j \text{ linked from } E} \text{score}(E), i \neq j $$ (3)

**Co Relevance (Relevance-weighted Co-Coupling):** Accumulating rank scores for entries $E$ that are links to from knowledge base entries of both $e_i$ and $e_j$ (analogously to Equation 3).

### 3.4 Rankings of Passages and Knowledge Base Entries

To define the relevance of co-occurrences our approach relies on a ranking of passages $P$. For relevance weighted coupling measures we need a ranking of knowledge base entries $E$.

We create two search indexes, one from passages in a background corpus and one for knowledge base entries. In this work, we use the BM25 retrieval model to produce rankings of the top 100 passages or entries.

### 3.5 Ranking Entities by Relations and Learning to Rank

Our approaches determine the relevance of a link between two entities. We score each entity by accumulating the relevance of its incident link relevance, akin to degree centrality. The highest ranking entities are selected for inclusion in the topic-specific knowledge base.

As some approaches might demonstrate their strengths only when combined, we study combinations with a learning-to-rank approach (L2R). Each of our approaches produces a separate feature for each entity. While many different methods are available, we opt for a list-wise learning-to-rank approach implemented in RankLib [Dang et al., 2013], optimizing with coordinate ascent to optimize mean-average precision (MAP), which has demonstrated robust state-of-the-art performance on small feature spaces in the past.

---

1. Lucene’s BM25 implementation with default parameters, which demonstrated good performance on the benchmark we use in the evaluation.
3.6 Constructing Topics from Multiple Queries

We construct topics from multiple queries. We build the result of topics by combining the top 100 ranking entities for each query and accumulating the relevance scores of the entities. We keep top 100 entities from the merged entities. We experimented with top 20 ranking entities instead of 100 and reached at the same conclusion as shown in Table 4 of Appendix B.

4. Evaluation

We are evaluating the three claims in Section 3 on a large-scale benchmark from TREC Complex Answer Retrieval (CAR) [Dietz and Gamari, 2018] which is designed for high-recall entity retrieval.

4.1 Benchmark and Evaluation Paradigm

The evaluation is based on the TREC Complex Answer Retrieval “Y1 test” benchmark for evaluation and “Y1 train” for training the combination with learning-to-rank. The benchmark includes a knowledge base that is derived from a Wikipedia dump that preserves meta information about names (title, disambiguation, and anchor text) and the first paragraph as well as the entity ids of inlinks and outlinks. The benchmark also includes a background corpus, which is derived from passages on Wikipedia. To ensure reproducibility, we use hyperlinks provided with the text instead of an entity linker.

The TREC CAR benchmark includes queries that are derived from Wikipedia article outlines. Articles for queries are removed from the knowledge base by the organizers. The benchmark includes section-level queries formed by concatenating the article title and heading text. In line with the task statement in Section 1, we define topics through the list of section-level queries associated with the same article. For example, the topic “Radiocarbon dating” includes among others, the query “Radiocarbon dating / Use in archaeology / Notable applications / Pleistocene / Holocene boundary in Two Creeks Fossil Forest”.

The benchmark includes automatic assessments for 1952 queries spanning 132 topics. The validity of the automatic assessments were confirmed by manual assessors [Dietz and Dalton, 2020].

We evaluate the quality of the resulting entity set using the F1 measure. The quality of the relative relevance scores are evaluates as a ranking via the measures mean-average Precision (MAP) and Rprecision. With \( R \) being the number of true entities, Rprecision measures the set precision among the first \( R \) entities of a predicted ranking. In contrast, in MAP averages precision of ranks at which relevant entities are found, awarding partial credit for low ranking entities.

4.2 Per-query Results

We first evaluate the quality of the knowledge base when only a single query is given. The results are presented in Table 1. We find that relevance-based co-occurrence approach is the most effective approach for entity ranking. In fact, the best three approaches are all based on a passage ranking (supporting Claim 2). This improvement can be noticed in the quality of the selected set, where relevance-weighted co-occurrence obtains a significantly
higher F1. Overall the F1 scores are low, because we choose to select the top 100 entities, but for most queries, fewer than 100 entities are relevant. (Rprecision reflects the quality when selecting the right number of entities.)

