

# Document knowledge base linking for information retrieval

[Blind]

## ABSTRACT

Entity linking and wikification are the tasks of matching mentions of an entity, such as a person, place, or organization, or a concept with its representation in a knowledge base such as Wikipedia. While there has been some investigation into use of entity linking in information retrieval, its usage may be hampered by the computational expense of constructing accurate entity annotations on large corpora, the frequent need for training data to construct entity links, and the ambiguity involved in real-world entity linking. We present a method by which a “bag of links” from a document to a knowledge base may be generated using standard information retrieval techniques with Wikipedia. We also suggest a document expansion approach that employs these links, which is effective in improving retrieval results.

## Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]

## 1. INTRODUCTION

Entity linking and wikification have been the subject of much research, especially from the Text Analysis Conference (TAC) Knowledge Base Population (KBP) entity linking track. Systems are designed to link mentions of entities or concepts found in the text to their representations in a knowledge base such as Wikipedia<sup>1</sup> or Freebase<sup>2</sup>.

These links enrich documents by connecting knowledge base information such as hyperlink graphs, entity membership in categories, and additional text to the annotated documents. It seems reasonable that information retrieval (IR) systems might take advantage of this added source of information to improve the quality of search results. Unfortunately, producing high quality entity annotations is often a computationally expensive process. In addition, these systems generally require training data, which increases the

barrier to use. Much entity linking and wikification research has also assumed more amenable circumstances that complicate their application to real-world text. For example, until 2014, the KBP entity linking track supplied explicit mention boundaries to participants.

In this paper, we present a method of using standard information retrieval techniques to produce knowledge base links for a corpus of documents. Our approach benefits by removing the requirement that links connect specific mentions from the text. We argue that such a “bag of links” provides many of the same benefits as more fine-grained entity linking or wikification processes for the purposes of ad hoc document retrieval. To demonstrate this benefit, we employ our knowledge base links using a simple document expansion model, which yields improvement over IR baselines.

## 2. RELATED WORK

### 2.1 Entity Linking, Wikification, and Entity Retrieval

Entity linking and wikification have been studied extensively, particularly in the context of TAC KBP [8]. Variations on these tasks exist: in some instances, entity mentions are assumed to be available, essentially reducing the tasks to disambiguation; more recently, the tasks includes identification of mentions in raw document text. If a system must identify entity mentions, it is common to use a named entity recognition tagger [7, 12]

Entity linking systems often exploit knowledge base structure to help match and disambiguate entities. For example, Cucerzan [4] employs Wikipedia disambiguation pages and redirection pages to help identify various entity surface forms. The use of anchor text and hyperlink graphs is also a frequent feature of these systems [11, 12, 13]. Most systems employ some form of context, which often refers to the co-occurrence of entities as evidence for disambiguation [13, 5]. Another use of context, relating most closely to our approach in this paper, refers to the text surrounding an entity mention, which can be used to disambiguate knowledge base representations [11, 5].

The related area of entity retrieval is also relevant to our work, e.g. [1, 3]. Entity retrieval refers to the task of retrieving entities, rather than documents, in response to a specified information need. This was studied at the TREC entity retrieval track [2]. Though entity retrieval tasks differ from ours in that entities are explicitly requested by the user, the querying of a corpus to retrieve entities is conceptually similar to much of our knowledge base linking approach.

<sup>1</sup><http://wikipedia.org>

<sup>2</sup><http://freebase.org>

## 2.2 Document Expansion in Information Retrieval

Investigation into the use of knowledge base links for document retrieval have, to our knowledge, been limited. Those applications most similar to ours use entity links in a feedback process to expand queries. Dalton and Dietz [6] use a number of sources, including entity-annotated queries and pseudo-relevance feedback (PRF) on entity-annotated document collections. Their process is underpinned by entity links produced by their KB Bridge system [5]. Similarly, Xiong and Callan expand queries using PRF with the descriptions and categories of Freebase objects linked to a retrieved document in the FACC1<sup>3</sup> annotation data. They employ a supervised machine learning model to select expansion terms, trained on the basis of the terms' influence on the ranking scores of relevant documents [15].

