"The Less Is More" for Text Classification

Rima Türker[1,2], Lei Zhang[1], 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}@fiz-karlsruhe.de
{firstname.lastname}@kit.edu

1 Introduction

Text Classification [2,5] is gaining more attention due to the availability of a huge number of text data, such as blog articles and news data. Traditional text classification methods [1] use all the words present in a given text to represent a document. However, the high number of words mentioned in documents can tremendously increase the complexity of the classification task and subsequently make it very costly. Moreover, long (natural language text) documents usually include a different variety of information related to the topic of a document. For example, encyclopedic articles such as the life of a scientist[3], contain besides topic related content also detailed biographical information. Often, in such articles after the first paragraph (or first a few sentences), words or entities appear, which are not related to the main topic (or category[4]) of the article. We assume that the most informative part of such articles is limited to a few starting sentences. In other words, instead of considering the complete document, only its beginning can be exploited to classify a document accurately.

In this study, we design a *Knowledge Based Text Classification* method, which is able to classify a document by using only a few starting sentences of the article. Since the length of the considered text is rather limited, ambiguous words might lead to inaccurate classification results. Therefore, instead of words, we consider entities to represent a document. In addition, entities and categories are embedded into a common vector space, which allows capturing the semantic similarity between them. Moreover, the similarity based approach does not require any labeled training data as a prerequisite. Instead, it relies on the semantic similarity between a set of predefined categories and a given document to determine which category the given document belongs to. The study has been validated with preliminary experiments on text classification for encyclopedic articles, which show that our method achieves comparable and even better results using only the first few sentences of a document than using the entire document.

2 Knowledge-Based Text Classification (KBTC)

Given a Knowledge Base KB containing a set of entities $E = \{e_1, e_2, .., e_n\}$ and a set of hierarchically related categories $C = \{c_1, c_2, .., c_m\}$, where each entity

[3] http://scihi.org/albert-einstein-revolutionized-physics/
[4] https://en.wikipedia.org/wiki/Category:Physics

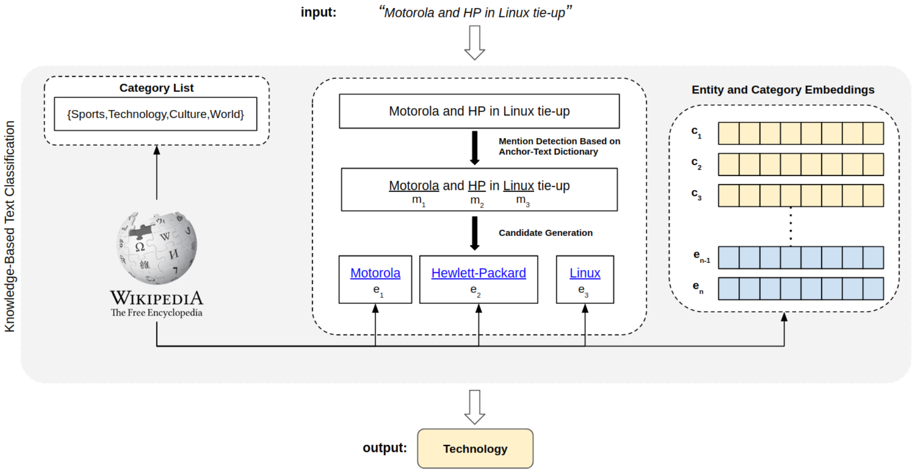
input:
“Motorola and HP in Linux tie-up”
Knowledge-Based Text Classification
Category List
{Sports,Technology,Culture,World}
Wikipedia
The Free Encyclopedia
Motorola and HP in Linux tie-up
Mention Detection Based on Anchor-Text Dictionary
Motorola and HP in Linux tie-up
m1 m2 m3
Candidate Generation
Motorola e1
Hewlett-Packard e2
Linux e3
Entity and Category Embeddings
c1
c2
c3
en-1
en
output:
Technology

Fig. 1. The work flow of the Knowledge Based Text Classification approach (best viewed in color)

$e_i \in E$ is associated with a set of categories $C' \subseteq C$. The input is a text t, which contains a set of mentions $M_t = \{m_1, \ldots, m_k\}$ that uniquely refer to a set of entities. Then, the output is the most relevant category $c_i \in C'$ for the given text t.

KBTC Overview. The general work flow of Knowledge Based Text Classification is shown in Figure 1. The first step is *“Mention Detection Based on Anchor-Text Dictionary”*, where each entity mention present in t is detected based on a *“Anchor-Text Dictionary”* prefabricated from Wikipedia. The Anchor-Text Dictionary contains all mentions and their corresponding Wikipedia entities. In order to construct an Anchor-Text Dictionary all the anchor texts of hyperlinks in Wikipedia articles referring to another Wikipedia article are extracted, whereby the anchor texts serve as mentions and the Wikipedia article links refer to the corresponding entities. In the second step, for each detected mention in the given input text candidate entities are generated based on the Anchor-Text Dictionary. In our example these are “Motorola”, “Hewlett-Packard” and “Linux”. Likewise, the predefined categories are mapped to Wikipedia categories. Finally, with the help of entity and category embeddings [3] that have been precomputed from Wikipedia, the output is the semantically most related category for the given entities. Thereby, in the given example the category *Technology* will be determined.

