A Web-centric entity-salience based system for determining Notability of entities for Wikipedia

Gokul Thota
IIIT Hyderabad

Rahul Khandelwal
IIIT Hyderabad

Vasudeva Varma
IIIT Hyderabad

Abstract

The content on Wikipedia is growing rapidly, and it is essential to moderate this uploaded content, for the efficient functioning of Wikipedia. We utilize the "Notability" guidelines, decided by the editors of Wikipedia, and design a system to automatically identify if an entity warrants its own article on Wikipedia or not. This system is based on web-based entity features and their text-based salience encodings, and functions irrespective of an entity's category.

Keywords: Notability, Entity Salience, Web Search, Dataset, Classification

Introduction

An average of 559 new articles are being uploaded to Wikipedia every day. With such a large rate of content creation, it is key to ensure that only important information is included. In order to identify the importance of the article of any given entity, a test called Notability is defined by the editors of Wikipedia. Notability detection can be viewed as a binary classification task, where any given entity would have its own label, indicating if it is notable or not. Although, performing this check for a large number of entities of varying categories, requires high manual effort.

We categorize any named entity to belong to one of the two partitions - **Generic** and **Abstract**. Reliable web domains can be identified for categories belonging to the Generic setting, such as Film Actors, Cricketers, etc., whereas such a pin-pointing of domains cannot be done for categories of Abstract setting, such as Biological concepts. We define a system centered around content corresponding to an entity on the web, which detects the Notability label irrespective of the entity's category and its setting (Generic or Abstract). We construct two datasets to validate our approach, one for each setting.

Methods

For the Generic setting, we initially collect a list of reliable web domains for a category (such as IMDb for films) and verify an entity's coverage in this domain. For identifying this coverage, we utilize a set of hand-crafted entity salience features (Pochampally and Karlapalem, 2017). Additionally, we focus on a more generalizable set of signals from the web and augment these features to the domain-specific features obtained above. We extract details about an entity's presence in the Wikipedia ecosystem - relevant documents count on Wikidata, relevant images count on Wiki-commons, and relevant articles in Wikipedia. We formulate a search query with entity title, category name and its related keywords, to ensure disambiguation of retrieved results. Jaccard threshold is utilized to perform entity-entry matching, to check if a webpage corresponds to an entity. A summary of query logs pertaining to an entity (over a period of time) is obtained from Google Trends, which is also included in the feature set. To extract signals from online news, a set of reliable international news web domains are chosen, and the number of news articles corresponding to the entity of interest is extracted. Further, an entity's presence in social media is also gauged using the followers count feature of Instagram and Twitter.

The 'Abstract' setting is similar to the above case, except instead of domain-specific features (which are absent here), the distribution of information about the entity, on the web, is extracted. This information is extracted by analyzing the salience of a selected set of documents, which are in the top few retrieved results on a search engine. Additionally, social media and online news-based features discussed in the Generic setting are omitted, due to their ineffective applicability.

The content extracted is passed through a classification pipeline, which utilizes BERT (Devlin et al., 2018) to obtain text encodings. Text encodings are extracted from relevant documents about an entity, to capture the quality of information centered around the entity. Categorical embeddings and entity descriptions are also obtained. These salience encodings, along with individual count features above (such as the number of relevant documents) are passed through a feed-forward neural network. This network produces a single label output, which indicates if the entity is notable or not. This architecture is depicted in figure 1. Binary cross-entropy is the loss function used, for training the model.

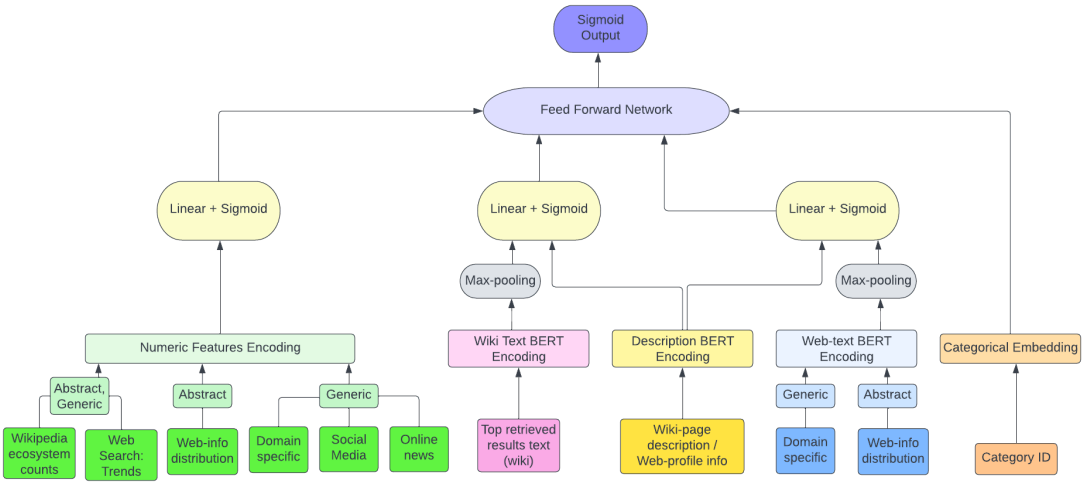

Figure 1: Classification Architecture

We construct 2 datasets for validation, one for the Generic setting and one for the Abstract setting. The Generic setting dataset includes 9 diverse categories such as Films, Cricketers, Medicinal plants, etc., and contains 30k samples in total (16.5k are notable). The Abstract setting dataset includes 5 diverse categories such as Biological concepts, Ragas in the Carnatic system, etc., and contains nearly 5k samples (3.1k are notable). The train-test-validation random split is 70-15-15.

Results

We compare our approach with the baseline which utilizes only domain-specific features (Pochampally and Karlapalem, 2017). The standard accuracy metrics such as Precision, Recall, F1 score, and Accuracy are used for comparing the performance. The results achieved in both settings can be visualized in table 1.

Table 1 Generic Setting (first) and Abstract setting (second) results

	ACC	PR	REC	F1
Baseline	85.38	85.29	85.12	85.20
Our system	**94.86**	**94.93**	**94.68**	**94.79**

	ACC	PR	REC	F1
Baseline	66.75	61.97	59.12	59.19
Our system	**80.79**	**79.63**	**76.84**	**77.86**

It can be observed that there is an improvement in performance accuracy (about 9% in Generic setting and 14% in Abstract setting) for our proposed approach, in comparison with the baseline approach of domain-specific features, in both settings. Additionally, the performance was observed to be consistent across all categories in both datasets, validating the extendability of the system to new categories.

Discussion/Conclusions

Our system significantly outperforms the previous work in the field of Notability Determination of entities for Wikipedia. It also addresses the shortcomings of the baseline, of handling entities representing abstract concepts, which is achieved by the Abstract setting.

A drawback of our system is that it is primarily designed for singular-named entities. There exist articles in Wikipedia that have complex titles, which could have multiple entities or concepts, or could have a non-trivial relationship with the category it belongs to (such as an Island's article in the "Birds" category). Hence, it is essential to first identify if a given title is simple or complex, by examining the attributes of entities involved in it. We are working on incorporating these aspects into the system, by using a graph-based approach for correlating entities and concepts.

References

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

[Pochampally and Karlapalem2017] Yashaswi Pochampally and Kamalakar Karlapalem. 2017. Notability determination for wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 1641–1646, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.