Improving Zero-Shot Text Classification with Graph-based Knowledge Representations

Fabian Hoppe[1,2]

[1]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[2]Karlsruhe Institute of Technology, Institute AIFB, Germany

Abstract

Insufficient training data is a key challenge for text classification. In particular, long-tail class distributions and emerging, new classes do not provide any training data for specific classes. Therefore, such a zero-shot setting must incorporate additional, external knowledge to enable transfer learning by connecting the external knowledge of previously unseen classes to texts. Recent zero-shot text classifier utilize only distributional semantics defined by large language models and based on class names or natural language descriptions. This implicit knowledge contains ambiguities, is not able to capture logical relations nor is it an efficient representation of factual knowledge. These drawbacks can be avoided by introducing explicit, external knowledge. Especially, knowledge graphs provide such explicit, unambiguous, and complementary, domain specific knowledge. Hence, this thesis explores graph-based knowledge as additional modality for zero-shot text classification. Besides a general investigation of this modality, the influence on the capabilities of dealing with domain shifts by including domain-specific knowledge is explored.

Keywords

Zero-Shot Learning, Text Classification, Knowledge Graph

1. Introduction

A crucial element for the current success of Deep Learning in Natural Language Processing (NLP) is the capability of models to be used in a transfer learning regime. Instead of training large models with millions to billions of parameters from scratch, they are pre-trained on similar tasks with large amounts of labeled data and fine-tuned by learning a small subset of parameters leveraging a smaller task-specific dataset. Especially language modeling, i.e. learning from natural language texts as pre-training task, has improved on the state of the art in many NLP tasks [1].

However, traditional transfer learning still requires task-specific data to adapt for the distribution shift by learning the small subset of parameters, making it infeasible for few- or zero-shot settings. Unfortunately, these settings are particularly common in many text classification tasks due to long-tail class distributions, emerging classes and generally high labelling costs. In these transfer learning scenarios, it is necessary to use additional, external knowledge about the task

Doctoral Consortium at ISWC 2022 co-located with 21st International Semantic Web Conference (ISWC 2022)
fabian.hoppe@fiz-karlsruhe.de (F. Hoppe)
0000-0002-7047-2770 (F. Hoppe)

to enable out of distribution generalization. For text classification, this kind of task agnostic transfer learning uses class definitions and is addressed by Zero-Shot Text Classification (ZSTC).

Instead of classifying documents by comparing them to other documents of a specific class, ZSTC exploits the explicitly defined semantics of classes to determine if a document belongs to a class, mimicking how humans are able to classify objects without considering examples. Naturally, the class definitions, that convey the semantics of a class, are a crucial part of ZSTC. They must provide sufficient context to enable knowledge transfer from pre-training tasks to downstream classification tasks.

Recently, Large Language Models (LLMs) utilizing natural language descriptions of tasks excel in the zero-shot setting by exploiting the implicit external knowledge of the LLM weights [2, 3]. However, providing sufficient context with natural language faces several challenges. Understanding the exact specifications of a class based on informal, natural language definitions is difficult even for humans. First, these informal class definitions are ambiguous, e.g. using *spam* as definition for annoying emails could lead a classifier to search for emails about canned meat. Secondly, the notion of distributional similarity reflected in language models differs from semantic similarity, e.g. logical relations like negations remain a challenge for language models. Finally, the classes use knowledge implicitly encoded in LLMs, which are notorious expensive to train and might include problematic input data. Consequently, new factual knowledge cannot be introduced easily and the classification results might even be based on counterfactual knowledge due to biased input data.

These challenges can be addressed by graph-based knowledge representations as an additional modality. It provides explicit knowledge to resolve ambiguity, with consideration of semantic similarity and verifiable factual knowledge, which can be updated more easily. Additionally, due to the Semantic Web, large Knowledge Graphs (KGs) providing external knowledge are available. Hence, this additional modality would add complementary, explicit, external knowledge. However, the usage of another modality is not a straightforward process and requires detailed investigations. This thesis contributes to that line of research by focusing on graph-based knowledge representations in ZSTC.

1.1. Importance

The ramifications of reducing the amount of required training data are easily visible by considering the impact of transfer learning in the past. It became easier to train a classifier, because it didn't require that much training data anymore, which allowed using text classification in more scenarios and significantly improved the results. Enhancing ZSTC will continue this trend within text classification. Additionally, the possibility of classifying classes with zero training examples would provide a solution for unbalanced classification problems.

Given the close connection between classification and understanding, an improvement of text classification would likely propagate beyond it to many other NLU tasks. Furthermore, utilizing graph-based knowledge representations in combination with implicit knowledge gathered by language models would be a practical step towards the combination of symbolic and subsymbolic systems.

