Overview of the 2016 ALTA Shared Task: Cross-KB Coreference

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

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

Diego Mollá
Macquarie University
Sydney, Australia
diego.molla-aliod@mq.edu.au

Abstract

This paper presents an overview of the 7th ALTA shared task that ran in 2016. The task was to disambiguate endpoints by determining whether two URLs were referring to the same entity. We present the motivation for the task, the description of the data and the results of the participating teams.

1 Introduction

Entity endpoints are URLs which reliably disambiguate named entity mentions on the web. For example, the URL `en.wikipedia.org/wiki/Barack_Obama` may be used in reference to US president Barack Obama. Inlinks to this page are unlikely to refer to some other entity, so we should consider this a disambiguating endpoint.

While Wikipedia has been used extensively for automated entity recognition and disambiguation, many other endpoints may exist for an entity on the web. For example, `nytimes.com/topic/person/barack-obama` and `twitter.com/BarackObama` may be used equivalently. This style of systematic entity indexing is characteristic of social sources (e.g. `facebook.com/*`), news aggregation endpoints (e.g. `nytimes.com/topic/person/*`) and organisation directories (e.g. `gtlaw.com/People/*`). These resources present a valuable and largely untapped source of entity information, both in the content they host and semantic resources that may be extracted from inbound links.

The ALTA 2016 Shared task addresses the problem of Cross-KB coreference resolution. Given two candidate endpoint URLs, systems must determine whether they refer to the same underlying entity. This pairwise version of the task serves as an important precursor to the general problem of clustering web endpoints into coreferent sets. These clusters act as aggregation points for information about entities, and may used for entity centric information extraction which is not limited by the coverage constraints of any single structured KB.

This paper is structured as follows. Section 2 describes the shared task. Section 3 gives a short survey of related research. Section 4 describes the data set that was used. Section 5 details the evaluation process. Section 6 briefly describes the participating systems. Section 7 presents and discusses the results. Finally, Section 8 concludes this paper.

2 The 2016 ALTA Shared Task

The 2016 ALTA Shared Task is the 7th of the shared tasks organised by the Australasian Language Technology Association (ALTA). Like the previous ALTA shared tasks, it is targeted at university students with programming experience. The general objective of these shared tasks is to introduce university students to the sort of problems that are the subject of active research in a field of natural language processing.

There are no limitations on the size of the teams or the means that they can use to solve the problem, as long as the processing is fully automatic — there should be no human intervention.

There are two categories: a student category and an open category.

- All the members of teams from the **student category** must be university students. The teams cannot have members that are full-time employed or that have completed a PhD.
- Any other teams fall into the **open category**.

The prize is awarded to the team that performs best on the private test set — a subset of the evaluation data for which participant scores are only revealed at the end of the evaluation period (see Section 5).

Andrew Chisholm, Ben Hachey and Diego Mollá. 2016. Overview of the 2016 ALTA Shared Task: Cross-KB Coreference. In *Proceedings of Australasian Language Technology Association Workshop*, pages 161–164.

3 Related Work

Entity disambiguation work has traditionally focused on the reconciliation of textual mentions to records in a centralised KB like Wikipedia (Cucerzan, 2007) or Freebase (Zheng et al., 2012). In this case, the domain of linkable entities is limited by the coverage of the target knowledge base and those which fall outside this domain are classified as NILs. NIL mention clustering is often addressed separately, and has been the focus of Text Analysis Conference (TAC) Knowledge Base Population shared tasks since 2011 (Ji et al., 2011).

A more generalised approach to resolving mention ambiguity is that of cross-document coreference resolution — where systems cluster mentions of the same entity together without reference to a central KB (Bagga and Baldwin, 1998; Singh et al., 2011). Both NIL clustering and cross-document coreference deal with ambiguity at the mention level. In contrast, the task of cross-KB coreference resolution deals with entity coreference at the KB level, by attempting to cluster entity records across distinct KBs.

This task is similarly structured to that of record linkage (Fellegi and Sunter, 1969; Xu et al., 2013). But, where record linkage commonly operates over structured databases, cross-KB coreference relies primarily on unstructured nodes as input. Cross-KB coreference can draw on the content of the entity endpoint, including the URL, the text and any structured or semi-structured data inside the endpoint page. It can also draw on the web's hyperlink graph, e.g., collecting mentions in context from pages that link to an entity endpoint.

