Hierarchical Supervised Topic Selection for Indian Knowledge Resource Building

Team Zero - Arghya B. , Neha M., Sanjana S., Sathvik B.

September 2019

Identifying topics/pages to be included in Indian Language Wikipedia

1 Introduction

Wikipedia has identified around 10,000 Vital Articles that need to be present in every language [1]. There is a vital need to improve the size of Wikipedia of Indian Languages, as they are only a very tiny fraction in size when compared to English Wikipedia. This project aims to improve the content of Indian Topics in Wikipedia, by identifying new and relevant topics to be added to any Indian Language.

1.1 Problem Statement

Use Indian language news papers, text books and current affairs content as a set of sources to identify the new topics to be added to any Indian language Wikipedia. In order to qualify to be a Wikipedia page, there has to be enough evidence or set of references - what are the topics that has these kinds of evidences/references but does not have a Wikipedia page

1.2 Project Plan Outline

Using the above 10,000 topics, we would like to build a classifier, with these topics as positive samples. New topics in the context of regional languages are then extracted from various sources using various heuristic methods and the classifier is run again to obtain the final list of relevant topics.

2 Related Work

Not much research has been done in terms of identifying topics that are currently trending and would be relevant enough for an informational platform like

[1] https://simple.wikipedia.org/wiki/Wikipedia:List_of_articles_all_languages_should_have/Expanded

Wikipedia. Additionally, most implementations are focused on the English language.

Cucerzan (2007) described recognition of named entities based on information extracted from Web search results. Upadhyay et al. (2016) introduced an event extraction approach that helps identify the dominant event of a particular news article using a rule based classifier. Choubey et al. (2018) tackled a similar problem by mining event conference relations.

3 The HIRB model

The HIRB model can be divided into two work flows:

- Identifying candidates by training
- Finalization of topics

3.1 Training Work Flow

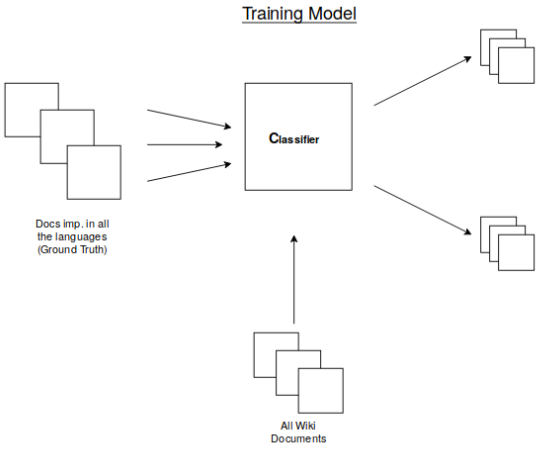

Figure 1: Training Work Flow

The training model (as pictured in Fig. 1) consists of a classifier. The topics that have been identified as necessary, regardless of language, are the positive samples. All of the existing Wikipedia topics can be run through the classifier and the positive predicted topics become the candidates.

3.2 Finalization of topics

To extract new topics in the context of local languages, we would need to

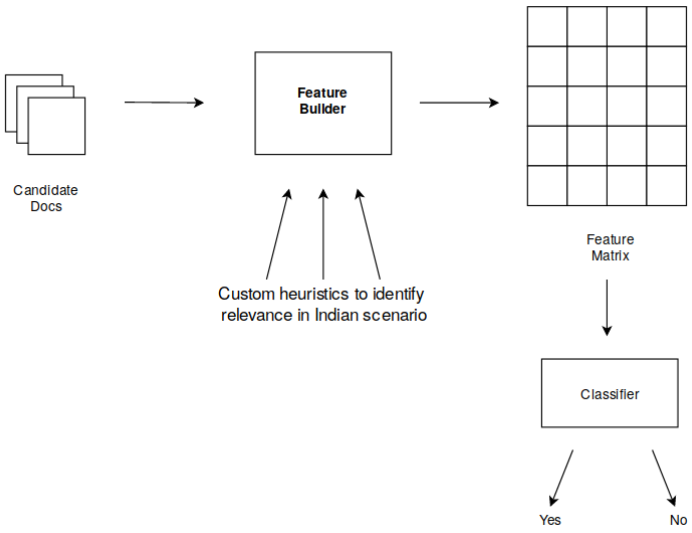

Figure 2: Finalizing Topics

- Scrape local language newspapers for news
- Blogs can be scraped for current affairs

Text books are a good source of fundamental topics that need to be available but OCR would be needed to extract text from PDF. These topics would give a brief idea. Additionally data in newspapers, blogs and books is unstructured in nature, so after scraping techniques such as Named Entity Recognition, semantic analysis and context interpretation to determine the event or main subject which in turn constitute a set of possibly important topics.

A custom feature set (as pictured in Fig. 2) is built from this set and the classifier is run again to extract topics that are relevant in an Indian Context.

3.3 Tools

- Python (Programming language)
- Selenium (for crawling Websites of News Papers)
- Pytorch + Keras + Tensorflow (Deep learning libraries)
- scikit-learn (Machine Learning Library)
- NLTK

4 References

[1] Cucerzan, Silviu. "Large-scale named entity disambiguation based on Wikipedia data." *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).* 2007.

[2] Upadhyay, Shyam, Christos Christodoulopoulos, and Dan Roth. " "Making the News": Identifying Noteworthy Events in News Articles." *Proceedings of the Fourth Workshop on Events.* 2016.

[3] Choubey, Prafulla Kumar, Kaushik Raju, and Ruihong Huang. "Identifying the most dominant event in a news article by mining event coreference relations." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).* 2018.