Pairwise Webpage Coreference Classification Using Distant Supervision

S. Shivashankar[1*], Timothy Baldwin[2*], Julian Brooke[3*], and Trevor Cohn[4*]

[*]Computing and Information Systems, The University of Melbourne, Australia
[1]shivashankar@student.unimelb.edu.au
[2,3,4]{tbaldwin,julian.brooke,t.cohn}@unimelb.edu.au

ABSTRACT

A person or other entity is often associated with multiple URL endpoints on the web, motivating the task of determining whether a given pair of webpages is coreferent to a given entity. To strike a balance between unsupervised and supervised methods that require annotated data, we build a positive and unlabelled (PU) learning model, where we obtain positive examples using web search-based distant supervision. We evaluate our proposed approach using the SemEval-2007 WePS and ALTA-2016 shared task datasets.

Keywords

Webpage Coreference; End-point Disambiguation; Distant Supervision; PU Learning; Semi-supervised Learning

1. INTRODUCTION

Entity endpoints are URLs which reliably disambiguate named entity mentions on the web [1]. Target entity can be a person, organisation, location, event or concept. Entity linking systems typically rely on semantic resources such as Wikipedia or DBPedia as end-points. Though they provide rich context for entities, such sources do not have high coverage for many domains. In addition to traditional sources, fortunately many other end-points exist for an entity on the web: for instance, social network sites (e.g. `facebook.com/*`), news aggregation endpoints (e.g. `nytimes.com/topic/person/*`) and organisation directories (e.g. `gtlaw.com/People/*`) are largely untapped sources of valuable entity information. Each source can be treated as a knowledge base (KB) containing entity-related unstructured data (webpages) in the form of URL end-points. Building on automatic approaches for discovering web KBs [1], we focus on the downstream challenge of end-point reconciliation across KBs. Specifically, similar to the ALTA 2016 shared task[1], given two candidate endpoint URLs, the task is to determine whether they refer to the same underlying entity. This can be seen as a precursor to grouping web endpoints into coreferent sets. This relates to work on web person search (WePS), i.e. the unsupervised task of clustering entity mention pages instead of entity endpoint pages, i.e., the pages can have content related to multiple entities.

Rather than building a completely unsupervised model [1, 2, 5] or a fully supervised model as in the ALTA shared task, we propose an approach to generate web search-based distantly supervised positive examples, and employ a positive and unlabelled (PU) learning algorithm [3, 4]. The intuition underlying our method is that, if we query for a target entity name e with the addition of a disambiguating keyword extracted from URL_a, and URL_b appears in the top search results, then URL_a and URL_b are highly likely to refer to the same entity e (i.e. be coreferent). If, on the other hand, URL_b is not included in the top search results, it is not possible to draw any conclusion as to whether the two URLs are coreferent or not. For example, while Wikipedia and `biography.com` both contain end-point URLs for singer George Clinton in the form of W1 and B1, respectively, the search query `"George Clinton" AND P-Funk` — where `P-Funk` is a disambiguating term extracted from W1 — does not list B1 as one of the top-ranked results.

[1]`http://alta.asn.au/events/sharedtask2016/`

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054224

2. PROPOSED APPROACH

We assume access to web URL end-point KBs based on automatic crawling [1]. From such KBs, given a training dataset D with web URL end-point ("URL" henceforth) pairs from different KBs that refer to the same entity name (e.g. *Michael Jordan*) but are not necessarily coreferent to the same entity, our objective is to learn a pairwise URL classifier. For a URL pair U_i and U_j, we learn a model $f(x) \rightarrow y$ where the target $y \in \{0, 1\}$ denotes whether the URL pair refers to the same entity ($y = 1$) or not ($y = 0$). x is a feature map for generating pairwise URL document features ($x = \phi(U_i, U_j)$). Initially, all pairs are unlabelled ($D_u = D$), and the positive and negative labelled sets are empty ($D_p, D_n = \emptyset$).

2.1 Distant Supervision

We construct web search queries for distant supervision as follows: Q_i**:** Using the target entity name and context information from U_i, in the form of person name (PER) and organisation (ORG) named entity instance extracted from U_i by the Stanford CoreNLP NER toolkit. Similar to [5], this takes the form of each of the first 8 NEs occurring in the title and document body (for a total of up to 8 separate queries), based on the intuition that the NEs that are most

relevant to the target entity will be mentioned earlier in the document. Q_j: Similar to the above, but we generate the context information from U_j.

All of our experiments are based on the DuckDuckGo search engine. For each query in Q_i, we check to see if U_j is present in the top-K search results (we use $K = 30$), and conversely if U_i is present in the top-K results for each query in Q_j. If the given URL is found in the result set for at least one of these (up to 16) queries, we assign the distant label $\hat{y_{ij}} = 1$ to U_i and U_j. These positive instances are added to D_p and removed from D_u.

