Discovering Entity Knowledge Bases on the Web

Andrew Chisholm
Hugo Australia
Sydney, Australia
achisholm@hugo.ai

Will Radford
Hugo Australia
Sydney, Australia
wradford@hugo.ai

Ben Hachey
Hugo Australia
Sydney, Australia
bhachey@hugo.ai

Abstract

Recognition and disambiguation of named entities in text is a knowledge-intensive task. Systems are typically bound by the resources and coverage of a single target knowledge base (KB). In place of a fixed knowledge base, we attempt to infer a set of endpoints which reliably disambiguate entity mentions on the web. We propose a method for discovering web KBs and our preliminary results suggest that web KBs allow linking to entities that can be found on the web, but may not merit a major KB entry.

1 Introduction

Entity linking (EL) resolves textual mentions to the correct node in a knowledge base (KB). Linking systems typically rely on semantic resources like Wikipedia as endpoints for disambiguation. These sources provide context for entity modelling, but impose an upper bound on recall based on their domain of coverage. Wide domain KBs like Wikipedia constrain coverage based on notability, while narrow domain sources like IMDb[1] or MusicBrainz[2] give depth at the expense of breadth. While it is possible to merge resources from multiple KBs in some applications, an explicit reconciliation of distinct entity sets and KB schemata is often problematic.

We explore a relaxed definition of a KB – any URI which reliably disambiguates linked mentions on the web. This covers resources which both work as a KB by design (e.g. a Wikipedia article) and those which do so implicitly by disambiguating mentions. We focus on the latter case, by trying to identify and exploit *class-instance* URI patterns. Figure 1 shows these patterns extracted from a website URIs listing classes of entity and instances of them – the entity endpoints.

CLASSES: URI PATTERNS
gtlaw.com/People
nytimes.com/topic/person
INSTANCES: ENTITY ENDPOINTS
gtlaw.com/People/Magdalena-Gad
nytimes.com/topic/person/madonna

Figure 1: Example of class and instance URIs.

We start by reviewing existing views of KBs, then discussing the content editing and publishing behaviours that we seek to exploit. To actually exploit these resources, we must first infer their existence on the web. We refer to this task as Knowledge Base Discovery (KBD) and introduce a supervised classification setting for endpoint discovery leveraging information from inbound links and silver standard mention annotation. We evaluate performance for this task using crowdsourced judgements over a held out set of candidate URIs.

This paper introduces web KBs extracted from a collection of news articles and we plan to release evaluation data, code and crowdsourced annotation. While our initial extraction is not perfect, we propose that web KBs make for compelling endpoints against which to disambiguate mentions of less prominent entities. Furthermore, we believe that a mixture of domain-specific KBs can assist entity linking to traditional KBs.

[1] http://www.imdb.com
[2] https://musicbrainz.org

Proceedings of AKBC 2016, pages 7–11,
San Diego, California, June 12-17, 2016.

2 Related Work

Entity linking and wikification have typically relied on Wikipedia (Cucerzan, 2007; Milne and Witten, 2008) or a subset (McNamee et al., 2009), or a larger structured resource such as Freebase (Zheng et al., 2012). 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. Hachenberg and Gottron (2012) 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.

Other tasks cluster mentions of the same entity, but without reference to a central KB, namely Cross Document Coreference (Bagga and Baldwin, 1998; Singh et al., 2011) and Web Person Search (Artiles et al., 2007). 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 (Hearst, 1992) document structure (Wang and Cohen, 2007; Bing et al., 2016) and site structure (Yang et al., 2010). 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 (Ji et al., 2011). 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 (Chisholm and Hachey, 2015), but does not include non-Wikipedia entities.

We examine whether we can successfully extract informal web KBs by exploiting the structure of individual URLs and the structure of the sites they describe. Like traditional linking KBs, they identify reference points against which mentions can be linked, but lack the information commonly expected in KBs.

3 Analysis of Linking Behaviour

We identify patterns of web linking behaviour producing endpoints for entity disambiguation.

Web News Some publishers maintain topic pages that aggregate structured and unstructured content on entities, e.g., `nytimes.com/topic/person/barack-obama` . These provide a landing page for search engine optimisation and enable some semantic analytics (e.g. "Do users click more on people than organisations?"). They also provide a link target to contextualise mentions in news articles and help prevent navigation away from the site. Notably, these may not include description of an entity, merely aggregate content.

Social Networks Social sites are a very rich source of entity information, e.g., `facebook.com/barackobama` . Our analysis identifies some of these endpoints. However, many links to social profiles have anchors that are not mentions of the target entity, e.g., "Find me on [Twitter]{ `twitter.com/BarackObama` }." Identifying these patterns is beyond the scope of the current work.