2. Related Work

Initially, the zero-shot setting for text classification was explored by the dataless classification framework [4]. The document as well as the classes representations use Explicit Semantic Analysis (ESA) as a latent representation. The semantics of a class is based on the class name as external knowledge resource. This classifier performs strict zero-shot learning, which refers to not using any task specific training data. It is achieved by comparing classes and documents in the same latent space by the L^2 norm, making alignment unnecessary. This work got extended using cosine similarity and several neural network based word embeddings, like Word2Vec [5]. Due to the simplicity of this approach, it is already frequently applied in scenarios without training data like categorizing scholarly articles [6] and German archival documents [7].

Recently, more contextualized information are utilized to improve zero-shot learning. One approach reformulates text classification as textual entailment problem by generating a hypothesis for each class, like "This text is about sports" [8]. It enables further pre-training with generic NLI datasets and makes task-specific fine-tuning possible. It is notable that the additional textual context does not add background knowledge, but supports the alignment between document and class representations. As shown in [3] such an binarization approach generalizes even to a heterogeneous set of text classification tasks, e.g. topic detection, sentiment analysis. A similar approach interprets classification as cloze task using a masked pattern, like "[Category: ___]". Based on this, a language model can predict the class based on most-likely class names. The zero- and few-shot setting in LLMs is explored even beyond text classification. GPT-3 shows a remarkable performance on several NLP task by only providing a brief task description.

These approaches continue to represent classes by only using the class name or small descriptions. However, as shown by [9] for image classification, considering Wikipedia articles as additional external knowledge to generate GloVe based class representations in combination with a linear transformation layer to align image and class representations reaches state-of-the-art performance. In the image domain more external knowledge is already included for zero-shot classification. In [10] the authors particularly highlight that KGs are an important resource to address the issue of semantic insufficiency and provide datasets connecting the classes used in zero-shot image classification to KGs, which includes hierarchical information and logical expression such as disjointness.

In ZSTC models include only limited amounts of explicit knowledge sources. Early on [5] included hierarchical knowledge by performing top-down or bottom-up classification. In recent studies, the hierarchy was exploited by training traditional supervised classifiers for course-grained classes [11]. Initial steps towards integrated KGs as explicit external knowledge are made by [11], too. They use the ConceptNet KG to extract explicit relations between words. In summary, current ZSTC lacks methods to include explicit knowledge sources.

3. Research Questions

The explicit, external knowledge provided by graph-based knowledge representations as additional modality for zero-shot classification poses several questions awaiting further investigations. This thesis limits itself to exploring the basic concept and the impact on out of domain

generalization for these models.

RQ 1 *Can graph-based knowledge representations improve ZSTC?*
ZSTC can be divided into three main components which can be adapted to consider graph-based knowledge representations.

RQ 1.1 *How can graph-based knowledge extend class representations?*
Many approaches encoding KGs to capture different semantic aspects of entities exist. Identifying which of these models address the aforementioned challenges of class representations sufficiently is a key aspect of achieving improvements.

RQ 1.2 *How can graph-based knowledge extend document representations?*
Similar to class representations graph-based knowledge can be used to enhance document representations by including complementary knowledge from KGs. However, the focus for documents is on the additional factual knowledge provided by KGs. Therefore, the challenge of biased or out-dated language models is addressed with this research questions.

RQ 1.3 *How can class and document representations be aligned efficiently?*
ZSTC depends on relating document representations with class representations. Consequently, in order to propose a model utilizing graph-based knowledge, an efficient way to enable relating both representations needs to be investigated.

RQ 2 *What is the impact of utilizing explicit, external knowledge in ZSTC on the capabilities of dealing with domain shift?*
The out of domain generalization to unseen classes is limited. However, due to explicit domain knowledge provided by (domain-specific) KGs, this capability could improve compared to classifiers which are not utilizing additional domain knowledge.

The effort of answering this set of questions yields the following main contributions for this thesis.

- A novel ZSTC model utilizing explicit, external knowledge from KGs as part of the class and document representations.
- A detailed evaluation of the influence of domain shifts on the performance of zero-shot classification.

4. Preliminary Results

Already conducted investigations of the defined research questions are focused on providing a starting point to answer RQ 1.1 and RQ 1.3. The following sections summarize this work.

4.1. Intrinsic Evaluation of Class Representations

Extending class representations by utilizing graph-based knowledge needs to add relevant semantic relations among classes in order to be useful for ZSTC. However, extrinsic evaluation

limits the understanding of the semantic relatedness, requires computationally expensive training and evaluates all components of the model. Therefore, we proposed an intrinsic evaluation enabling an independent investigation of semantic relatedness for class representations [12].

