Knowledge-Based Dataless Text Categorization

Rima Türker

FIZ Karlsruhe, Leibniz Institute for Information Infrastructure,
Eggenstein-Leopoldshafen, Germany &
AIFB, Karlsruhe Institute of Technology (KIT),
Karlsruhe, Germany
{rima.tuerker}@fiz-karlsruhe.de

Abstract. Text categorization is an important task due to the rapid growth of online available text data in various domains such as web search snippets, news documents, etc. Traditional supervised methods require a significant amount of training data and manually labeling such data can be very time-consuming and costly. Moreover, in case the text to be labeled is of a specific domain, then only the expensive domain experts are able to fulfill the manual labeling task. This thesis focuses on the problem of missing labeled data and aims to develop a novel and generic model which does not require any labeled training data to categorize text. Instead, it utilizes the semantic similarity between documents and the predefined categories by leveraging graph embedding techniques.

Keywords: Text Categorization, Dataless Classification, Network Embeddings

1 Introduction

Text categorization plays a fundamental role in many Natural Language Processing applications such as web search, question answering, etc. Traditional 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 [10,19] 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 [6].

To address the lack of labeled data problem, a number of *dataless text classification* methods [14,1] have been proposed. These methods do not require any labelled data to perform text classification. Rather, they rely on the semantic similarity between a given document and a set of predefined categories to determine which category the given document belongs to. In other words, documents and categories are represented in a common semantic space based

on the words contained in the documents and category labels, which allows to calculate a meaningful semantic similarity between documents and categories based on their vector representation. However, the most prominent and successful dataless classification approaches are designed for long documents such as news documents, i.e in case of short text most of them fail to classify the text properly as the available context is rather limited.

This thesis aims to address mentioned challenges by developing a **K**nowledge **B**ased **D**ataless **T**ext **C**lassification (KBDTC) approach for documents of arbitrary length without requiring any labeled training data. The method utilizes Knowledge Bases (KBs) as an external source. Moreover, to determine the category of a given text, KBDTC exploits the semantic similarity between the document and the predefined categories by leveraging graph embedding techniques.

The rest of this paper is structured as follows: Sec. 2 discusses related work. The research problems and expected contributions are presented in Sec. 3., while Sec. 4 outlines the research methodology and the approach. Sec. 5 and 6 describe the experimental setup for the evaluation as well as discuss the achieved results. Last, Sec. 7, concludes the paper with a discussion of open issues and future work.

2 State of the Art

The aim of this thesis is to develop a Dataless Text Categorizing method which does not require any labeled data for training, instead it utilizes KBs as an external knowledge. Thus, in this section several Dataless Text Classification methods are presented. Since the proposed method (see Sec. 4) has been already applied to short text, the studies related to short text classification are also discussed in the subsequent section.

Dataless Text Classification. In order to address the problem of missing labeled data, [1] introduced a dataless text classification method by representing documents and category labels in a common semantic space. As source, Wikipedia was utilized supported with Explicit Semantic Analysis (ESA)[3] to quantify semantic relatedness between the labels to be assigned and the documents. As a result, it was shown that ESA is able to achieve better classification results than the traditional BOW representations. Further, [14] proposed a dataless hierarchical text classification by dividing the dataless classification task into two steps. In the semantic similarity step, both labels and documents were represented in a common semantic space, which allows to calculate semantic relatedness between documents and labels. In the bootstrapping step, the approach made use of a machine learning based classification procedure with the aim of iteratively improving classification accuracy.

In contrast to these approaches, our goal differs in two main aspects. First, all the mentioned studies were designed for long text such as news articles. However the main purpose of this thesis is categorization of documents of arbitrary length without the necessity of labeled training data. Second, none of the mentioned approaches made use of the entities. Rather, to represent a document, they consider the words contained in the document.

Short Text Classification. To overcome the data sparsity problem of short text, recent works [17,18] proposed deep learning based approaches for short text classification. The results of these approaches have been compared with traditional supervised classification methods, such as SVM, multinomial logistic regression, etc., where the authors showed that in most of the cases their approach achieved superior results.While performing well in practice, the aforementioned approaches are slow both in the training and in the test phases. In addition, their performance highly depends on the size of training data, its distribution, and the chosen hyper parameters. By contrast, our approach does not require any training data nor any parameter tuning.

3 Problem Statement and Contributions

This section presents the research questions that are intended to be answered in this thesis as well as the related hypotheses.