In Section 2.4 we claimed that keyword matching is not sufficient for identifying a set of relevant entities. To support our claim, we evaluate the ranking of entities via their knowledge base entries as described in Section 3.4 (KG-Entity). However, this is one of the worst performing approaches. This also proves our Claim 1 that relevance-based entity mentions count approach is more effectual in giving relevant entities than searching for entity entries in the knowledge base.

For Claim 3, we study whether using a retrieval over knowledge base entries as a relevance-weight in bibliographic and co-coupling patterns, but obtain mixed results: While weighting shared inlinks by relevance (Co-coupling Relevance), it decreases the performance when weighting shared outlinks (Biblio Relevance).

To analyze how multiple approaches work in combination, we use learning-to-rank using the four best approaches: three context-based approaches and undirected links in the knowledge base. In this combination we obtain a small, but not significant improvement. The combination of all features performs worse, as some features work extremely well for some cases, but not consistently.

### 4.3 Topic-specific KB from Multiple Queries

A topic is specified through a list of search queries. In the context of the TREC CAR benchmark, all queries that are derived from the same article form one topic. We evaluate the performance obtained by each approach when using multiple queries, as described in Section 3.6. The results are given in Table 2.

To demonstrate our Claim 1, we include two baselines: Using each query to retrieve from knowledge base entries, then combine the results as in Section 3.5 (KG-Entity) and using the name of the topic to search for knowledge base entries (KG-Entity-Topic). With

---

**Table 1: Results on selecting entities for 1952 individual queries.**

<table>
<thead>
<tr>
<th>Features</th>
<th>MAP</th>
<th>Rprecision</th>
<th>F1</th>
<th>In L2R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-occ Relevance</td>
<td>0.1485 ± 0.0052</td>
<td>0.1427 ± 0.0053</td>
<td>0.0689 ± 0.0017</td>
<td>X</td>
</tr>
<tr>
<td>Co-occ Count</td>
<td>0.0955 ± 0.0036</td>
<td>0.1033 ± 0.0039</td>
<td>0.0623 ± 0.0016</td>
<td>X</td>
</tr>
<tr>
<td>Mention-Freq</td>
<td>0.1111 ± 0.0035</td>
<td>0.1120 ± 0.0038</td>
<td>0.0723 ± 0.0016</td>
<td>X</td>
</tr>
<tr>
<td>Outlinks</td>
<td>0.0740 ± 0.0027</td>
<td>0.0731 ± 0.0029</td>
<td>0.0621 ± 0.0014</td>
<td></td>
</tr>
<tr>
<td>Inlinks</td>
<td>0.0724 ± 0.0025</td>
<td>0.0733 ± 0.0029</td>
<td>0.0652 ± 0.0014</td>
<td></td>
</tr>
<tr>
<td>Bidi</td>
<td>0.0407 ± 0.0015</td>
<td>0.0341 ± 0.0018</td>
<td>0.0608 ± 0.0014</td>
<td>X</td>
</tr>
<tr>
<td>Undirected</td>
<td>0.0788 ± 0.0027</td>
<td>0.0811 ± 0.0030</td>
<td>0.0640 ± 0.0014</td>
<td>X</td>
</tr>
<tr>
<td>Biblio Count</td>
<td>0.0712 ± 0.0023</td>
<td>0.0711 ± 0.0028</td>
<td>0.0655 ± 0.0015</td>
<td></td>
</tr>
<tr>
<td>Co-coupling Count</td>
<td>0.0389 ± 0.0015</td>
<td>0.0365 ± 0.0019</td>
<td>0.0587 ± 0.0013</td>
<td></td>
</tr>
<tr>
<td>Biblio Relevance</td>
<td>0.0501 ± 0.0020</td>
<td>0.0497 ± 0.0024</td>
<td>0.0526 ± 0.0012</td>
<td></td>
</tr>
<tr>
<td>Co-coupling Relevance</td>
<td>0.0766 ± 0.0028</td>
<td>0.0777 ± 0.0031</td>
<td>0.0624 ± 0.0014</td>
<td></td>
</tr>
<tr>
<td>KG-Entity</td>
<td>0.0338 ± 0.0028</td>
<td>0.0355 ± 0.0028</td>
<td>0.0112 ± 0.0004</td>
<td></td>
</tr>
<tr>
<td>L2R</td>
<td><strong>0.1558 ± 0.0053</strong></td>
<td><strong>0.1461 ± 0.0054</strong></td>
<td><strong>0.0707 ± 0.0015</strong></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Results on selecting entities for 132 topics.