The evaluation task predicts given a triple of classes `<Anchor, A, B>` if class `A` or `B` is more similar to the `Anchor` class. This limits the need for manually annotated labels to a binary classification, instead of directly estimating the continuous similarity between two classes. An embedding space Θ can be evaluated by checking if the constraint $cosine\ similarity(\Theta(Anchor), \Theta(A)) > cosine\ similarity(\Theta(Anchor), \Theta(B))$ matches the human labels. The initial evaluation of class representations is based on a dataset of 31 *arXiv.org* computer science classes. Each of the $3,000$ triples of this dataset was annotated by 5 domain experts. The results are presented in Table 1 comparing textual representations (Word2Vec [13], BERT [14]), KG embeddings (TransR [15], TransE [16], RDF2Vec [17]) and Wikipedia2Vec [18] as hybrid model. As external resources class names, the Wikipedia abstracts and manually mapped KG entities (DBpedia and AI-KG as domain specific KG) are considered.

Table 1
Intrinsic evaluation of class representations from arXiv.org classes.

Model	External Knowledge Source	Precision	Recall	F-measure
Word2Vec	Name	0.668	**0.757**	0.709
Word2Vec	Wikipedia abstract	0.682	0.65	0.666
BERT	Name	0.591	0.593	0.592
BERT	Wikipedia abstract	0.658	0.727	0.691
Wikipedia2Vec	Wikipedia entity	**0.738**	0.74	**0.739**
TransR	DBpedia	0.548	0.573	0.56
TransR	AI-KG	0.498	0.55	0.523
TransE	DBpedia	0.508	0.513	0.511
TransE	AI-KG	0.501	0.597	0.545
RDF2Vec	DBpedia	0.496	0.573	0.532
Human Annotator		0.947	0.853	0.881

Overall, Wikipedia2Vec achieving the best results illustrates the potential value of graph-based knowledge representations to capture semantic relatedness beyond text. However, the text modality provides in general better results than KGs. This indicates that state-of-the-art KG embeddings of single entities are unlikely to extend the class representations for ZSTC in a meaningful manner without further adjustments. Consequently, exploiting the potential benefits of graph-based knowledge needs to accumulate more external knowledge by combining several entities or even by considering TBox knowledge.

4.2. Alignment of Representations by a Fully-Connected Layer

The intrinsic evaluation is only an approximation for the potential performance of a classifier utilizing KG embeddings. As addressed by RQ 1.3, the relations between document and class representations enable zero-shot classification. Consequently, a fundamental component of

ZSTC is the alignment of both spaces to enable relating documents and classes. In principal, this alignment step estimates the function $\gamma : (d, c) \rightarrow \{0, 1\}$, where $d \in D$ is a document and $c \in C$ is a class. This basic process was investigated by utilizing a minimal model to generate initial results and establish a ZSTC pipeline [19]. This model consists of two BiLSTMs, which generate normalized, textual representations for documents as well as classes. Additionally, single KG entities are represented by RDF2Vec embeddings of DBpedia entities. The alignment step uses one fully-connected layer to train an implicit alignment and the scoring function between both representations.

For the experiments the multi-class, multi-label *arXiv.org* dataset is used to train the BiLSTMs as well as the dense layer. Preliminary results of this evaluation are restricted to computer science classes which are already seen during training and are reported in Table 2.

Table 2
Evaluation of the minimal zero-shot model on seen computer science classes of the *arXiv.org* dataset.

Model	Hamming Loss	Precision		Recall		F-measure	
		micro	macro	micro	macro	micro	macro
Random Classifier	0.499	0.061	0.061	0.502	0.380	0.109	0.102
NLI ZSL	0.171	0.157	**0.293**	0.416	0.408	0.229	**0.250**
BiLSTM (label)	0.52	0.066	0.067	**0.578**	**0.438**	0.119	0.112
BiLSTM (KG)	0.105	0.253	0.234	0.373	0.292	0.301	0.236
BiLSTM (label + KG)	**0.096**	**0.275**	0.248	0.348	0.272	**0.307**	0.237

Again the combination of textual and KG embeddings achieves the best results of all BiLSTM settings. However, the baseline natural language inference approach proposed in [8] achieves a good performance without any task specific training, especially considering the macro-averaged scores. This indicates that the trivial architecture of this zero-shot classifier depends on the amount of available training data for each class.

5. Evaluation

Measuring the improvement of ZSTC to answer RQ 1 uses the well-established evaluation metrics of classification tasks: precision, recall and f-measure. Additionally, as measurement for the accuracy in the multi-label multi-class setting, the hamming loss is used. However, the identification of useful benchmark datasets is more challenging in the zero-shot setting. Typical supervised text classification datasets assume independent and identical distribution (iid) for training and test set. This assumption does not hold for zero-shot learning. In this setting in particular the generalization beyond the training distribution should be evaluated. One proposed benchmark is generated by introducing a new train-test split for existing supervised text classification datasets, like Yahoo! Answers [8]. Naturally, this thesis will use the proposed Yahoo! Answers dataset and apply a similar approach to generate additional datasets for the evaluation of graph-based knowledge enhanced zero-shot classifier. Especially the investigation of domain shift to provide detailed answers about RQ 2 requires fine-grained classes with different degrees of domain shift. Therefore, the DBpedia dataset with its hierarchical structure should be utilized.