3.1 Problem Statement

In this research we aim to address the following questions:

RQ 1. *How can entities that are associated with hierarchically related categories from a KB be utilized for short text classification without requiring any labeled data as a prerequisite?*

In this thesis, as a first step, we have considered short documents for the categorization task. In short text the available context is rather limited and it is assumed that words tend to be ambiguous, however entities carry much more information [16]. Therefore, we have developed a Knowledge Based **S**hort Text Categorization (KB**STC**) method, which considers the semantic similarity between entities (present in a given short text) and a set of predefined categories to derive the category of the text.

Then, the corresponding sub-question is:

RQ 2. *How to capture the semantic relation between entities and categories?*

To calculate the meaningful semantic relatedness, the proper semantic representation of entities and categories in a common vector space is essential. For this reason, we have proposed a new entity and category embedding model which can leverage entities and categories from large KBs, furthermore can embed them into a common vector space.

RQ 3. *Can words along with entities present in a text be utilized to increase the accuracy of KBDTC?*

The subsequent step is to investigate how to incorporate words into KBDTC to enhance the classification accuracy. KBDTC relies on the semantic similarity between documents and categories to determine the category of a given text. In order to exploit words as well as entities, the semantic similarity between words, entities and categories should be captured by a new embedding model. In other words, in this phase of the thesis, the purpose is to develop a joint word, entity and category embedding model, then integrate this model to KBDTC to improve the classification accuracy.

RQ 4. *Can KBDTC be exploited to create labeled data for supervised classification methods?*
Supervised classification methods, especially, deep learning approaches perform very well in short text classification [18,2,4,5]. However, they all require a significant amount of labeled training data. We will investigate how to utilize KBDTC to generate training sets for supervised classification approaches.

RQ 5. *How to generalize KBDTC to perform categorization of arbitrary documents?*
The proposed approach has been assessed in the context of short text classification. The experiments (see Sec. 6) show that KBSTC can categorize short text in an unsupervised way with a high accuracy. As a subsequent step, we plan to generalize the proposed approach for categorization of arbitrary length documents including tweets (short text), search snippets (short text), news data (long text), etc.

RQ 6. *How to generalize KBDTC to be compatible with arbitrary KBs?*
As yet, a general KB, Wikipedia has been utilized for KBSTC task. However, domain specific knowledge is not appropriately covered by general KBs such as Wikipedia. Therefore, in this phase, the goal is to generalize KBDTC to be adaptable to arbitrary KBs from different domains.
The following hypotheses are deducted:

H 1. *"Exploiting a KB can help to conduct text classification without requiring any labelled data"*

H 2. *"Embedding entities and categories into a common vector space enables to calculate meaningful semantic relatedness between them"*

H 3. *"KBDTC can be used to generate training sets for supervised approaches"*

H 4. *"KBDTC can be extended to utilize arbitrary KBs to categorize documents"*

3.2 Contributions

The expected contributions of this research include:

1. A new paradigm for text categorization, based on a KB.
2. A new model for short text categorization so called KBSTC.
3. The development of an entity and category embedding model for calculating the semantic similarity between entities and categories.
4. The improvement of the embedding model by including words into the entity and category embeddings.
5. The development of a generic KBDTC approach, which is compatible with any KBs for categorizing arbitrary text.

4 Research Methodology and Approach

In this research, so far we have developed KBSTC **(RQ 1)** as well as a new entity and category embedding model **(RQ 2)**. Therefore, this section provides a

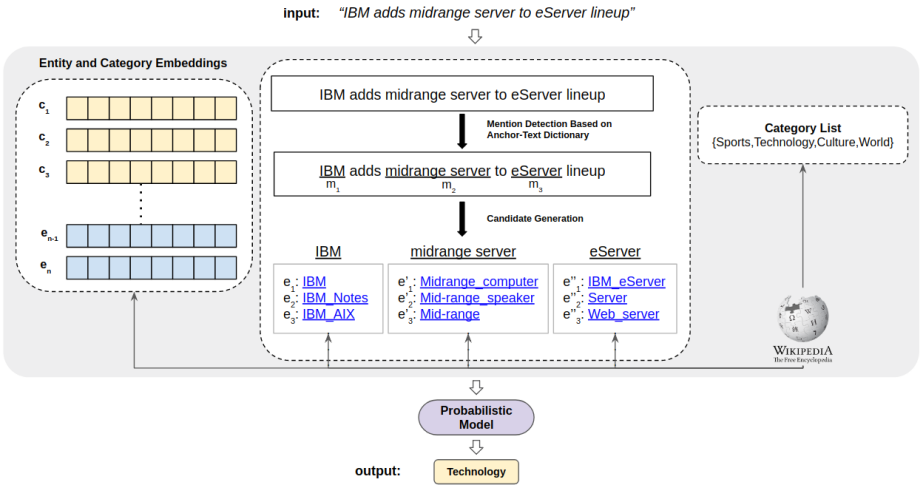

Fig. 1. The work flow of the proposed KBSTC approach (best viewed in color)