<table>
<thead>
<tr>
<th>Features</th>
<th>MAP</th>
<th>Rprecision</th>
<th>F1</th>
<th>In L2R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-occ Relevance</td>
<td>0.2140 ± 0.0111</td>
<td>0.3092 ± 0.0106</td>
<td>0.3106 ± 0.0096</td>
<td>X</td>
</tr>
<tr>
<td>Co-occ Count</td>
<td>0.1669 ± 0.0085</td>
<td>0.2636 ± 0.0095</td>
<td>0.2714 ± 0.0087</td>
<td>X</td>
</tr>
<tr>
<td>Mention-Freq</td>
<td>0.1971 ± 0.0085</td>
<td>0.3032 ± 0.0087</td>
<td>0.3189 ± 0.0086</td>
<td>X</td>
</tr>
<tr>
<td>Outlinks</td>
<td>0.1509 ± 0.0071</td>
<td>0.2568 ± 0.0078</td>
<td>0.2722 ± 0.0078</td>
<td></td>
</tr>
<tr>
<td>Inlinks</td>
<td>0.1566 ± 0.0065</td>
<td>0.2630 ± 0.0078</td>
<td>0.2839 ± 0.0079</td>
<td></td>
</tr>
<tr>
<td>Bidi</td>
<td>0.0538 ± 0.0043</td>
<td>0.1357 ± 0.0072</td>
<td>0.1665 ± 0.0077</td>
<td></td>
</tr>
<tr>
<td>Undirected</td>
<td>0.1640 ± 0.0071</td>
<td>0.2707 ± 0.0078</td>
<td>0.2856 ± 0.0081</td>
<td>X</td>
</tr>
<tr>
<td>Biblio Count</td>
<td>0.1469 ± 0.0068</td>
<td>0.2527 ± 0.0081</td>
<td>0.2700 ± 0.0079</td>
<td></td>
</tr>
<tr>
<td>Co-coupling Count</td>
<td>0.0812 ± 0.0052</td>
<td>0.1844 ± 0.0075</td>
<td>0.2129 ± 0.0077</td>
<td></td>
</tr>
<tr>
<td>Biblio Relevance</td>
<td>0.1099 ± 0.0064</td>
<td>0.2135 ± 0.0075</td>
<td>0.2320 ± 0.0080</td>
<td></td>
</tr>
<tr>
<td>Co-coupling Relevance</td>
<td>0.1339 ± 0.0068</td>
<td>0.2320 ± 0.0083</td>
<td>0.2448 ± 0.0086</td>
<td></td>
</tr>
<tr>
<td>KG-Entity</td>
<td>0.0172 ± 0.0020</td>
<td>0.0519 ± 0.0037</td>
<td>0.0499 ± 0.0031</td>
<td></td>
</tr>
<tr>
<td>KG-Entity-Topic</td>
<td>0.0187 ± 0.0023</td>
<td>0.0544 ± 0.0038</td>
<td>0.0495 ± 0.0035</td>
<td></td>
</tr>
<tr>
<td>L2R</td>
<td>0.2534 ± 0.0116</td>
<td>0.3552 ± 0.0104</td>
<td>0.3175 ± 0.0092</td>
<td></td>
</tr>
</tbody>
</table>

a MAP of 0.0187 and F1 of 0.05, the performance of both baselines is dramatically lower than any of the alternative approaches in our study—supporting Claim 1.

