Global Entity Ranking Across Multiple Languages

Prantik Bhattacharyya, Nemanja Spasojevic
Lithium Technologies | Klout
San Francisco, CA
{prantik.bhattacharyya, nemanja.spasojevic}@lithium.com

ABSTRACT

We present work on building a global long-tailed ranking of entities across multiple languages using Wikipedia and Freebase knowledge bases. We identify multiple features and build a model to rank entities using a ground-truth dataset of more than 10 thousand labels. The final system ranks 27 million entities with 75% precision and 48% F1 score. We provide performance evaluation and empirical evidence of the quality of ranking across languages, and open the final ranked lists for future research.

Keywords

entity ranking; entity extraction; knowledge base;

1. INTRODUCTION

In the past decade, a number of openly available Knowledge Bases (KBs) have emerged. The most popular ones include Freebase, Wikipedia, and Yago, containing around 48M, 25M, and 10M entities respectively. Many of the entities overlap across the KBs. In NLP entity linking [1], the task is to link mentioned entities within text to their identity within the KB. A foundational part of setting up a real-time entity linking system is to choose which entities to consider, as memory constraints prohibit considering the entire knowledge base [1]. Additionally, some entities may not be of relevance. In order to maximize quality of the NLP entity linking system, we need to include as many important entities as possible.

In this paper we identify a collection of features to perform scoring and ranking of the entities. We also introduce the ground truth data set that we use to train and apply the ranking function.

2. RELATED WORK

A large body of previous work has addressed ranking entities in terms of temporal popularity, as well as in the context

---

[1]also known as named entity linking (NEL), named entity disambiguation (NED) or named entity recognition and disambiguation (NERD)

of a query; however, little study has been done in terms of building the global rank of entities within a KB. Temporal entity importance on Twitter was studied by Pedro et. al. [4]. In [2], authors propose a hybrid model of entity ranking and selection in the context of displaying the most important entities for a given constraint while eliminating redundant entities. Entity ranking of Wikipedia entities in the context of a query, has been done using link structure and categories [5], as well as graph methods and web search [6].

3. OUR APPROACH

Given KB, we want to build a **global long-tailed ranking** of entities in order of socially recognizable importance. When building the NLP entity linking system, the N top ranked entities from KB should yield maximum perceived quality by casual observers.

3.1 Data Set

We collected a labeled data set by selecting $10,969$ entities. We randomly sampled as well as added some important entities, to balance the skewed ratio that KBs have of important / non-important entries. Each evaluator had to score the entities on scale 1 to 5; 5 being most important. Seven evaluators used the following guidelines regarding importance:

Public Persons important if currently major pro athletes, serving politicians, etc. If no longer active, important if influential (e.g. Muhammad Ali, Tony Blair).

Locations look at population (e.g. Albany, California vs. Toronto, Canada), historical significance (Waterloo).

Dates unimportant unless shorthand for a holiday or event (4th of July, 9/11).

Newspapers important, especially high-circulation ones (WSJ).

Sports Teams important if in pro league.

Schools important if recognised globally.

Films & Song major franchises and influential classics are important – more obscure are often not.

Laws important if they enacted social change (Loving v. Virginia, Roe v. Wade), unimportant otherwise.

Disambiguators entities that disambiguate are important because we want them in the dictionary (Apple, Inc. and Apple Fruit).

3.2 Features and Scoring

Features were derived from Freebase and Wikipedia sources. They capture popularity within Wikipedia links, and how important an entity is within Freebase. Some signals used are page rank, link in/out counts and ratio, number of categories a page belongs to in Wikipedia. We also use the

Table 1: Feature Performance For English Rankings

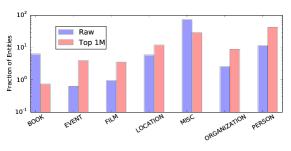|	Feature	P	R	F1	C	RMSE
Wikipedia	Page Rank	0.59	0.05	0.09	0.164	1.54
	Outlink Count	0.55	0.13	0.21	0.164	2.09
	Inlink Count	0.62	0.12	0.20	0.164	1.82
	In Out Ratio	0.75	0.19	0.31	0.164	1.54
	Category Count	0.65	0.21	0.36	0.164	1.89
Freebase	Subject #	0.28	0.06	0.10	1.000	2.39
	Subject Types #	0.42	0.10	0.16	1.000	2.25
	Object #	0.62	0.12	0.20	0.973	2.00
	Object Types #	0.46	0.11	0.17	0.973	2.25
	Klout Score	0.57	0.11	0.17	0.004	2.32
	$\mathcal{S}(e)$ - All Feat.	**0.75**	**0.37**	**0.48**	**1.00**	**1.15**

Figure 1: Entity Count by Type

number of objects, a given entity is connected to, i.e., object and object type count features, as well as the number of times a given entity was an object with respect to another entity, i.e., subject and subject type count features. We also extract social media identities mentioned in an entity's KB and use their Klout score [3] as a feature. The full set of features derived as well as their performance is listed in Table 1.

