Bag-of-Entities Representation for Ranking

Chenyang Xiong  
Language Technologies Institute  
Carnegie Mellon University  
Pittsburgh, PA 15213, USA  
cx@cs.cmu.edu

Jamie Callan  
Language Technologies Institute  
Carnegie Mellon University  
Pittsburgh, PA 15213, USA  
callan@cs.cmu.edu

Tie-Yan Liu  
Microsoft Research  
Beijing, 100080, P. R. China  
tie-yan.liu@microsoft.com

ABSTRACT

This paper presents a new bag-of-entities representation for document ranking, with the help of modern knowledge bases and automatic entity linking. Our system represents query and documents by bag-of-entities vectors constructed from their entity annotations, and ranks documents by their matches with the query in the entity space. Our experiments with Freebase on TREC Web Track datasets demonstrate that current entity linking systems can provide sufficient coverage of the general domain search task, and that bag-of-entities representations outperform bag-of-words by as much as 18% in standard document ranking tasks.

Keywords  
Text Representation, Document Representation, Knowledge Base, Bag-of-Entities

1. INTRODUCTION

In the earliest information retrieval systems, query and documents were represented by terms manually picked from predefined controlled vocabularies [6]. The controlled vocabulary representation conveys clean and distilled information, and can be ranked accurately by simple methods. However, it also requires manual annotations and suffers from the small size of controlled vocabularies, thus is mainly used for specific search domains. As full-text search became popular, query and documents are mainly represented by their bag-of-words vectors, and more sophisticated ranking models are used to rank documents in the word space.

Recently, knowledge bases, as the modern version of controlled vocabularies but at larger scale, have provided a new opportunity to improve ranking. The rich semantic information in knowledge bases has been successfully used by ranking systems to better understand general domain queries, for example, to generate better expansion terms [8], richer learning to rank features [1], and additional connections between query and documents [5, 7]. Also, automatic entity annotation is made possible by entity linking research, and is becoming increasingly efficient and effective for both query and documents.

This paper presents a new bag-of-entities based representation for document ranking, as a heritage of the classic controlled vocabulary based representation, but with the aid of modern large scale knowledge bases and automatic entity linking systems. We represent query and documents by their bag-of-entities constructed from the annotations provided by three entity linking systems: Google’s FACC1 [3] with high precision, CMNS [4] with high recall, and TagMe [2] with balanced precision and recall. With the deeper text understanding provided by entity linking, documents can be ranked by their overlap with the query in the entity’s explicit semantic space.

To investigate the effectiveness of bag-of-entities representations, we conducted experiments with a state-of-the-art knowledge base, Freebase, and two large scale web corpora, ClueWeb09-B and ClueWeb12-B13, together with their queries from the TREC Web Track. Our evaluation results first confirm that current entity linking systems can provide sufficient coverage over general domain queries and documents. Then we compare bag-of-entities with bag-of-words in standard document ranking tasks and demonstrate that although the accuracy of entity linking is not perfect (about 50% – 60% on TREC Web Track queries), the ranking performance can be improved by as much as 18% with bag-of-entities representations.

2. BAG-OF-ENTITIES REPRESENTATION

We construct bag-of-entities representations for queries and documents using several entity linking systems. When annotating texts, an entity linking system does not just match the n-grams with entity names, but makes the decision jointly by also considering external evidences such as the entity’s descriptions and relationships in the knowledge base, and corpus statistics like commonness, linked probability and contexts of entities. As a result, representing texts by entities naturally incorporates deeper text understanding from the linking process: Entity synonyms are aligned, polysemy in entity mentions is disambiguated, and global coherence between entities is incorporated.

The choice of knowledge base in this paper is Freebase, one of the largest public knowledge bases frequently used in recent IR research [1, 5, 7, 8]. Several entity linking systems have been developed for it. This work explore the following three popular ones to annotate queries and documents with Freebase entities.
FACC1 is the Freebase annotation of TREC queries and ClueWeb corpora provided by Google [3]. It aims to achieve high precision, which is believed to be around 80-85% based on a small-scale human evaluation1.

TagMe is an entity linking system [2] widely used in prior research [4, 7]. It balances precision and recall, both at about 60% in various evaluations.

CMNS is an entity linking system that spots texts using surface forms from FACC1 annotation, and links all of them to their most frequently linked entities [4]. It can achieve almost 100% recall on some query entity linking datasets, but the precision may be lower [4].

Table 1: Entity linking performance on ClueWeb queries. All methods are evaluated by Precision, Recall and F1.