We see that approaches that consider the relevance of passage contexts are still superior (supporting Claim 2). However, while in the per-query case, Co-occurrence Relevance improved over undirected links by 80% in terms of MAP, the advantage is only 30% for multi-query topics.

In general we see the same pattern for multi-query topics as in the per-query analysis: Relevance-based weighting of shared nodes improves the performance of co-coupling but not bibliographic coupling. We suspect that one the reasons why no convincing improvement is seen with relevance-weighted bibliographic coupling, is that the quality of the underlying knowledge base ranking (KG-Entity) is inferior.

4.4 Narrative Evaluation of Radiocarbon Dating

We provide detailed results for the example topic “Radiocarbon Dating”. The list of queries for this topic is given in Appendix A. We first analyze the both the topic-specific knowledge base for this topic which consist of 17 queries as well as one particular query in Table 3.

We observe the similar performance differences as before: relevance-based co-occurrence performs better than link patterns. The exception is that Co-coupling Relevance is the best performing knowledge base approaches both for the topic at large and the single query.

So far we have observed in topic-level results that relevance-weight improved performance only for shared outlinks but not for shared inlinks. In this single topic, we observe that relevance-weight can be effective in improving the performance of both shared inlinks and outlinks. We suspect that entities with lower number of edges i.e., inlinks and outlinks, can benefit from relevant context.

For the example query we analyze which approach retrieves how many of the seven relevant entities in the top 100 (cf. column “Entities” in Table 3). We see that the relevance-
The Pleistocene is a geological epoch that began about 2.6 million years ago. The Holocene, the current geological epoch, begins about 11,700 years ago, when the Pleistocene ends. Establishing the date of this boundary—which is defined by sharp climatic warming—as accurately as possible has been a goal of geologists for much of the 20th century. At Two Creeks, in Wisconsin, a fossil forest was discovered (Two Creek Buried Forest State Natural Area), and subsequent research determined that the destruction of the forest was caused by the Valders ice readvance, the last southward movement of ice before the end of the Pleistocene in that area. Before the advent of radiocarbon dating, the fossilized trees had been dated by correlating sequences of annually deposited layers of sediment at Two Creeks with sequences in Scandinavia. This led to estimates that the trees were between 24,000 and 19,000 years old, and hence this was taken to be the date of the last advance of the Wisconsin glaciation before its final retreat marked the end of the Pleistocene in North America. In 1952 Libby published radiocarbon dates for several samples from the Two Creeks site and two similar sites nearby; the dates were averaged to 11,404 BP with a standard error of 350 years. This result was uncalibrated, as the need for calibration of radiocarbon ages was not yet understood. Further results over the next decade supported an average date of 11,350 BP, with the results thought to be most accurate averaging 11,600 BP. There was initial resistance to these results on the part of Ernst Antevs, the paleobotanist who had worked on the Scandinavian varve series, but his objections were eventually discounted by other geologists. In the 1990s samples were tested with AMS, yielding (uncalibrated) dates ranging from 11,640 BP to 11,800 BP, both with a standard error of 160 years. Subsequently, a sample from the fossil forest was used in an interlaboratory test, with results provided by over 70 laboratories. These tests produced a median age of 11,788 ± 3 BP (2σ confidence) which when calibrated gives a date range of 13,730 to 13,550 cal BP. The Two Creeks radiocarbon dates are now regarded as a key result in developing the modern understanding of North American glaciation at the end of the Pleistocene.

Figure 2: First ranked relevant feedback passage for the example single query “Radiocarbon dating/Use in archaeology/Notable applications/Pleistocene/Holocene boundary in Two Creeks Fossil Forest”. The linked entities in the passage are underlined.

Table 3: For the example topic “Radiocarbon Dating” after combining all 17 queries in comparison to the result of a single query “Radiocarbon dating/Use in archaeology/Notable applications/Pleistocene/Holocene boundary in Two Creeks Fossil Forest”. For the query, seven entities are according to the ground truth. The columns are number of predicted relevant entities in the top 100, and the three evaluation scores obtained by the approaches.