Probabilistic Model. The proposed classification task is formalized as estimating the probability of $P(c|t)$ of each predefined category c and an input text t. Based on Bayes' theorem, the probability $P(c|t)$ can be rewritten as follows:

$$P(c|t) = \frac{P(c,t)}{P(t)} \propto P(c,t) \tag{1}$$

where the denominator $P(t)$ has no impact on the ranking of the categories. For an input text t, a *mention* is a term in t that can refer to an entity e and the *context* of e is the set of all other mentions in t except the one for e. For each candidate entity e in t, the input text t can be decomposed into the mention and context of e, denoted by m_e and C_e, respectively. Based on the above introduced concepts, the joint probability $P(c,t)$ is given as follows:

$$\begin{aligned} P(c,t) &= \sum_{e\in E_t} P(e,c,t) = \sum_{e\in E_t} P(e,c,m_e,C_e) \\ &= \sum_{e\in E_t} P(e)P(c|e)P(m_e|e)P(C_e|e) \end{aligned} \tag{2}$$

where E_t represents the set of all possible entities contained in the input text t. Here, we simply apply a uniform distribution to calculate $P(e)$ for each entity e. The probability $P(c|e)$ models the relatedness between an entity e and a category c, which is estimated by using the prefabricated entity-category embeddings. Moreover, the probability of $P(m_e|e)$ is calculated based on the anchor text dictionary. Finally, the probability $P(C_e|e)$ models the relatedness between the entity e and its context C_e. Each mention in C_e refers to a context entity e_c from the given knowledge base. The probability $P(C_e|e)$ can be calculated with the help of entity-category embeddings. More details about the probability estimation can be found in [4].

3 Results and Discussion

Dataset. The proposed text classification approach is evaluated on articles of *SciHi*[5], a web blog on the history of science. From that dataset[6] 1452 articles associated to a single category have been considered. The different categories supported in the dataset are 45 and the average number of sentences per article is 32.96.

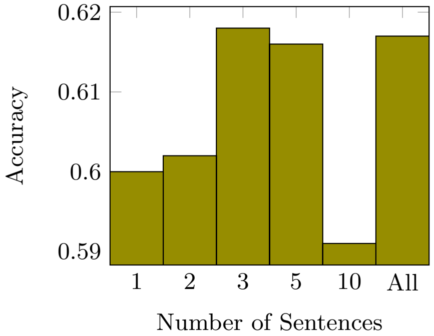

Fig. 2. Performance of the approach for different size of sampled sentences.

Number Of Sentences	Execution Time
1	18 min
2	23 min
3	28 min
5	38 min
10	64 min
All	215 min

Table 1. Execution Time for different size of sampled sentences.

[5] http://scihi.org/

[6] https://github.com/ISE-FIZKarlsruhe/Submission-under-review

Experimental Results. The proposed approach does not require any training phase. Therefore, only test sets are generated for the classification task from SciHi data. To show the impact of the number of the starting sentences of the articles on the classification accuracy, the data set has been sampled in different sizes. From each article, the first sentence, first 2, first 3, first 5, first 10 sentences and complete documents have been collected. For each sampled datasets the proposed approach has been applied to the classification task. The results are depicted in Fig. 2. The results show that a few starting sentences (in this case 3 sentences) are rather informative and have huge impact on the classification accuracy. During the experiments it has been observed that most of the times irrelevant entities to the corresponding category tend to appear after the first 2 or 3 sentences. Hence, after the 3^{rd} sentence the accuracy starts to drop (Fig. 2). Note that usually in such documents the frequency of relevant entities is higher in comparison to irrelevant entities. Therefore, complete documents help to obtain reasonable classification accuracy. However, the classification of complete documents is computationally very expensive (cp. Table 1). The classification of the whole documents takes 215 minutes while the classification of a sentence requires no more than 18 minutes for the entire dataset. The best results have been obtained with first 3 sentences (Fig. 2), where the execution time was 23 minutes, which is almost 90% faster. As expected, the complexity significantly increases when the number of sentences is increased.

4 Conclusion and Future Work

In this study, a probabilistic text classification approach has been used to analyze the influence of the text length for a text classification task. Based on the obtained results we can conclude that considering complete document does not always increase the classification accuracy. Instead, the accuracy depends on the nature of the considered part of the documents. In this study, it has been observed that the most informative part of encyclopedic documents is the first 3 sentences for the classification based on entity and category embeddings. Moreover, as anticipated, the complexity of the classification task decreases by considering only a few starting sentences. As for future work, we plan to apply the proposed approach to the different domains such as patent data to be able to classify patents.

References

1. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98 pp. 137–142 (1998)
2. Song, Y., Roth, D.: On dataless hierarchical text classification. In: AAAI (2014)
3. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. CoRR (2015)
4. Türker, R., Zhang, L., Koutraki, M., Sack, H.: Short text categorization using joint entity and category embeddings - (under review), https://github.com/ISE-FIZKarlsruhe/Submission-under-review
5. Zhang, X., Zhao, J.J., LeCun, Y.: Character-level convolutional networks for text classification. In: NIPS (2015)