formal definition of the KBSTC task, followed by the description of the proposed probabilistic approach for KBSTC. Finally, the plan for tackling the rest of the research questions **(RQ 3, RQ 4, RQ 5, RQ 6)** is presented.

– **RQ 1 and RQ 2**
Definition (KBSTC task). 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 $c_i \in C'$.
KBSTC Overview. The general workflow of KBSTC is shown in Fig. 1. In the first step, each entity mention present in a given short text t is detected. Next, for each mention, a set of candidate entities are generated based on a pre-fabricated Anchor-Text Dictionary, which contains all mentions and their corresponding Wikipedia entities. In order to detect entity mentions, first all n-grams from the input text are gathered and then the extracted n-grams matching surface forms of entities (based on the Anchor-Text dictionary) are selected as entity mentions. To construct the Anchor-Text Dictionary, all the anchor texts of hyperlinks in Wikipedia pointing to any Wikipedia articles are extracted, whereby the anchor texts serve as mentions and the links refer to the corresponding entities. Given the short text t as *"IBM adds midrange server to eServer lineup"*, the detected mentions are *"IBM"*, *"midrange server"* and *"eServer"*. Likewise the predefined categories, $C' = \{Sports, Technology, Culture, World\}$, are mapped to Wikipedia categories. Finally, applying the proposed probabilistic model by utilizing the entity and category embeddings that have been precomputed from Wikipedia, the output of the KBSTC task is the semantically most relevant category for the entities present in t. Thereby, in the given example the category *Technology* should be determined.

Probabilistic Approach
The KBSTC task is formalized as estimating the probability of $P(c|t)$ of each predefined category c and an input short text t. The result of this probability estimation can be considered as a score for each category. Therefore, the most relevant category c for a given text t should maximize the probability $P(c|t)$. 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 $P(t)$ can be ignored as it has no impact on the ranking of the categories. To calculate $P(c|t)$, the proper semantic representation of entities and categories in a common vector space is essential. Hence, in this thesis we have also proposed an entity and category embedding model that embeds entities and categories from Wikipedia into a common vector space. Readers can refer to our research paper [15] for more technical detail of the parameter estimation of Eq. (1) and proposed embedding model.

- **RQ 3**
  The next research step is to extend KBSTC towards the additional inclusion of word embeddings into the common entity and category vector space. KBSTC exploits only entities for short text classification, however, words might have a positive impact on the model performance. Since the proposed embedding model is flexible to adopt new type of relations, the inclusion of word embedding is straightforward. In other words, words and word-category relations can be included as an additional type of vertices and edges in already constructed heterogeneous network [15].
- **RQ 4**
  As already mentioned supervised methods, especially, deep learning approaches perform very well in short text classification. However, they require million-scale labelled documents [8]. Since KBSTC can classify a given short text without requiring any labelled data, the most confidently classified documents can be collected as a training set for the deep learning phase. Next, the rest of the documents (that could not be properly classified by KBSTC) can be classified with the trained deep learning model.
- **RQ 5 and RQ 6**
  Final goal of this thesis is to have a full fledged generic KBDTC approach. The ultimate approach should be capable to categorize a given text (which are of arbitrary length) without requiring any training data and compatible with any KBs.

5 Evaluation Plan

As already mentioned, so far we have developed KBSTC and a new entity and category embedding model. Hence, this section provides a description of the datasets and the baselines for evaluating KBSTC and the embedding model.

5.1 Datasets

The experiments have been conducted on the following two benchmarks (the data distribution of both dataset can be found in [15]):
AG News (AG)[1]***:*** This dataset is adopted from [20], which contains both titles and short descriptions (usually one sentence) of news articles. In our experiments, the dataset has two versions, where one contains only titles and the other contains both titles and descriptions. The total number of entities and the average number of entities and words per text in the test datasets are shown in Table 1.
Google Snippets (Snippets)[2]***:*** This is a well-known dataset for short text classification, which was introduced in [12] and contains short snippets from Google search results. As shown in Table 1, the test dataset has in total 20,284 entities, an average of 8.9 entities and an average of 17.97 words in each snippet.