In contrast, our approach is concerned with *document* expansion. It is based on two-stage smoothing [16]; however, while two-stage smoothing generally uses the collection as its background language model, we introduce the knowledge base description text. Our idea is closely related to Wei and Croft's LDA-based document model, which smooths the document language model with latent Dirichlet allocation probabilities [14]. Liu and Croft's *CBDM* model performs a similar type of smoothing by interpolating the probability of a query term in a document cluster with its probability in the document [10].

## 3. CONSTRUCTING KNOWLEDGE BASE LINKS

### 3.1 Underlying Retrieval Model

Throughout this paper we rely on the language modeling retrieval framework [9], though this is not strictly necessary and imposes no particular mathematical constraints on our approach.

More specifically, our framework for all of the retrievals carried out in this work is the query likelihood (QL) ranking method. Given a query  $Q$  and a document  $D$ , we rank documents on  $P(Q|\theta_D)$ , where  $\theta_D$  is the language model (typically a multinomial over the vocabulary  $V$ ) that generated the text of document  $D$ . Assuming independence among terms and a uniform distribution over documents, each document is scored by

$$\log P(Q|D) = \prod_{w \in Q} P(w|Q) \cdot \log P(w|\theta_D). \quad (1)$$

We follow standard procedures for estimating the probabilities in Eq. 2. We simply use the maximum likelihood estimate of  $\hat{P}(w|Q) = \frac{c(w,Q)}{|Q|}$  where  $c(w,Q)$  is the frequency of word  $w$  in  $Q$ . For  $P(w|\theta_D)$  we estimate a smoothed language model by assuming that document language models in a given collection have a Dirichlet prior distribution:

$$\hat{P}(w|\theta_D) = \frac{c(w,D) + \mu \hat{P}(w|C)}{|D| + \mu} \quad (2)$$

where  $\hat{P}(w|C)$  is the maximum likelihood estimate of the probability of seeing word  $w$  in a "background" collection

<sup>3</sup><http://lemurproject.org/clueweb09/FACC1/>

$C$  (typically  $C$  is the corpus from which  $D$  is drawn), and  $\mu \geq 0$  is the smoothing hyper-parameter.

### 3.2 Linking with Document Pseudo-Queries

To find candidate entities to "link" to a given document  $D$ , we begin by treating the text of  $D$  as a pseudo-query which we pose against a collection of entities  $C_E$ . To transform a document into a pseudo-query we apply two transformations. First we remove all terms from  $D$  that appear in the standard Indri stoplist<sup>4</sup>. Next, we prune our pseudo-query by retaining only the  $0 < k \leq |D|$  most frequent words in the stemmed text of  $D$ . The integer variable  $k$  is a parameter that we choose empirically. Let  $Q_D$  be the pseudo-query for  $D$ , consisting of the text of  $D$  after our two transformations.

With our pseudo-query for document  $D$  in hand, we obtain a list of candidate entities by running  $Q_D$  over an index of our knowledge base,  $C_E$ , where each entry in this index is the text of an entity  $E$ 's knowledge base node. More formally, we rank the entities in our knowledge base against  $D$  using Eq. 2, substituting  $Q_D$  for the query and  $E_i$ —the text of the  $i^{th}$  entity—for the document. Let  $\pi_i$  be the log-probability for entity  $E_i$  with respect to  $D$  given by equation 2.

We now have a ranked list of tuples  $\{(E_1, \pi_1), (E_2, \pi_2), \dots, (E_N, \pi_N)\}$  relating knowledge base entry  $E_i$  to  $D$  with log-probability  $\pi_i$ . Our final step is to take the top  $n$  entries where  $0 \leq n \leq N$ . We call these top entries  $\mathcal{E}_D$  and designate them as our knowledge base links for  $D$ . Finally, we exponentiate each  $\pi_i$  and normalize our entity scores so they sum to 1 over the  $n$  retained entities. Assuming a uniform prior over entities, we now have a probability distribution over our  $n$  retained entities:  $P(E|D)$ .