<table>
<thead>
<tr>
<th>Features</th>
<th>MAP</th>
<th>Rprecision</th>
<th>F1</th>
<th>Entities</th>
<th>MAP</th>
<th>Rprecision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-occ Relevance</td>
<td>0.2619</td>
<td>0.3390</td>
<td>0.3670</td>
<td>7</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.1308</td>
</tr>
<tr>
<td>Co-occ Count</td>
<td>0.2102</td>
<td>0.2966</td>
<td>0.3211</td>
<td>3</td>
<td>0.2921</td>
<td>0.2357</td>
<td>0.0561</td>
</tr>
<tr>
<td>Mention-Freq</td>
<td>0.2105</td>
<td>0.3136</td>
<td>0.3394</td>
<td>7</td>
<td>0.3663</td>
<td>0.2857</td>
<td>0.1308</td>
</tr>
<tr>
<td>Undirected</td>
<td>0.1795</td>
<td>0.2881</td>
<td>0.3119</td>
<td>3</td>
<td>0.1760</td>
<td>0.1429</td>
<td>0.0561</td>
</tr>
<tr>
<td>Biblio Count</td>
<td>0.1373</td>
<td>0.2712</td>
<td>0.2936</td>
<td>4</td>
<td>0.0690</td>
<td>0.1429</td>
<td>0.0748</td>
</tr>
<tr>
<td>Co-coupling Count</td>
<td>0.0500</td>
<td>0.1695</td>
<td>0.1835</td>
<td>2</td>
<td>0.0105</td>
<td>0.0000</td>
<td>0.0374</td>
</tr>
<tr>
<td>Biblio Relevance</td>
<td>0.1405</td>
<td>0.2712</td>
<td>0.2936</td>
<td>3</td>
<td>0.0382</td>
<td>0.0000</td>
<td>0.0561</td>
</tr>
<tr>
<td>Co-coupling Relevance</td>
<td>0.2073</td>
<td>0.2881</td>
<td>0.3119</td>
<td>3</td>
<td>0.3125</td>
<td>0.2857</td>
<td>0.0561</td>
</tr>
</tbody>
</table>

Based co-occurrence approach is able to retrieve all relevant entities at higher or equal recall than the other approaches with an exception of Biblio Count.

The presence of all relevant entities in the top most ranking feedback passage of the single query as shown in the Figure 2 leads to such high recall in relevance based co-occurrence approach. We assume that entities mentioned in higher ranking passages are
more relevant to the query and hence are important than other entities. As seen in Figure 2 all the relevant entities are present in the first ranking passage and hence are given higher importance in Co-occurrence Relevance. The list of all the relevant entities retrieved for this single query is given in Table 5.

The relevance-based co-occurrence approach also performs better in terms of MAP and Rprecision (with and exception of Co-coupling Relevance).

5. Conclusion

We study the problem of selecting a high-recall set of topically relevant entities for topic-specific knowledge base construction. We find that links in the knowledge base are not effective in selecting relevant entities, possibly because the majority of links are not relevant for the topic at hand. In contrast, using entity co-occurrences in retrieved text passages—especially when the relative relevance of passages is incorporated as link strength—is between 80% and 30% more effective than the best link-based approach.

While incorporating the relevance of shared entities can improve coupling approaches, this requires a high-quality ranking of entities. We found that one of the most commonly used approaches, namely retrieving entities through names and abstract of their knowledge base entry, is an ineffective approach for topic-specific knowledge base construction.

References


Appendix A. Queries for Topic Radiocarbon Dating

The topic "Radiocarbon Dating" consists of several sections as follows:
1. Radiocarbon dating
2. Radiocarbon dating/Background/History
3. Radiocarbon dating/Background/Physical and chemical details
4. Radiocarbon dating/Background/Principles
5. Radiocarbon dating/Dating considerations/Atmospheric variation
6. Radiocarbon dating/Dating considerations/Isotopic fractionation
7. Radiocarbon dating/Dating considerations/Reservoir effects
8. Radiocarbon dating/Measurement and results/Accelerator mass spectrometry
9. Radiocarbon dating/Measurement and results/Beta counting
10. Radiocarbon dating/Measurement and results/Calibration
11. Radiocarbon dating/Measurement and results/Errors and reliability
12. Radiocarbon dating/Measurement and results/Reporting dates
13. Radiocarbon dating/Samples/Preparation and size
14. Radiocarbon dating/Use in archaeology/Impact
15. Radiocarbon dating/Use in archaeology/Interpretation
16. Radiocarbon dating/Use in archaeology/Notable applications/Dead Sea Scrolls
17. Radiocarbon dating/Use in archaeology/Notable applications/Pleistocene/Holocene boundary in Two Creeks Fossil Forest