In the web domain, work on finding links associated with existing KB entities (Hachenberg and Gottron, 2012) and web person search (WePS) (Artiles et al., 2007) is also closely related. WePS takes the output of a web search for some entity name and attempts to cluster the results that refer to the same underlying entity. The ultimate aim of cross-KB coreference is also to cluster web pages. By contrast, however, it focuses on clustering entity endpoint pages instead of entity mention pages.

The task builds in part on the Knowledge Base Discovery (KBD) system of Chisholm et al. (2016), where the existence of web endpoints may be inferred from their usage on the web. Shared task data and evaluation are described below.

4 Data

Constructing a balanced corpus of endpoint URL pairs which present non-trivial cases of entity ambiguity is a challenging task. Randomly sampling from a corpus of web links is insufficient as any two URLs are unlikely to refer to the same entity, leading to a highly imbalanced dataset of negative samples. Conversely, if we constrain our sampling to pairs linked from similar anchor text, almost all pairs will be coreferent since entity mentions follow a Zipf-like distribution corresponding to notability.

To address these challenges, we target entity names at the low end of the notability distribution where the ratio of URLs per entity is small in comparison to the general corpus. We train the KBD system of Chisholm et al. (2016) over a corpus of 14.5 million news article outlinks and extract high-confidence endpoints where $P(entity|url) \geq 0.825$. We construct a bipartite anchor-endpoint graph and keep only those anchors that link to only one endpoint URL. These anchors constitute a corpus of long-tail entities names.

We sample 1,000 names from this collection and use the Bing Web Search API[1] to search the web for links corresponding to each anchor span. From each search, we take the first two result URLs which are classified by KBD as entity endpoints and use this as a candidate entity URL pair for the shared task. We also record the page title and search engine snippet returned by the Bing Search API for each instance. Next, we filter out instances of result pairs which both come from the same domain, as these samples typically represent trivial cases of non-coreference. Finally, URL pairs are manually annotated to filter out erroneous endpoint classifications and judge coreference. We randomly sample 200 positive and 200 negative pairs from this set and shuffle them into equal train and test splits.

We observe most endpoints originate from social sources `linkedin.com` and `twitter.com`, while a moderate amount come from more traditional KB-style sites like `imdb.com` and `tripadvisor.com`. The remainder come from a variety of news sources (e.g. `sports.yahoo.com`, `forbes.com`) and small online directories (e.g. `psychology.nova.edu`).

[1] `http://www.bing.com/toolbox/bingsearchapi`

5 Evaluation

The shared task was managed and evaluated using the Kaggle in Class framework, with the name "ALTA 2016 Challenge"[2]. The Kaggle in Class site was created as an invitation-only competition, where the participants could post questions and comments, and submit trial runs and the final submission.

As is standard in Kaggle-in-class competitions, the data set was partitioned into a training set, a public test set, and a private test set. The training set contained 200 pairs of URLs and their labels, and was made available to the participants. The public and private test sets contained 100 new unlabeled pairs of URLs each and were combined into a single test file. The participants were asked to submit the labels of the combined test set. The evaluation results of the public test set were available as soon as the results were submitted, and the evaluation results of the private test set were not made available until after the final deadline. This way, the participants could obtain instant feedback with the public test set, and the risk of overfitting to the final results was diminished.

Evaluation uses the F1 score for the positive class. This is similar to pairwise F1 sometimes used for evaluation of entity resolution (Winkler, 2006), except calculated here over pairs listed in the data only. Precision is the ratio of true positives tp (the number of pairs of endpoints that were correctly labelled as coreferring) to all predicted positives (the total number of pairs of endpoints that the system labelled as coreferring, computed as the sum of true positives and false positives fp). Recall is the ratio of true positives to all actual positives (the number of pairs of endpoints that are coreferring according to the test data, computed as the sum of true positives and false negatives fp). The formula of the F1 score is:

$$F1 = 2\frac{p \cdot r}{p + r}$$

where

$$p = \frac{tp}{tp + fp}, \quad r = \frac{tp}{tp + fn}$$

The product $p \cdot r$ in the numerator of the formula will tend to reward systems that are moderately good in both recall and precision, whereas systems that do extremely well in one and poorly in the other would achieve a lower F1 score.

[2] https://inclass.kaggle.com/c/alta-2016-challenge

6 Systems

This section presents short descriptions of some of the participating systems. For further details, refer to the shared task section of the proceedings of the 2016 ALTA workshop.