Table 1. Statistical analysis of the test datasets

Dataset	#Entities	Avg. #Ent	Avg. #Word
AG News (Title)	24,416	3.21	7.14
AG News (Title+Description)	89,933	11.83	38.65
Google Snippets	20,284	8.90	17.97

As the focus of this work is the KBSTC task, where the goal is to derive the most relevant category from the knowledge base for a given short text, we need to adapt these datasets by aligning the labels/categories with the categories in the used knowledge base. More specifically, each label/category in these datasets is manually mapped to its corresponding Wikipedia category, e.g., the category *Sports* from the AG dataset is mapped to the Wikipedia category *Sports*[3]. Furthermore, as KBSTC does not depend on any training/labeled data, the training datasets of AG and Snippets are only used for the training of the supervised baseline methods. Lastly, to measure the performance of KBSTC, the classification accuracy (the ratio of correctly classified data over all the test data) was used.

5.2 Baselines

To demonstrate the performance of the KBSTC approach, the following dataless and supervised classification methods have been selected as baselines:
Dataless ESA and Dataless Word2Vec: As described in Sec. 2, the dataless approaches do not require any labeled data or training phase, therefore, they can be considered as the most similar approaches to KBSTC. Two variants of the state-of-the-art dataless approach [14] are considered as baselines, which are based on ESA [3] and Word2Vec [9], respectively.

[1] http://goo.gl/JyCnZq [2] http://jwebpro.sourceforge.net/data-web-snippets.tar.gz
[3] https://en.wikipedia.org/wiki/Category:Sports

Table 2. The classification accuracy of KBSTC against baselines (%)

Model	AG (title)	AG (title+description)	Snippets
Dataless ESA [14]	53.5	64.1	48.5
Dataless Word2Vec [14]	49.5	52.7	52.4
NB+TF-IDF	86.6	90.2	64.4
SVM+TF-IDF	**87.6**	**91.9**	69.1
LR+TF-IDF	87.1	91.7	63.6
KBSTC+Our Embedding	67.9	80.5	**72.0**

NB, SVM, LR: Additional baselines include the traditional supervised classifiers, i.e., Naive Bayes (NB), Support Vector Machine (SVM) and Logistic Regression (LR), with the features calculated based on the term frequency and the inverse document frequency (TF-IDF).

6 Intermediate Results

This section provides experimental results as well as a comparison to existing state-of-the-art approaches.

6.1 Evaluation of KBSTC

Table 2 shows the accuracy of the proposed probabilistic KBSTC approach based on our entity and category embedding model in comparison to the baselines on the AG and Snippets datasets.

It is observed that the KBSTC approach considerably outperforms the dataless classification approaches. While Dataless ESA and Dataless Word2Vec have been assessed with longer news articles and achieved promising results in [14], they cannot perform well with short text due to the data sparsity problem.

Remarkably, KBSTC performs better than all the baselines on the Snippets dataset, however, all supervised approaches outperform KBSTC on the AG dataset. The reason here can be attributed to the different characteristics of the two datasets. AG is a larger dataset with more training samples in comparison to Snippets. Moreover, the AG dataset provides only 4 different categories in comparison to 8 categories of the Snippets dataset. Those differences might be the reason of the significant decrease in accuracy for the supervised approaches on the Snippets dataset in comparison to the AG dataset. This could be an indicator that the size of the training data and the number of classes make a real impact on the classification accuracy for the supervised approaches. Since KBSTC does not require or use any labeled data, the number of the available training samples has no impact on its accuracy.

Regarding the results of KBSTC, the AG (title+description) dataset yields better accuracy than the Snippets dataset, which in turn, results in better accuracy than the AG (title) dataset. The reason might be found in the nature of the datasets. As shown in Table 1, the average number of entities per text in AG (title+description) is greater than Snippets, followed by AG (title). Often a richer context with more entities can make the categorization more accurate.

Table 3. The classification accuracy of KBSTC with different embedding models (%)

Model	AG (title)	AG (title+description)	Snippets
KBSTC+HCE	67.0	79.6	**72.3**
KBSTC+DeepWalk	57.1	74.2	64.3
KBSTC+RDF2Vec	62.7	77.5	68.2
KBSTC+Our Embedding	**67.9**	**80.5**	72.0

Overall, the results in Table 2 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.