<table>
<thead>
<tr>
<th></th>
<th>CW09 Query</th>
<th>CW12 Query</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
</tr>
<tr>
<td>FACC1</td>
<td>0.274</td>
<td>0.236</td>
</tr>
<tr>
<td>TagMe</td>
<td>0.581</td>
<td>0.597</td>
</tr>
<tr>
<td>CMNS</td>
<td>0.577</td>
<td>0.596</td>
</tr>
</tbody>
</table>

Indexing: We indexed both corpora with Indri, using their bag-of-words. Default stemming and stopword removal were used. Spam in ClueWeb09 was filtered using the default threshold (70%) of Waterloo spam scores. Spam filtering was not used for ClueWeb12 because its effectiveness is unclear. The evaluation of entity linking, and the re-ranking using bag-of-entities are performed on the top 100 documents per query retrieved by Indri’s language model with default Dirichlet smoothing (µ = 2500).

Entity Linking Systems: We used TagMe software provided by Ferragina et al. [2] to annotate queries and documents with Wikipedia entities, which are then aligned to Freebase entities using the Wikipedia ID field in Freebase.

CMNS is implemented by ourselves, following Hasibi et al. [4]. The boundary overlaps of surface forms are resolved by only linking the earliest and then the longest one [4].

FACC1 entity annotations for ClueWeb documents are provided by Google [3]. They also annotated ClueWeb09 queries’ intent descriptions, but not the queries. We used the descriptions’ annotations as approximations of queries’ annotations, and manually filtered out entities that did not appear in the original queries to reduce disturbance. ClueWeb12’s queries are not annotated by Google so we are only able to study FACC1 annotations on ClueWeb09.

Baselines: We used two standard unsupervised bag-of-words ranking models as baselines: Indri’s unstructured language model (Lm) and sequential dependency model (SDM), both with default parameters: µ = 2500 for Lm, and query weights (0.8, 0.1, 0.1) for SDM. Typically these baselines do well in competitive evaluations such as TREC. There are better rankers, for example learning to rank methods. However, methods that combine many sources of evidence usually outperform methods that use a single source of evidence, thus such comparison would not reveal much about the value of the bag-of-entities representation.

There were three representations for ClueWeb09 (FACC1, TagMe and CMNS), and two for ClueWeb12 (TagMe and CMNS). All annotations are done automatically. All annotated entities are used, since filtering using the annotation score lowers the annotation accuracy in our experiments. CMNS and EF were used to re-rank the top 100 documents per query retrieved by Lm. Ties were broken by Lm’s score.

Evaluation Metrics: The entity annotations were evaluated by the lean evaluation metric from Hasibi et al. [4]. It averages the annotation performances (precision or recall) of all the whole query level, e.g. whether all entities in a query are correctly annotated or not, and on the individual entity level. The ranking performances were evaluated by the TREC Web Track Ad-hoc Task’s official evaluation metrics: ERR@20 and NDCG@20. Statistical significance was tested by the Fisher randomization test (permutation test) with p < 0.05.

1http://lemurproject.org/clueweb09/FACC1/
Table 2: Coverage of annotations. \textbf{Freq} and \textbf{Dens} are the average number of entities linked per query/document and per word respectively. \textbf{Missed} is the percentage of queries or documents that have no annotation at all. ClueWeb12 queries do not have FACC1 annotations as they are published later than FACC1.

<table>
<thead>
<tr>
<th></th>
<th>ClueWeb09</th>
<th>ClueWeb12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Query</td>
<td>Document</td>
</tr>
<tr>
<td></td>
<td>Freq</td>
<td>Dens</td>
</tr>
<tr>
<td>FACC1</td>
<td>0.42</td>
<td>0.20</td>
</tr>
<tr>
<td>TagMe</td>
<td>1.54</td>
<td>0.70</td>
</tr>
<tr>
<td>CMNS</td>
<td>1.50</td>
<td>0.69</td>
</tr>
</tbody>
</table>

4. EVALUATION RESULTS

This section evaluates the accuracy and coverage of entity annotations, and bag-of-entities’ performance in ranking.

4.1 Annotation Accuracy and Coverage

Table 1 shows the precision, recall, and F1 of FACC1, TagMe, and CMNS on ClueWeb queries. TagMe performs the best on ClueWeb09 queries with higher precision, while CMNS performs better on ClueWeb12 queries. The ClueWeb09 queries are more ambiguous because they needed to support the TREC Web Track’s Diversity task; TagMe’s disambiguation was more useful on this set. ClueWeb12 queries needed to support risk minimization research, and have been shown to be harder; both systems perform worse on them. FACC1 query annotation does not perform as well as its goal was to annotate the query’s description, not the query itself.

There is no gold standard entity annotation for ClueWeb documents. Nevertheless, our manual examination confirms that FACC1 has high precision; TagMe performs a little better on documents with more contexts; and CMNS performs worse than TagMe on documents as it only uses the surface forms.