6.1 EOF

The system by team EOF (Khirbat et al., 2016) follows a two-stage approach. First, in the entity endpoint determination stage, the system determines the most likely underlying entities being referred to by each URL. Second, in the entity disambiguation stage, the two endpoints are disambiguated. Entity endpoint determination is achieved by extracting the named entities of the text pointed by the URL using the Stanford NER, and ranking the entities using logistic regression. The top 3 entities are passed to the entity disambiguation stage, together with additional features based on the URLs, anchor texts of the URLs, and the text pointed by the URLs. This information is processed by a tree ensemble classifier.

6.2 NLPCruise

The system by team NLPCruise (Shivashankar et al., 2016) also follows a two-stage approach but in a different manner. The first stage is a filtering step that rules out cases of dissimilar entities. Those URL pairs which pass the filter pass through to the second stage for more sophisticated processing. The filtering step uses the Stanford NER for detecting the named entities of the titles and the URL pairs. The second stage uses an ensemble of 3 classifiers: one based on Bing search results of the named entities, another classifier based on short-text semantic similarity, and a third classifier that uses additional features extracted from the text pointed to by the URL. An interesting aspect of their system is the use of machine translation as a means to compute semantic similarity, by computing the probability that one text translates into the other.

6.3 BCJR

The systen by team BCJR (Yu et al., 2016) uses a statistical classifier that takes as input features from the pair of URLs. The features are based on the word, character and character bigram embeddings of the text pointed by the URLs. The team has also made available an expanded training data set with about 1700 training pairs.

System	Category	Public	Private
EOF	Student	**0.91**	**0.86**
NLP-Cruise	Student	0.86	0.78
LookForward	Student	0.89	0.78
BCJR	Student	0.75	0.69
ZZ	Student	0.81	0.67
(Baseline 1)		0.67	0.66
STEM	Open	0.79	0.64

Table 1: F1 on the public and private test sets.

6.4 Baseline 1

This is a trivial baseline system provided by the organisers of the shared task. This system returned 1 (the URLs are co-referring) for every instance.

7 Results

Table 1 shows the results of the public and the private test sets. The results are sorted by the outcome of the private test set.

Results from the top three systems in the range $[0.78, 0.86]$ are encouraging, suggesting that cross-KB coreference can be performed with good accuracy even for long-tail entities. Overfitting is a particular challenge with the small data set here and we observe changes from -0.05 to -0.15 F1. Some changes affect system ordering between public and private data, indicating that good generalisation is important to success on this task.

8 Conclusions

The 2016 ALTA Shared Task was the 7th of the series of shared tasks organised by ALTA. This year's shared task focused on cross-KB coreference, and the participants were asked to determine whether two URLs were referring to the same entity. Teams used an array of techniques including logistic regression, ensemble classifiers, and training set aggregation.

The training data set was small, with only 200 pairs of URLs. The small training and test sizes might have caused some of the systems to overfit to the public test set, but overall very good results were achieved. The winning team EOF achieved an F1 score of 0.91 in the public test set, and 0.86 in the private test set, while the second and third teams achieved 0.78 in the private test set.

For full details on participating systems, refer to the shared task section of the 2016 ALTA workshop proceedings.

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.

Andrew Chisholm, Will Radford, and Ben Hachey. 2016. Discovering entity knowledge bases on the web. In *NAACL Workshop on Automated Knowledge Base Construction*, pages 7–11.

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

Ivan P Fellegi and Alan B Sunter. 1969. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

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.

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

Gitansh Khirbat, Jianzhong Qi, and Rui Zhang. 2016. Disambiguating entities referred by web endpoints using tree ensembles. In *ALTA*.

S. Shivashankar, Yitong Li, and Afshin Rahimi. 2016. ALTA shared task 2016: Filter and match approach to pair-wise web URI linking. In *ALTA*.

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.

William E. Winkler. 2006. Overview of record linkage and current research directions. Technical Report Statistics #2006-2, U.S. Bureau of the Census.

Ying Xu, Zhiqiang Gao, Campbell Wilson, Zhizheng Zhang, Man Zhu, and Qiu Ji. 2013. Entity correspondence with second-order markov logic. In Xuemin Lin, Yannis Manolopoulos, Divesh Srivastava, and Guangyan Huang, editors, *WISE*, pages 1–14.

Cheng Yu, Bing Chu, Rohit Ram, James Aichinger, Lizhen Qu, and Hanna Suominen. 2016. Pairwise text classifier for entity disambiguation. In *ALTA*.

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