Knowledge-Based Short Text Categorization Using Entity and Category Embedding

Presentation of work originally published in the Proc. of the 16th ESWC 2019

Rima Türker,[1] Lei Zhang,[1] Maria Koutraki,[2] Harald Sack[1]

Abstract: Short text categorization is an important task due to the rapid growth of online available short texts in various domains such as web search snippets, news feeds, etc. Most of the traditional methods suffer from sparsity and shortness of the text. Moreover, supervised learning methods require a significant amount of training data and manually labeling such data can be very time-consuming and costly. In this study, we propose a novel probabilistic model for Knowledge-Based Short Text Categorization (KBSTC), which does not require any labeled training data to categorize a short text [Tü].

Keywords: Short Text Categorization; Dataless Text Classification, Network Embeddings

1 Introduction

Short text categorization [Tü18b, Tü18a] plays a fundamental role in many Natural Language Processing applications such as web search, question answering, etc. Although, traditional text classification methods perform well on long text such as news articles, yet, by considering short text, most of them suffer from issues such as data sparsity and insufficient text length. Moreover, most text classification approaches require a significant amount of labeled training data and a sophisticated parameter tuning process. Manual labeling of such data can be a rather time-consuming and costly task. Especially, if the text to be labeled is of a specific scientific or technical domain, crowd-sourcing based labeling approaches do not work successfully and only expensive domain experts are able to fulfill the manual labeling task. Alternatively, semi-supervised text classification approaches have been proposed to reduce the labeling effort. Yet, due to the diversity of the documents in many applications, generating small training set for semi-supervised approaches still remains an expensive process.To address the lack of labeled data problem, we propose a novel probabilistic model for Knowledge-Based Short Text Categorization (KBSTC), which does not require any labeled training data. It is able to capture the semantic relations between the entities represented in a short text and the predefined categories by embedding both into a common vector space using the proposed network embedding technique. Finally, the appropriate category for the given text can be derived based on the semantic similarity between entities

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany {name.surname}@fiz-karlsruhe.de
[2] L3S Research Center, Leibniz University of Hannover, Hannover, Germany {surname}@l3s.de

doi:10.18420/inf2019_45

present in the given text and the set of predefined categories. The similarity is computed based on the vector representation of entities and categories.

2 Knowledge-Based Short Text Categorization (KBSTC)

Given an input short text t that contains a set of entities $E_t \subseteq E$ as well as a set of predefined categories $C' \subseteq C$ (from the underlying knowledge base KB), the output of the KBSTC task is the most relevant category $c_i \in C'$ for the given short text t, i.e., we compute the category function $f_{cat}(t) = c_i$, where $c_i \in C'$.

The proposed categorization 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. Moreover, we define a novel graph embedding that exploits (a) the entity-entity graph defined via Wikilinks and (b) the entity-category graph defined by the Wikipedia category system to implement KBSTC. More details about the probability estimation can be found in [Tü].

3 Experimental Results and Conclusion

To demonstrate the performance of the KBSTC approach, it has been compared against several text classification approaches. The experimental results have proven that by utilizing KBSTC it is possible to categorize short text in an unsupervised way with a high accuracy. Further, to assess the quality of the proposed entity and category embedding model, we have compared it with state-of-the-art embedding approaches in the context of the KBSTC task. The results indicate that our embedding model enables to capture better semantic relations between entities and categories from Wikipedia. Overall, all the experimental results have demonstrated that for short text categorization, KBSTC achieves a high accuracy without requiring any labeled data, a time-consuming training phase, or a cumbersome parameter tuning step.

References

[Tü] Türker, Rima; Zhang, Lei; Koutraki, Maria; Sack, Harald: Knowledge-Based Short Text Categorization Using Entity and Category Embedding. In: ESWC 2019.

[Tü18a] Türker, Rima; Zhang, Lei; Koutraki, Maria; Sack, Harald: TECNE: Knowledge Based Text Classification Using Network Embeddings. In: EKAW. 2018.

[Tü18b] Türker, Rima; Zhang, Lei; Koutraki, Maria; Sack, Harald: “the less is more” for text classification. In: SEMANTiCS. 2018.