Organisation Directories Universites and law firms maintain directories of employee profiles, e.g., `gtlaw.com/People/Matthew-Galati` . These collect fewer inlinks than news site topic pages and social profile pages. They are nevertheless a promising source of information for entities that don't meet Wikipedia's notability requirements.

4 Knowledge Base Discovery (KBD)

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.

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, we extend this approach to leverage inlinks across a col-

lection of automatically discovered web KBs. 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.

4.1 Endpoint Inference

We explore a simple supervised classifier for KBD. For a web anchor span linking to a URI u, we wish to model the probability that it both references an entity e and is a true named entity mention m.

$$\begin{aligned} P(e, m|u) &= \frac{P(e, m, u)}{P(u)} \\ &= P(e|m, u)P(m|u) \end{aligned}$$

We approximate $P(e|m, u) \approx 1$ by assuming all mentions are entity references independent of their target URI. This allows for an estimation of our target distribution via a model which predicts the probability that links targeting u are a mention m.

$$P(e, m|u) \approx P(m|u)$$

In practice, we find this achieves good results.

4.2 Features

We represent endpoint patterns as a bag of binary features hashed to 500,000 dimensions to help manage model size. This section describes the two major categories of features used to represent instances.

Path Features We tokenize endpoint patterns by splitting on forward slash characters and include path component uni-gram and bi-grams as features. We find path tokens are good predictors of entity mentions and often generalize across KBs. For example, it is common to observe links to entity pages prefixed by terms like `profile` or `wiki`. Similarly, terms like `news` or date patterns `YYYY/MM/DD` in a URI can provide negative evidence.

Domain Features In many cases, patterns are not sufficient to identify a KB endpoint without prior knowledge. For example, `twitter` entities are only observed via a common `<domain>/<eid>` pattern. We allow the model to explicitly memorise likely KB URIs by including as features the conjunction of domain name with each bi-gram feature.

	Total	Aligned
$\|Mentions\|$	14.5	3.4
$\|URIs\|$	5.4	1.0
$\|Anchors\|$	4.4	0.6
$\|Patterns\|$	1.5	0.3

Table 1: Statistics of the corpus in millions. The first column includes all corpus links. The second column includes links whose anchor text aligns to an NER span.

While this subset of features cannot generalise to unseen domains, we are able to achieve high precision for known KBs observed in the seed corpus.

5 Experimental Setup

We validate the KBD approach described above on an internal corpus of links collected from 2,948,841 web news articles (Cadilhac et al., 2015). We leverage named entity recognition to identify likely entity references as link anchors that align to predicted mentions for person, location and organisation entity types. And we convert target URIs to endpoint patterns by normalising to lower case, removing protocol (e.g., http) and domain (e.g., sfgate.com), and removing entity identifiers (e.g., query="Elon+Musk").

Table 1 includes statistics of the full link corpus (Total) and the NER-aligned subset (Aligned). The full corpus includes a total of 14,462,659 links. 3,436,033 of these align to NER mentions, yielding 1,029,405 candidate entity endpoints across 309,182 URI patterns.

5.1 Estimating $P(m|u)$

We estimate $P(m|u)$ via logistic regression using a sample of (u, m) pairs that act as a silver standard. We consider all URI patterns with at least ten inlinks as possible training instances. We treat a URI pattern as a positive instance if a majority of inlinks from our corpus are aligned to mentions. If not, we treat it as a negative instance. To measure performance on unseen URI patterns, we group instances by domain name before partitioning. This produces a silver standard training set of 100,852 instances (10% positive), and a development test set of 10,404 (12% positive). Before training, we subsample positive instances to equal the number of negative instances.

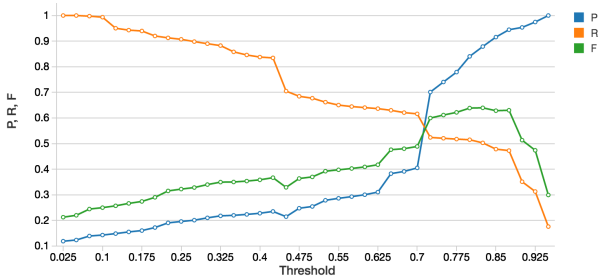

Figure 2: Precision-recall tradeoff across thresholds.

Endpoint	Entities
linkedin.com/in	3,246
variety.com/t	2,871
data.cnbc.com/quotes	2,958
si.com/nfl/player	1,426
ign.com/stars	933
cyclingnews.com/riders	899
gtlaw.com/people	257

Table 2: Sample of predicted URI patterns and entity counts.