Since this procedure does not depend on the query, we may compute  $\mathcal{E}_D$  once at indexing time and reuse our knowledge base links across queries.

## 4. KB-LINKED RETRIEVAL MODEL

We would now like to incorporate our knowledge base links into a retrieval model over documents. Though many knowledge bases provide structured information such as hyperlink graphs and entity category information, in this work we focus only on the textual content supplied for each entry.

We assume that a query is generated by a mixture of the document model  $\theta_D$  and a model  $\theta_K$  representing the concepts linked from the knowledge base. We assume that  $\theta_K$  can be estimated using the linked knowledge base concepts  $\mathcal{E}_D$ . This mixture model may be expressed as:

$$\hat{P}^\lambda(Q|D) = \prod_{i=1}^m (1 - \lambda)P(q_i|D) + \lambda P(q_i|\mathcal{E}_D) \quad (3)$$

The larger  $\lambda$  is, the more we believe that the knowledge base concepts are responsible for generating  $Q$ , and the less we believe that the document is responsible for generating  $Q$ . We estimate  $P(q_i|\mathcal{E}_D)$  in expectation:

$$P(q_i|\mathcal{E}_D) = \sum_{E \in \mathcal{E}_D} P(q_i|E)P(E|D). \quad (4)$$

Like  $P(q_i|D)$ , we estimate  $P(q_i|E)$  as a Dirichlet-smoothed query likelihood, but using the knowledge base's text for entry  $E$  to estimate of the underlying model. By virtue of

<sup>4</sup><http://www.lemurproject.org/stopwords/stoplist.dft>

our entity-document scoring and normalization, we also have  $P(E|D)$ .

## 5. EVALUATION

### 5.1 Data

To perform knowledge base linking, we make use of the September 1, 2015 dump of English Wikipedia. We build an Indri<sup>5</sup> index over the Wikipedia page text. The text of each Wikipedia page also serves as the “description text” used in Eq. 4.

We test our approach using the TREC 2004 robust topics. These 250 topics are used with data from TREC disks 4 and 5. In addition, we use the AP newswire collection from TREC disks 1 and 2 with topics 101-200.

### 5.2 Runs

We produce two runs per collection:

- *baseline-ql*, a baseline query likelihood run
- *kb-ql*, incorporating knowledge base links using Eq 3.

For the two *kb* runs, we retrieve the top 1000 documents per query based on the default Indri query likelihood implementation. We then re-rank these documents by incorporating their knowledge base links as described in Section 4.

### 5.3 Parameters

Param	Meaning	Value
$k$	The maximum number of document terms to use in constructing $D_Q$ .	20
$n$	The maximum number of knowledge base entries in $\mathcal{E}_D$ .	10
$\lambda$	Mixing parameter controlling the weights of $P(q D)$ and $P(q E)$	0.0-1.0
$\mu$	Used for Dirichlet smoothing of both $P(q D)$ and $P(q E)$ .	2500

Table 1: Parameter settings for the entity linking procedure and retrieval model

The various parameters required for our approach, along with their meanings and values are shown in Table 1.

We sweep across values of  $\lambda$  at intervals of 0.1 to investigate how much weight should be given to the entity model. The results of these sweeps are shown in Figure 1 and discussed further in Section 6.

For this work, we set  $k$  and  $n$  heuristically. In principle, both parameters need not be limited beyond the length of the document and size of the corpus respectively; however, this would increase computation time significantly, so we have opted to set them to the specified values.

## 6. RESULTS

Retrieval performance of the top-scoring knowledge base runs and baselines are shown in Table 2. Bolded mean average precision (MAP) scores are greater than their corresponding baseline (e.g. *kb-ql* compared to *baseline-ql*) with

<sup>5</sup><http://www.lemurproject.org/indri/>

	Run	$\lambda$	MAP
Robust	<i>baseline-ql</i>	–	0.2183
	<i>kb-ql</i>	0.9	<b>0.2275</b>
AP	<i>baseline-ql</i>	–	0.2346
	<i>kb-ql</i>	0.9	<b>0.2515</b>

Table 2: The top-scoring knowledge base runs and baselines by mean average precision (MAP).  $\lambda$  is specified for knowledge base runs. Bolded values indicate statistical significance compared to the baseline.