We model the evaluator's score using simple linear regression. The feature vector $\mathcal{F}(e)$ for an entity e is represented as: $\mathcal{F}(e) = [f_1(e), f_2(e), ..., f_m(e)]$ where $f_k(e)$ is the feature value associated with a specific feature f_k. Normalized feature values are denoted by $\hat{f}_k(e)$. Features are normalized as: $\hat{f}_k(e) = \frac{log(f_k(e))}{\max\limits_{e_i \in KB} log(f_k(e_i))}$. Importance score for an entity is denoted by $\mathcal{S}(e)$ and is computed as the dot product of a weight vector $\mathbf{w}$ and the normalized feature vector: $\mathcal{S}(e) = \mathbf{w} \cdot \hat{\mathcal{F}}(e)$ (1). Weight vector is computed with supervised learning techniques, using labeled ground truth data (train/test split of 80/20).

4. EXPERIMENTS

Table 1 shows precision, recall, F1 and the population coverage for the full list of features and the final system. The importance score was calculated using Eq.1 where final score was rounded to an integer value so it can be compared against the labels from ground-truth data.

We observe that Wikipedia features have the highest precision among all the features. The Freebase features have the highest coverage values. The Klout score feature also has one of the highest individual precision values. While this feature has the lowest coverage, it helps boost the final score and floats up a few relevant entities for final system application in social media platforms. We also look at root mean squared error (RMSE) of the entity scores against assigned labels. The final model shows the lowest RMSE value.

We also plot the distribution of entity types in the top 1 million ranked entities and the unranked list for the English language. 11% of entities are of type 'person' in the global

Entity	Image	EN	AR	ES	FR	IT
Vogue		2	6,173	200	2,341	62
World Bank		322	103	3,747	2,758	5,704
Morocco		1,277	2	527	544	232
Donald Duck		10,001	9,494	7,444	10,380	4,575
Balkans		36,753	109	17,456	9,383	2,854
Bed		109,686	23,809	68,180	66,859	52,713
Bunk Bed		992,576	64,399	330,669	906,988	416,292

Table 2: Entity Ranking Examples For Different Languages

list while the top ranked list contains 42% entities of type 'person'. The percentage of 'MISC' entity types drop from 72% to 29%. These difference in coverage highlight that entities are ranked relevantly in the corpus.

In Table 2, we provide examples of entities with their ranks in a particular language. We see that the entity ranks are regionally sensitive in the context of their language, e. g. 'Morocco' is ranked 2 in the ranking for 'Arabic' language. We also observe the rankings are sensitive with respect to the specificity of the entity, for example 'bunk bed' is ranked magnitudally lower than the more generic entity 'bed'.

5. SUMMARY

We make the ranked list of top $500,000$ entities available as an open source data set at `https://github.com/klout/opendata`. To conclude, in this work, we built a global ranking of entities across multiple languages combining features from multiple knowledge bases. We also found that combination of multiple features yields the best results. Future work in this direction is to include new signals such as Wikipedia page view statistics and edit history.

6. REFERENCES
[1] P. Bhargava, N. Spasojevic, and G. Hu. High-throughput and language-agnostic entity disambiguation and linking on user generated data. In *Proceedings of the 26th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017.

[2] A. Gionis, T. Lappas, and E. Terzi. Estimating entity importance via counting set covers. In *18th Intl. Conf. on Knowledge Discovery and Data Mining*, 2012.

[3] A. Rao, N. Spasojevic, Z. Li, and T. Dsouza. Klout score: Measuring influence across multiple social networks. In *IEEE Intl. Conf. on Big Data*, 2015.

[4] P. Saleiro and C. Soares. Learning from the news: Predicting entity popularity on twitter. In *International Symposium on Intelligent Data Analysis*, 2016.

[5] A.-M. Vercoustre, J. A. Thom, and J. Pehcevski. Entity ranking in wikipedia. In *ACM symposium on Applied computing*, 2008.

[6] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking very many typed entities on wikipedia. In *16th ACM conference on Conference on Information and Knowledge Management*, 2007.