2.2 Labelling Unlabelled Pairs

Since the proportion of positive examples may not be large, we supplement D_p as follows:

- Step 1: Randomly select N instances from D_p, and held them out in S_p.
- Step 2: Train a binary classifier θ, taking $D'_p = D_p \setminus S_p$ as positive instances and D_u as negative instances. We use a linear-kernel SVM classifier in our experiments.
- Step 3: Compute the average (probabilistic) score of instances in S_p assigned by θ (using Platt scaling), $\mu_p = \frac{1}{|S_p|} \sum_{i:S_p} p(x_i = 1)$. Form the set D^*_p of instances in D_u where the θ-assigned score exceeds this threshold, i.e. $D^*_p = x_u \in D_u : p(x_u = 1) > \mu_p$.

This is a variant of the approaches of [3] and [4], and results in a training set of positive instances ($D_p \cup D^*_p$) and negative instances ($D_u \setminus D^*_p$), over which any standard binary classifier can be trained. In our experiments, we use a linear kernel SVM, which we refer to as "DP-SVM" in Table 1, since it uses propagated distant labels.

3. EXPERIMENTS

We present results over the SemEval-2007 WePS[2] development set and ALTA-2016 shared task datasets. For WePS, we constructed a balanced 1000 URL-pair dataset from webpages for 49 people. We only included pairs sharing the same person name from webpages that still exist. Additionally, we considered end-point pages only (filtered using features from [1]). We used a random split of 70/30 for training and testing. The ALTA dataset is a balanced set which contains 400 pairs of URLs that can refer to any entity – person, organisation, location, etc. We used 300 pairs for training, and used the heldout test set of 100 pairs for testing.

Our features are based on those proposed by the best performing systems in the respective shared tasks. In brief, the 8 features used for WePS include: unigram cosine similarity, n-gram cosine similarity, named entity-based cosine similarity (obtained using Stanford CoreNLP NER), title unigram Jaccard Coefficient score, URL character 4-gram Jaccard coefficient, document-level cosine similarity based on an average word-level word2vec representation, document length difference, and URL path length difference. We used the same set of features for the ALTA dataset, in addition to semantic similarity[3] and machine translation metric-based similarity[4] (BLEU, METEOR and TER) between pairs of Bing search snippet text provided in the dataset (12 features in total).

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

[3] https://dandelion.eu/semantic-text/text-similarity-demo/

[4] https://github.com/jhclark/multeval

Dataset	BSVM	Spy-SVM	SPUL	HC	DP-SVM
WePS	0.472	0.630	0.516	0.408	**0.653**
ALTA	0.500	0.540	0.587	0.481	**0.608**

Table 1: Micro-averaged F-score

Using distant supervision (Section 2.1) we got around 10% of positive examples for WePS and around 25% for ALTA. With D_p as positive and D_u as negative examples, we use biased SVM ("BSVM") with different costs for positive and negative classes as a simple baseline. PU Learning baselines with D_p and D_u include: Spy-SVM with recommended parameters [4] and SPUL [3]. All SVM-based models are based on a linear kernel. We also include an unsupervised approach based on hierarchical clustering (referred to as "HC") which has been shown to perform well for the WePS dataset [2]. We set the number of clusters to 2, and use a brute-force technique to minimise overall error in assigning clusters to classes. Hyperparameters for the SVM (C) and biased-SVM (C and class weights) were tuned based on cross-validation. For our proposed method ("DP-SVM"), based on Sections 2.1 and 2.2, we provide average scores across five runs with five random positive sets S_p. We present the micro-averaged F-score for the test set of the two datasets in Table 1.

The distant supervision-based model based on Section 2.1 ("BSVM") performs better than the unsupervised approach (HC). SPUL is sensitive to the ratio of positive examples ($|D_p|$ vs. $|D_u|$), and performs better on the ALTA dataset than the WePS dataset. Overall, our proposed approach performs better the other competing methods. We evaluated the upper bound performance of a supervised SVM model built with manually annotated labels, and found it to be around 5% greater than our proposed approach in absolute terms on both datasets.

4. CONCLUSIONS

We have proposed an approach to determining whether two endpoint URLs refer to the same entity, with two key contributions: (a) the use of distant supervision; and (b) the application of PU Learning to the task. To the best of our knowledge, this is the first attempt to leverage distant supervision in conjunction with PU Learning.

5. REFERENCES

[1] A. Chisholm, W. Radford, and B. Hachey. Discovering entity knowledge bases on the web. In *Workshop on Automated Knowledge Base Construction*, 2016.

[2] A. D. Delgado, R. Martínez, V. Víctor Fresno, and S. Montalvo. A data driven approach for person name disambiguation in web search results. In *COLING*, 2014.

[3] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *SIGKDD*, 2008.

[4] B. Liu, W. S. Lee, and X. Li. Partially supervised classification of text documents. In *ICML*, 2002.

[5] 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.