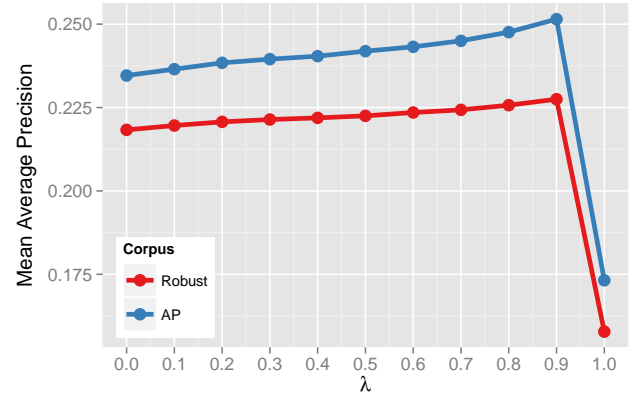


Figure 1: Sweeps over values of  $\lambda$  for robust (red) and AP (blue) runs.

statistical significance at  $p < 0.001$  using a paired t-test. Note that baselines correspond to  $\lambda = 0.0$ .

Though Table 2 shows only the top-scoring values of  $\lambda$ , performance for all knowledge base runs improved over the baseline for values of  $0.0 < \lambda < 1.0$ , as shown in Figure 1. All improvements were statistically significant with at least  $p < 0.05$ .

## 7. CONCLUSIONS

The results indicate that our approach for constructing knowledge base links between documents and Wikipedia produces useful data for document retrieval purposes. Our simple document expansion model that incorporates these links performs well compared to a query likelihood baseline. These outcomes support our argument that a “bag of links” to a knowledge base can provide helpful information for a document retrieval task. Our proposed method using standard information retrieval techniques to build these links is efficient compared to usual entity linking and wikification systems and can be accomplished without the need for complex natural language processing.

In this paper, we have limited ourselves to using only knowledge base description text. However, knowledge base links provide a great deal more information. Future work may benefit from harnessing this knowledge base data such as hyperlink graphs and entity categories. Since our retrieval model performs *document* expansion, we also plan to investigate its utility when paired with *query* expansion techniques that employ knowledge base links.

## 8. ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation under Grant No. [blind]. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## 9. REFERENCES

- [1] S. F. Adafre, M. de Rijke, and E. T. K. Sang. Entity retrieval. *Recent Advances in Natural Language Processing*, 2007.
- [2] K. Balog and P. Serdyukov. Overview of the trec 2011 entity track. Technical report, National Institute of Standards and Technology, 2011.
- [3] M. Bron, K. Balog, and M. De Rijke. Ranking related entities: Components and analyses. In *Proceedings of the 19th ACM international conference on information and knowledge management*, pages 1079–1088. ACM, 2010.
- [4] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716, 2007.
- [5] J. Dalton and L. Dietz. A neighborhood relevance model for entity linking. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 149–156. Le Centre De Hautes Etudes Internationales D’informatique Documentaire, 2013.
- [6] J. Dalton, L. Dietz, and J. Allan. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval*, pages 365–374. ACM, 2014.
- [7] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics, 2011.
- [8] H. Ji, J. Nothman, and B. Hachey. Overview of tac-kbp2014 entity discovery and linking tasks. In *Proc. Text Analysis Conference (TAC2014)*, 2014.
- [9] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*, pages 111–119. ACM, 2001.
- [10] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193. ACM, 2004.
- [11] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on information and knowledge management*, pages 233–242. ACM, 2007.
- [12] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics, 2011.
- [13] V. Stoyanov, J. Mayfield, T. Xu, D. W. Oard, D. Lawrie, T. Oates, and T. Finin. A context-aware approach to entity linking. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 62–67. Association for Computational Linguistics, 2012.
- [14] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, pages 178–185. ACM, 2006.
- [15] C. Xiong and J. Callan. Query expansion with freebase. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 111–120. ACM, 2015.
- [16] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.