Pair-wise Web URI classification for KBD using Distant Supervision and Unlabeled Data

ABSTRACT

Entity linking establishes a mapping from entities in given text to URI end-points in a knowledge base (KB), or NIL if the entity is not in the KB. It is an useful task to not only collate information about entities but also to disambiguate them. We explore the relaxed definition of KB from [4], that includes any URI which reliably disambiguates linked mentions on the web. To exploit these resources, we must first infer their existence on the web. We refer to this task as Knowledge Base Discovery (KBD). Web endpoints also yield disambiguated entity mentions. For every entity endpoint we discover, we may recover thousands of entity mentions via inlinks. While the effectiveness of inlink-driven entity disambiguation is known for a single KB setting, this can be extended to leverage inlinks across a collection of automatically discovered web KBs [5]. This process has the potential to both improve EL accuracy for well-covered entities and extend the coverage of EL systems by uncovering endpoints for previously unseen entities. We address the problem of pair-wise URI classification based entity disambiguation with distant supervision, by bootstrapping labeled examples using web search, and unlabeled data. We propose a learning to combine approach leveraging distantly supervised examples and unsupervised pair-wise similarity models on entire data. We empirically evaluate our proposed approach using WePS and ALTA shared task 2016 datasets.

Keywords

Entity Linking; Distant Supervision; Learning to Combine; Semi-supervised Learning

1. INTRODUCTION

Entity linking (EL) resolves textual mentions to the correct node in a knowledge base (KB). Linking systems typically rely on semantic resources such as Wikipedia, DBPedia to be used as end-points. Though they provide a rich context for entities, these sources do not achieve sufficient recall for all the domains and entities. On the other hand, domain centered sources such as DBLP, IMDb or MusicBrainz would have the depth in coverage for a target domain. It is understandable that a near complete solution would depend on merging resources from multiple knowledge bases (KBs). But an explicit reconciliation of distinct entity sets and heterogeneous sources' schema is a challenging problem in itself [5]. This work is important for growing KBs to include more entities about which we know less – the long tail. Other work shows that web links can produce models nearly as accurate as those built from richly structured KBs [4], but does not include non-Wikipedia entities. We define an entity endpoint as any URI for which inlinks reliably identify and disambiguate named entity mentions. For example, we may observe that inlinks to en.wikipedia.org/wiki/Barack_Obama are typically mentions of the entity Barack Obama. Links targeting this URI in reference to some other entity are unlikely, so we should consider this an endpoint for the entity Barack Obama. This can help to improve a variety of related challenges such as Web People Search task (WePS[1]) which is defined as a problem of organizing web search results for a given person name.

In-order to leverage web KB as end-points, valid URIs must be infered on the web, where inlink based disambiguation is a key step. Inlink based disambiguation can be handled as a pair-wise document classification. Pairs are classified as positive if they refer to the same entity and negative otherwise. Providing sufficient labeled examples will be a hard task as it requires enormous resources, but fortunately we can leverage web search based distant supervision to label pairs of URLs automatically. In this work, we propose to leverage web-search based distant supervision and similarity models built on unlabeled data to jointly label examples. This solution is motivated from [7], where relation labels from KB are used as queries. Here we use named entities from URLs as context keywords to query. In [7], the reverse task was addressed, to map entities in a KB to web URI end-points. Additionally, we evaluate an extended approach where we use distantly supervised positive examples and models built on unlabeled data to learn a meta classifier model. The intuition is that, if web-search results with context keywords from URL_a and domain name from URL_b, vice versa are used as queries, then a hit in top K search results would indicate that both URIs refer to same entity. But if the top K results do not include URL_b, then it may not rule out the possibility completely for both URLs to refer to same entity. Especially when K is a less value

© 2017 ACM. ISBN 978-1-4503-2138-9.

DOI: 10.1145/1235

[1] http://nlp.uned.es/weps/weps-3

to increase precision, such as 2 or 5, then due to domain popularity (based on pagerank, HITS) there are chances to miss the expected URI.

