Predicting Wikipedia Infobox Type Information using Word Embeddings on Categories

Russa Biswas[1,2], Maria Koutraki[1,2], and Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[2] Karlsruhe Institute of Technology, Institute AIFB, Germany
{firstname.lastname}@kit.edu

Abstract. Wikipedia has emerged as the largest multilingual, web based general reference work on the Internet. A huge amount of human resources have been invested in the creation and update of Wikipedia articles which are ideally complemented by so-called infobox templates defining the type of the underlying article. It has been observed that the Wikipedia infobox type information is often incomplete and inconsistent due to various reasons. However, the Wikipedia infobox type information plays a fundamental role for the RDF type information of Wikipedia based Knowledge Graphs such as DBpedia. This stimulates the need of always having the correct and complete infobox type information. In this work, we propose an approach to predict Wikipedia infobox types by using word embeddings on categories of Wikipedia articles, and analyze the impact of using minimal information from the Wikipedia articles in the prediction process.

Keywords: Wikipedia · Infobox · Word Embeddings · Text Classification

1 Introduction

Wikipedia has become the most widely used and largest multilingual open encyclopedia. Huge amount of human skills, expertise and efforts goes in for the creation of Wikipedia articles. It comprises of both structured and unstructured or free text. Structured data in Wikipedia is represented in the form of an *infobox* containing property value pairs summarizing the information content of the article. An infobox is a fixed-format table usually added to consistently present a summary of some unifying aspects that the articles share and sometimes to improve navigation to other interrelated articles. Furthermore, infobox information is widely used in different Knowledge Graphs (KGs) such as DBpedia.

Wikipedia infobox templates are created and assigned based on the categorical type of the article, i.e. articles belonging to a specific genre or type should be assigned the same template. The assignment of the infobox type to a Wikipedia article is executed based on the discussions between the contributors and the editors of the content of the Wikipedia article. However, no integrity tests are conducted to determine the correctness of the infobox assignment. This leads to

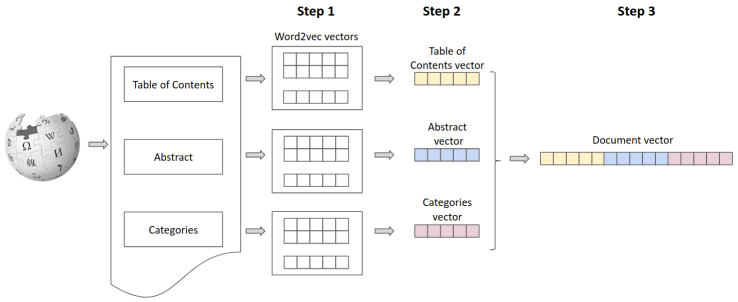

Fig. 1. Feature extraction from the Wikipedia articles (best viewed in color)

the assignment of incomplete and or incorrect infobox type information which eventually leads to erroneous RDF type information in the KGs [2].

The prediction of Wikipedia infobox type can be viewed upon as a text classification problem [7] in which the infobox types are nothing but the classes. In this paper, it is intended to classify the Wikipedia articles by exploiting word embeddings on the Wikipedia *categories* the articles belong to. Moreover, the impact of using minimal information such as only the first line of the abstract against the whole abstract in the classification process has also been studied. This work is inspired by the work done on Wikipedia infobox type prediction in [3] and extends the approach by making use of Wikipedia categories.

2 Related Work

Wu et al.[8] proposed KYLIN, a method of automatically creating new infoboxes and updating the existing incomplete ones, by learning a CRF extractor over common attributes. An automated Wikipedia infobox type prediction has been proposed by Sultana et al.[6] by training a SVM classifier over the first k-sentences of the articles as well as categories and named entities. Bhuniyan et al.[1] focuses on an automated NLP based infobox type prediction system.

The work presented in this paper is inspired by another work on Wikipedia infobox type prediction presented by Biswas et al.[3] in which the prediction problem is converted to a classification problem, where word and graph embeddings have been applied to generate the feature set for various classifiers. However, in this work, our aim is to predict the Wikipedia infobox types leveraging word embeddings on Wikipedia categories followed by a neural network based classification process. The proposed method does not focus on the creation of new infoboxes rather it helps to predict correct infobox types.

3 Infobox Type Prediction

In this work, the infobox type prediction problem is considered as a text classification problem, in which infobox types are regarded as classes.

Features. Three features from the Wikipedia articles are being used for the classification process.

- **Table of Contents (TOC)** is the collection of section headers and sub headers of the Wikipedia articles.
- **Abstract (A)** of the Wikipedia articles i.e. summary of the entire content.
- **Categories (C)** is a list of Wikipedia's main categorization system, intended to group together articles on similar subjects.