Appendix B. Different Ranking Entities

Table 4 shows per-query results by selecting top 20 instead of top 100 ranking entities for evaluation. The results shows that top 20 ranking entities does not contain all the relevant entities detected by each relation style approaches.
Table 4: Results on selecting top 20 entities for 1952 individual queries.

<table>
<thead>
<tr>
<th>Features</th>
<th>MAP</th>
<th>Rprecision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-occ Relevance</td>
<td>0.1269</td>
<td>0.1395</td>
<td>0.1202</td>
</tr>
<tr>
<td>Co-occ Count</td>
<td>0.0753</td>
<td>0.1007</td>
<td>0.0877</td>
</tr>
<tr>
<td>Mention-Freq</td>
<td>0.0915</td>
<td>0.1107</td>
<td>0.1147</td>
</tr>
<tr>
<td>Undirected</td>
<td>0.0584</td>
<td>0.0785</td>
<td>0.0821</td>
</tr>
<tr>
<td>Biblio Count</td>
<td>0.0480</td>
<td>0.0679</td>
<td>0.0801</td>
</tr>
<tr>
<td>Co-coupling Count</td>
<td>0.0199</td>
<td>0.0339</td>
<td>0.0447</td>
</tr>
<tr>
<td>Biblio Relevance</td>
<td>0.0349</td>
<td>0.0475</td>
<td>0.0553</td>
</tr>
<tr>
<td>Co-coupling Relevance</td>
<td>0.0623</td>
<td>0.0795</td>
<td>0.0611</td>
</tr>
</tbody>
</table>

Table 5: Relevant Entities retrieved by each indicator in top 100 ranking entities for the single query “Radiocarbon dating/Use in archaeology/Notable applications/Pleistocene/Holocene boundary in Two Creeks Fossil Forest”. The column Correct Relevant Entities are the ground truth entities and remaining columns specify the relevant entities retrieved by all the approaches.

<table>
<thead>
<tr>
<th>Correct Relevant Entities</th>
<th>Co-occ Relevance</th>
<th>Co-occ Count</th>
<th>Mention-Freq</th>
<th>Undirected</th>
<th>Biblio Count</th>
<th>Co-coupling Count</th>
<th>Biblio Relevance</th>
<th>Co-coupling Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ernst Antevs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Holocene</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Paleobotany</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Pleistocene</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Two Creeks Buried Forest</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>State Natural Area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two Creeks, Wisconsin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Wisconsin glaciation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Appendix C. Correct Entities Detected

Table 5 shows the relevant entities that are retrieved by each relation style approach for the single query “Radiocarbon dating/Use in archaeology/Notable applications/Pleistocene/Holocene boundary in Two Creeks Fossil Forest”. There are 7 ground truth entities for this single query and each of the column specify which relevant entities were retrieved by each of the relation style approach in top 100.

Appendix D. Entity Expansion Approach Results

We explore the approach of entity expansion, Table 6, by expanding the query with top 20 ranking entities retrieved from relevant passages (Entity-Expansion-Topic). A MAP of
Table 6: Results of Entity Expansion approach for 132 topics.

<table>
<thead>
<tr>
<th>Features</th>
<th>MAP</th>
<th>Rprecision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity-Expansion-Topic</td>
<td>0.0169 ± 0.0040</td>
<td>0.0457 ± 0.0067</td>
<td>0.0496 ± 0.0069</td>
</tr>
</tbody>
</table>

0.0169 and F1 of 0.0496 shows the performance of entity expansion approach is lower than the other approaches.