2. RELATED WORKS

Entity linking and disambiguation have typically relied on Wikipedia [6, 11] or a subset [10], or a larger structured resource such as Freebase [16]. Entries in the KB provide a point against which mentions that refer to that entity are clustered. In addition to this, the KBs provide extra information for an entity such as facts, text and other media. Other tasks cluster mentions of the same entity, but without reference to a central KB, namely Cross Document Coreference [2, 13] and Web Person Search [1]. The task can be more challenging, as we are unable to exploit priors inferred from the KB or leverage information about an entity for clustering. While an EL KB and a set of coreference clusters are quite different, they both act as aggregation points for mentions of their respective entities. Mining the content and structure to discover new entities is another important task. There is also substantial work in trying to identify instances of entity classes from text, exploiting language [8], document structure [14, 3] and site structure [15]. Clustering NIL entities (those that cannot be linked to the KB) has been a focus of the Text Analysis Conference (TAC) Knowledge Base Population shared tasks from 2011 [9].

In [7] authors address the reverse task of identifying good links that correspond to specific KB entities by searching for the entity name in a web search engine and refining the results. This work is important for growing KBs to include more entities about which we know less – the long tail. Other work shows that web links can produce models nearly as accurate as those built from richly structured KBs [5], but does not include non-Wikipedia entities.

In this work, we treat any valid web URI as an end-point that can be used to disambiguate to an entity. We consider the sub-task of knowledge base discovery on web which requires URI disambiguation based on in-links for entity linking. Since getting access to large amount of labeled data for task is very hard and annotating them manually would incur huge amount of resources, we propose to leverage web search based distant supervision for labeling a subset of URI pairs. We leverage distant supervision labels together with unsupervised pairwise similarity models for classifying all the URI pairs in a transductive fashion. We also evaluate an extended approach that uses distant supervision to provide examples only. Though web search results have been as a feature for clustering documents in WePS [12], we are interested to use it to label a given pair of documents as shown in [7] where web search based URL classification with entity labels as keywords is shown to work remarkably well.

Distant supervision is a learning scheme in which a classifier is learned given a weakly labeled training set where the training data is labeled automatically based on heuristics / rules. It has been successfully used for many applications such as relation extraction, sentiment analysis, emotion classification and so on. In this work, we use labeling strategy similar to [7] for obtaining labels automatically. Also to the best of our knowledge this is the first attempt to leverage distant supervision for labeling positive examples, and unsupervised similarity models on unlabeled data to jointly learn to combine different models.

3. PROPOSED APPROACH

4. EXPERIMENTAL SETUP

5. CONCLUSIONS

6. REFERENCES

[1] J. Artiles, J. Gonzalo, and S. Sekine. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 64–69. Association for Computational Linguistics, 2007.

[2] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 79–85. Association for Computational Linguistics, 1998.

[3] L. Bing, M. Ling, R. C. Wang, and W. W. Cohen. Distant ie by bootstrapping using lists and document structure. *arXiv preprint arXiv:1601.00620*, 2016.

[4] A. Chisholm and B. Hachey. Entity disambiguation with web links. *Transactions of the Association for Computational Linguistics*, 3:145–156, 2015.

[5] A. Chisholm, W. Radford, and B. Hachey. Discovering entity knowledge bases on the web. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 7–11, 2016.

[6] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716, 2007.

[7] C. Hachenberg and T. Gottron. Finding good urls: Aligning entities in knowledge bases with public web document representations. pages 17–28, 2012.

[8] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

[9] J. Heng, R. Grishman, and H. Dang. Overview of the tac2011 knowledge base population track. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*, 2011.

[10] P. McNamee and H. T. Dang. Overview of the tac 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, volume 17, pages 111–113, 2009.

[11] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[12] R. Nuray-Turan, Z. Chen, D. V. Kalashnikov, and S. Mehrotra. Exploiting web querying for web people search in weps2. In *Web People Search Evaluation Workshop, WWW*, 2009.

[13] S. Singh, A. Subramanya, F. Pereira, and A. McCallum. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the*

Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 793–803. Association for Computational Linguistics, 2011.

[14] R. C. Wang and W. W. Cohen. Language-independent set expansion of named entities using the web. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 342–350. IEEE, 2007.

[15] Q. Yang, P. Jiang, C. Zhang, and Z. Niu. Reconstruct logical hierarchical sitemap for related entity finding. Technical report, DTIC Document, 2010.

[16] Z. Zheng, X. Si, F. Li, E. Y. Chang, and X. Zhu. Entity disambiguation with freebase. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 82–89. IEEE Computer Society, 2012.