Feature Vector. Word2Vec [5] word embeddings are applied to generate the feature vectors. Word2Vec aims to learn the distributed representation for words reducing high dimensional word representations while keeping linguistic contexts of words.

In this paper, the Google pre-trained word vectors [3] of length 300 are used to generate word vectors for each word present in the TOC, abstract as well as for the categories. The Google pre-trained word2vec model includes word vectors for a vocabulary of three million words and phrases trained on roughly 100 billion words from a Google News dataset.

For each Wikipedia articlean abstract vector, a TOC vector and a category vector are generated by performing vector addition on all the word vectors of the abstract and normalized by the total number of words present in each of these features. Finally, a document vector is generated by concatenating these three vectors as shown in Figure 1.

Classification. Two classifiers have been trained to predict the Wikipedia infobox types. The aforementioned document vector is used as the feature vector in the classification method using a Random Forest(RF) Classifier. For a multi-label convolutional neural network(CNN), categories and TOC are considered as free text and sentence classification [4] method has been used where each Wikipedia article is considered as a sentence.

4 Results

The classifiers have been trained on the most popular 30 infobox types with 5000 articles for each type from the Wikipedia 2016 version. Features generated using TF-IDF have been used as a baseline.

The experiments established the fact that categories of Wikipedia articles play a vital role to determine the infobox type. With CNN, categories can predict the infobox types with a micro F_1-score of 96.8% which is 0.7% better than our previous results obtained when the prediction was based on the entire abstract and TOC combined as shown in Table 1. Furthermore, the word embeddings approach performs much better than the TF-IDF baseline, since word embeddings are able to capture semantic similarities. For instance, the Wikipedia article of the album *The Wall* by Pink Floyd is assigned to most of the categories containing the word *album*. Furthermore, using categories only results in higher scores than using the entire abstract. Moreover, the prediction

[3] `https://code.google.com/archive/p/word2vec/`

Feature Set	With Embedding			TF-IDF	
	RF(CV)	RF(Split)	CNN	RF(CV)	RF(Split)
TOC	65%	65.8%	76.5%	38%	32.3%
A(full)	86%	86.4%	95.1%	80%	80.4%
A(1stSent)	82.2%	82%	93.5%	70.4%	71%
C	88%	88.3%	**96.8%**	33%	34.4%
TOC + C	88.6%	89%	**97.6%**	81%	81.7%
A(full) + TOC	88%	88%	**96.1%**	83%	83.9%
A(1stSent) + TOC	82%	82.1%	**95%**	77.8%	78.2%
A(full) + C	88.6%	89%	**97.6%**	82%	82.4%
A(1stSent) + C	88.4%	89.3%	**98%**	80.8%	81%
A(full) + TOC + C	89%	89.7%	**98.3%**	84.6%	85.3%
A(1stSent) + TOC + C	86%	87.1%	**98.2%**	83.3%	84.2%

Table 1. Performance of classifiers using micro F1 score over the features

results are slightly better if categories are considered together with the first sentence of the abstract instead of considering the whole abstract, which means that less information is sufficient to infer Wikipedia infoboxes. Furthermore, in all the experiments word embeddings perform better than TF-IDF. However, rather similar results are obtained when TOC and categories are combined with the entire abstract as well as only with the first sentence of the abstract.

5 Conclusion

In this paper, the achieved results strengthen the fact that Wikipedia categories as well as minimal text plays a vital role in the prediction of infobox types. In future we intend to design a semi-supervised approach to correct existing Wikipedia infobox types and to predict infobox types for newly created articles.

References

1. Bhuiyan, H., Oh, K., Hong, M., Jo, G.: An Unsupervised Approach for Identifying the Infobox Template of Wikipedia Article. In: CSE (2015)
2. Biswas, R., Koutraki, M., Sack, H.: Exploiting Equivalence to Infer Type Subsumption in Linked Graphs. In: European Semantic Web Conference (2018)
3. Biswas, R., Türker, R., Moghaddam, F.B., Koutraki, M., Sack, H.: Wikipedia infobox type prediction using embeddings. In: DL4KGS@ ESWC (2018)
4. Kim, Y.: Convolutional Neural Networks for Sentence Classification. In: EMNLP (2014)
5. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. CoRR (2013)
6. Sultana, A., Hasan, Q.M., Biswas, A.K., Das, S., Rahman, H., Ding, C.H.Q., Li, C.: Infobox Suggestion for Wikipedia Entities. In: CIKM (2012)
7. Türker, R., Zhang, L., Koutraki, M., Sack, H.: "the less is more" for text classification,. In: SEMANTiCS 2018 (2018)
8. Wu, F., Weld, D.S.: Autonomously Semantifying Wikipedia. In: CIKM (2007)