6. Conclusion

The preliminary results show that adding KGs as additional modality for ZSTC requires more complex models. Especially the intrinsic evaluation demonstrates that a one-to-one mapping between KG entities and classes does not provide sufficient external knowledge. Consequently, one important part of the future work is to improve semantic relatedness by considering multiple entities or in a next step use knowledge from TBoxes for KG based class representations. Given the performance of a fully connected layer to align class and document representations, RQ 1.3 needs further investigations as well. The preliminary results show that the training of such an alignment component is able to achieve better than random results, but lacks generalizability. More complex models with a stronger inductive bias must be considered in this respect. This could be addressed with explicit alignment of class and document representation space combined with a similarity based binary classification. Extending document representations to include KGs (RQ 1.2) is another important future step of this work in order to provide an extensive answer to RQ 1.

Naturally, the answer of RQ 1 provides the basis for a closer investigation of the possible domain shift and whether the explicit knowledge provided by KGs improves this (RQ 2) and thereby makes a potential contribution towards better transfer learning in ZSTC.

Acknowledgments

This thesis is supervised by Prof. Dr. Harald Sack.

References

[1] S. Ruder, M. E. Peters, S. Swayamdipta, T. Wolf, Transfer learning in natural language processing, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials, 2019, pp. 15–18.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[3] K. Halder, A. Akbik, J. Krapac, R. Vollgraf, Task-aware representation of sentences for generic text classification, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 3202–3213.

[4] M. W. Chang, L. Ratinov, D. Roth, V. Srikumar, Importance of semantic representation: Dataless classification, 27th AAAI conference on Artificial Intelligence (2008).

[5] Y. Song, D. Roth, On dataless hierarchical text classification, in: Proceedings of the 28th AAAI conference on Artificial Intelligence, 2014, p. 1579–1585.

[6] A. A. Salatino, F. Osborne, T. Thanapalasingam, E. Motta, The CSO classifier: Ontology-driven detection of research topics in scholarly articles, in: International Conference on Theory and Practice of Digital Libraries, 2019, pp. 296–311.

[7] F. Hoppe, T. Tietz, D. Dessì, N. Meyer, M. Sprau, M. Alam, H. Sack, The challenges of German archival document categorization on insufficient labeled data, in: Proceedings of

the Third Workshop on Humanities in the Semantic Web, co-located with 15th Extended Semantic Web Conference, 2020, pp. 15–20.
[8] W. Yin, J. Hay, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 3914–3923.
[9] S. Bujwid, J. Sullivan, Large-scale zero-shot image classification from rich and diverse textual descriptions, in: Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), 2021, pp. 38–52.
[10] Y. Geng, J. Chen, Z. Chen, J. Z. Pan, Z. Yuan, H. Chen, K-zsl: resources for knowledge-driven zero-shot learning, arXiv preprint arXiv:2106.15047 (2021).
[11] J. Zhang, P. Lertvittayakumjorn, Y. Guo, Integrating semantic knowledge to tackle zero-shot text classification, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 1031–1040.
[12] F. Hoppe, D. Dessì, H. Sack, Understanding class representations: An intrinsic evaluation of zero-shot text classification, in: Workshop on Deep Learning for Knowledge Graphs (DL4KG@ ISWC2021), CEUR WS, 2021, pp. 55–65. URL: http://ceur-ws.org/Vol-3034/paper8.pdf, event-place: Virtual Conference.
[13] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv (2013).
[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv (2018).
[15] Y. Lin, Z. Liu, M. Sun, Y. Liu, X. Zhu, Learning entity and relation embeddings for Knowledge Graph completion, in: 29th AAAI conference on Artificial Intelligence, 2015, pp. 2181–2187.
[16] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, Advances in neural information processing systems (2013).
[17] P. Ristoski, H. Paulheim, Rdf2Vec: RDF graph embeddings for data mining, in: International Semantic Web Conference, 2016, pp. 498–514.
[18] I. Yamada, A. Asai, J. Sakuma, H. Shindo, H. Takeda, Y. Takefuji, Y. Matsumoto, Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 23–30.
[19] F. Hoppe, D. Dessì, H. Sack, Deep learning meets knowledge graphs for scholarly data classification, in: Companion Proceedings of the Web Conference 2021, 2021, pp. 417–421.