6.2 Evaluation of Entity and Category Embedding

To assess the quality of the proposed entity and category embedding model, we compared it with HCE [7], DeepWalk [11] and RDF2Vec [13] in the context of the KBSTC task.

While the Wikipedia entity and category embeddings generated by HCE can be directly used, DeepWalk has been applied on the network constructed using Wikipedia and RDF2Vec has been applied on the RDF graph of DBpedia to obtain the needed embeddings. Then, these embeddings are integrated into KBSTC to compute the entity-category relatedness. The results of KBSTC with different embedding models are shown in Table 3. The proposed entity and category embedding model outperforms all other embedding models for the KBSTC task on the AG dataset, while HCE performs slightly better than our model on the Snippets dataset.

As HCE is a more specific embedding model that has been designed to learn the representation of entities and their associated categories from Wikipedia, it is not flexible to be adapted to other networks. In contrast, our model can deal with more general networks.

Although DeepWalk and RDF2Vec aim to learn the representation of vertices in general networks and RDF graphs, respectively, they have been either designed for homogeneous networks or treated each type of vertices and edges in a RDF graph equally. The results also indicate that our embedding model enables to capture better semantic representation of vertices by taking into account different types of networks.

7 Conclusions and Lessons learned

In this thesis, the main goal is to address the labeled data scarcity problem. For this purpose we have proposed a new paradigm for text categorization based on KBs so called KBDTC. Furthermore, a novel KBSTC model which is originated from the proposed paradigm has been presented. The experimental results have proven that it is possible to categorize short text in an unsupervised way with a high accuracy by utilizing a KB. As for future work, we would like to tackle **RQ 3**, **RQ 4**, **RQ 5** and **RQ 6**. In other words, the next research steps includes the

extension of KBDTC towards the additional inclusion of words and arbitrary document categorization. Moreover, we intend to utilize KBDTC to generate training set for other supervised classifiers such as deep learning methods.

Acknowledgement. This thesis is supervised by Prof. Harald Sack and Dr. Lei Zhang.

References

1. Chang, M.W., Ratinov, L.A., Roth, D., Srikumar, V.: Importance of semantic representation: Dataless classification. In: AAAI (2008)
2. Conneau, A., Schwenk, H., Barrault, L., LeCun, Y.: Very deep convolutional networks for natural language processing. CoRR (2016)
3. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI (2007)
4. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP (2014)
5. Lee, J.Y., Dernoncourt, F.: Sequential short-text classification with recurrent and convolutional neural networks. In: CoRR (2016)
6. Li, C., Xing, J., Sun, A., Ma, Z.: Effective document labeling with very few seed words: A topic model approach. In: CIKM (2016)
7. Li, Y., Zheng, R., Tian, T., Hu, Z., Iyer, R., Sycara, K.P.: Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. In: COLING (2016)
8. Meng, Y., Shen, J., Zhang, C., Han, J.: Weakly-supervised neural text classification. In: ACM (2018)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
10. Nigam, K., McCallum, A., Thrun, S., Mitchell, T.M.: Text classification from labeled and unlabeled documents using EM. Machine Learning (2000)
11. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: KDD (2014)
12. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: WWW (2008)
13. Ristoski, P., Paulheim, H.: RDF2Vec: RDF Graph Embeddings for Data Mining. In: ISWC (2016)
14. Song, Y., Roth, D.: On dataless hierarchical text classification. In: AAAI (2014)
15. Türker, R., Zhang, L., Koutraki, M., Sack, H.: Knowledge-based short text categorization using entity and category embedding. In: ESWC (2019)
16. Wang, C., Song, Y., Li, H., Zhang, M., Han, J.: Text classification with heterogeneous information network kernels. In: AAAI (2016)
17. Wang, J., Wang, Z., Zhang, D., Yan, J.: Combining knowledge with deep convolutional neural networks for short text classification. In: IJCAI (2017)
18. Wang, P., Xu, B., Xu, J., Tian, G., Liu, C.L., Hao, H.: Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing (2016)
19. Xuan, J., Jiang, H., Ren, Z., Yan, J., Luo, Z.: Automatic bug triage using semi-supervised text classification. In: SEKE (2010)
20. Zhang, X., LeCun, Y.: Text understanding from scratch. CoRR (2015)