5.2 Development Experiments

We select a threshold on held out instances from our development split. Figure 5.2 shows the precision-recall tradeoff across possible threshold values. We select a threshold of $P(m|u) >= 0.825$ here as this maximises F-score at 0.64 and is in the middle of the threshold range. Table 2 shows a sample of URI patterns predicted by this model and the number of corresponding entity endpoints discovered from the seed corpus. Encouragingly, apart from general news, we see two of the behaviour categories from Section 3: domain-specific news topic pages from Sports Illustrated and Cycling News, and professional profile pages like LinkedIn and legal web sites, which can inform disambiguation models for long-tail entities.

6 Evaluation

To evaluate how well our model for $P(m|u)$ estimates $P(e, m|u)$, we construct a corpus of human-annotated endpoint URIs. While it would be possible to randomly sample URIs, this would give us a highly imbalanced set with very few positive instances. We design a crowd task to collect pairwise identity judgements within clusters of candidate coreference pairs. To build clusters, we retrain our model over combined silver standard data (train + test) and use it to collect endpoints from the complete seed corpus with classification confidence above our threshold. We use the anchor-URI graph to build candidate clusters from randomly selected seed URIs by enumerating inlink anchors and then collecting all target URIs linked to from these anchors. We repeat this a second time to create candidate clusters based on various names for the seed URI to help account for synonymy. Ambiguity means clusters also include endpoints corresponding to different underlying entities that share a name with the seed entity. Finally, we randomly select a pair of URIs from the cluster for evaluation.

We post 500 URI pairs to Crowdflower[3] and ask three workers to judge whether each endpoint is an entity page. We also ask whether they refer to the same underlying entity. The evaluation shows that 71.2% of the 1,000 endpoints are confirmed as entities. Of the 277 pairs that include two true endpoints, 70.8% are judged as coreferent providing reasonably balanced data for evaluating future endpoint reconciliation experiments.

Finally, we estimate the extent to which our approach can be used to extend knowledge beyond standard Wikipedia KBs. We sample 100 endpoints validated in the crowd annotation and search for a corresponding Wikipedia page. 20% of endpoints represent entities that are not in Wikipedia. This suggests that the approach does discover useful knowledge further down the tail of notability.

7 Conclusion

We described an approach for discovering knowledge bases on the web — endpoints that disambiguate entity mentions. An initial endpoint classifier trained on automatically created silver standard data was validated over a corpus of 2.9 million news articles. A crowd-sourced evaluation of 1,000 endpoints found that the classifier has precision of 71.2%. Acquiring new entities is a key aspect of populating KBs, and investigation of discovered endpoints finds that approximately 20% are not in Wikipedia. Rather than simply identifying a new NIL mentions, therefore, we identify new entities to add to the KB. We hope to refine this model and apply it to larger, more diverse web corpora.

[3] http://www.crowdflower.com

References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The SemEval 2007 WePS Evaluation: Establishing a benchmark for the Web People Search task. In *SemEval*, pages 64–69.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL*, pages 79–85.

Lidong Bing, Mingyang Ling, Richard C. Wang, and William W. Cohen. 2016. Distant IE by bootstrapping using lists and document structure. In *AAAI*. to appear.

Anaïs Cadilhac, Andrew Chisholm, Ben Hachey, and Sadegh Kharazmi. 2015. Hugo: Entity-based news search and summarisation. In *CIKM Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 51–54.

Andrew Chisholm and Ben Hachey. 2015. Entity disambiguation with web links. *Transactions of the Association for Computational Linguistics*, 3:145–156.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *EMNLP-CoNLL*, pages 708–716.

Christian Hachenberg and Thomas Gottron. 2012. Finding good URLs: Aligning entities in knowledge bases with public web document representations. In *ISWC Workshop on Linked Entities*, pages 17–28.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING*, pages 539–545.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. 2011. Overview of the TAC 2011 knowledge base population track. In *TAC*.

Paul McNamee, Heather Simpson, and Hoa Trang Dang. 2009. Overview of the TAC 2009 Knowledge Base Population Track. In *TAC*.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *CIKM*, pages 509–518.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *ACL*, pages 793–803.

Richard C. Wang and William W. Cohen. 2007. Language-independent set expansion of named entities using the web. In *ICDM*, pages 342–350.

Qing Yang, Peng Jiang, Chunxia Zhang, and Zhendong Niu. 2010. Reconstruct logical hierarchical sitemap for related entity finding. In *TREC*.

Zhicheng Zheng, Xiance Si, Fangtao Li, Edward Y. Chang, and Xiaoyan Zhu. 2012. Entity disambiguation with freebase. In *WI-IAT*, pages 82–89.