One concern of controlled vocabulary based search systems is the low coverage on general domain queries, restricting their usage mainly to specific domains. With the much larger scale of current knowledge bases, it is interesting to study whether they can influence the majority of general domain queries. Table 2 shows the coverage results of our entity annotations on ClueWeb queries and their top 100 retrieved documents. \textbf{Freq} and \textbf{Dens} are the average number of entities linked per query/document, and per word respectively. \textbf{Missed} is the percentage of queries/documents that have no linked entity. The results show that TagMe and CMNS have good coverage on ClueWeb queries and documents. Almost all queries and documents have at least one linked entity. The annotations are no longer sparse. There can be up to 324 entities linked per document on average. However, precision and coverage have not been achieved together yet. FACC1 has the highest precision but provides very few entities per document and misses many documents.

These results show that the entity linking is still an open research problem. Precision and coverage can not yet be achieved at the same time. Thus, the ranking method must be robust and able to accommodate a noisy representation.

4.2 Ranking Performance

The performances of bag-of-entities in ranking are shown in Table 3a. \textbf{EF} and \textbf{COOR} rerank top retrieved documents using the bag-of-entities from FACC1, TagMe and CMNS. The percentages are relative performances over SDM. W/T/L refer to the number of queries improved (Win), unchanged (Tie) and hurt (Loss) comparing with SDM.

On ClueWeb09, both TagMe and CMNS work well with \textbf{EF} and \textbf{COOR}, and outperform all baselines on all evaluation metrics. The best method, TagMe-EF, outperforms SDM as much as 18% on ERR@20. On ClueWeb12, COOR outperforms all baselines on all evaluation metrics by about 12%. These results demonstrate that even with current imperfect entity linking systems, bag-of-entities is a valuable representation on which very basic ranking models can significantly outperform standard bag-of-words based ranking.

Bag-of-entities’ representation power correlates with the entity linking system’s accuracy. ClueWeb09 queries are more ambiguous, favoring TagMe in annotation accuracy, and TagMe provides the most improvements when ranking for ClueWeb09 queries. ClueWeb12 queries have lower annotation quality, and bag-of-entities based ranking is not as powerful as on ClueWeb09 queries. FACC1’s coverage is too low and can not well represent the documents. This also explains why prior research mainly uses it as a pool to select query entities [1, 7, 8].

Entity linking is a rapidly developing area; improvement in the future is likely. To study how the bag-of-entities can benefit from improvements in annotation accuracy, we used the manual query annotations [1, 5] to divide queries into two groups: Correctly Annotated, whose ground truth entities are all correctly linked, and Mistakenly Annotated, whose entities are not all correctly linked. Table 3b shows the ranking performances of TagMe and CMNS on the two groups in ClueWeb09. We omit FACC1 as it always hurts, and ClueWeb12 queries as there are not enough queries in either group to provide reliable observations. The relative performance, W/T/L and statistical significance over SDM are calculated on the same group of queries for each method. The results are as expected: On Correctly Annotated queries, bag-of-entities provides more accurate ranking; on Mistakenly Annotated queries, the improvements are smaller, and sometimes bag-of-entities reduces accuracy.

Our experiments show that intuitions developed for bag-of-words representations do not necessarily apply directly to bag-of-entities representations. A long line of research shows that frequency based (e.g., tf.idf) ranking models are superior to Boolean ranking models. Thus, one might expect \textbf{EF} to provide consistently more accurate ranking than \textbf{COOR}, however that is not the case in our experiments. We found that the majority of annotation errors are missed annotations, which makes entity frequency counts less reliable. However, it is rare for the entity linker to miss every mention of an important entity in a document, thus the Boolean model is robust to this majority type of errors.

We also examined the effectiveness of other ranking intuitions, such as inverse document frequency (idf) and document length normalization. In our bag-of-entities represen-
tions they did not provide improvements when used individually or in language modeling and BM25 rankers. We speculate that idf had less impact because most queries contained just one or two entities, thus most of the queries were ‘short’ in the entity space; idf is known to be less important in short texts with Wikipedia pages.

Entity linking is a rapidly-developing research area; its further improvements is likely to improve ranking accuracy, for example by providing more reliable entity frequencies. The bag-of-entities representation provides new evidence from knowledge bases, but also introduces new types of errors and uncertainties. How to better utilize bag-of-entities’ strength and handle its noises is an important future research direction.

Prior research on using knowledge bases for search was mainly query-based, for example, selecting a few entities for the query and using them to enhance the bag-of-words based ranking [1, 5, 7, 8]. This work focuses more on the document representation and operates directly in the entity space. How to combine the earlier work with the work described in this paper is another open problem.

6. ACKNOWLEDGMENTS

This research was supported by National Science Foundation (NSF) grant IIS-1422676. Any opinions, findings, and conclusions expressed in this paper are the authors’ and do not necessarily reflect those of the sponsors.

7. REFERENCES