ProZe: Explainable and Prompt-guided Zero-Shot Text Classification

Ismail Harrando[*]**, Alison Reboud**[*]**, Thomas Schleider**[*]**, Thibault Ehrhart and Raphael Troncy**
EURECOM, Sophia Antipolis, France

***Abstract*—As technology accelerates the generation and communication of textual data, the need to automatically understand this content becomes a necessity. In order to classify text, being it for tagging, indexing or curating documents, one often relies on large, opaque models that are trained on pre-annotated datasets, making the process unexplainable, difficult to scale and ill-adapted for niche domains with scarce data. To tackle these challenges, we propose *ProZe*, a text classification approach that leverages knowledge from two sources: prompting pre-trained language models, as well as querying ConceptNet, a common-sense knowledge base which can be used to add a layer of explainability to the results. We evaluate our approach empirically and we show how this combination not only performs on par with state-of-the-art zero shot classification on several domains, but also offers explainable predictions that can be visualized.**

***Keywords:* Text classification, zero-shot, explainability, common sense knowledge graph, prompting language models**

Introduction

The Natural Language Processing (NLP) and Information Extraction (IE) fields have seen many recent breakthroughs, especially since the introduction of Transformer-based approaches and BERT [1], which has become the *de-facto* family of models to tackle most NLP tasks. Over the last years, few-shot and zero-shot learning approaches have gained momentum, particularly for the cases with little data and where uncommon or specialized vocabularies are being used. Fully zero-shot classification approaches do not require any training data and often show respectable performance. An interesting new paradigm is *prompt-based learning* which leverages pre-trained language models through prompts (i.e. input queries that are handcrafted to produce the desirable output) instead of training models on annotated datasets. However, a major downside of all these approaches based on transformer-based language models is that they suffer from a lack of explainability.

Recently, ZeSTE [2] tackled this lack of interpretability problem in text classification by departing from language models and relying instead on ConceptNet [3] and its explicit relations between words. With every word being a node in ConceptNet, ZeSTE can justify the relatedness between words in the document to classify its assigned label. While it shows state-of-the-art results in topic categorization, it does not offer ways to specialize the classifier beyond “common sense knowledge” (domain adaptation), nor does it offer the possibility to disambiguate labels. These challenges are important to solve for text classification of specific domains, especially since zero-shot classification is particularly useful for domain-specific use cases with little data to train

[*] Equal contribution

a model. As a consequence, this paper proposes *ProZe*, a Zero-Shot classification model which combines latent contextual information from pre-trained language models (via prompting) and explicit knowledge from ConceptNet. This method keeps the explainability property of ZeSTE while still offering a step towards label disambiguation and domain adaptation.

The remainder of this paper is structured as follows. First, we give an overview of the relevant state-of-the-art work. We then detail our proposed method called ProZE. Next, we present our results on common topic categorization datasets as well as on three challenging datasets from diverse domains: screenplay aspects for a crime TV series [4], historical silk textile descriptions [5], and the Situation Typing dataset [6]. We report and analyze the results of several empirical classification experiments, which includes a comparison to some state-of-the-art Zero-Shot approaches. Finally, we conclude and outline some future work.

Related Work

Language Models

Since the breakthrough performance by AlexNet on the 2012 ImageNet challenge[7], transfer learning via pre-trained models became a new standard in many machine learning tasks. With the introduction of the Transformers architecture [8], this paradigm shift made its way to the NLP field as well through the advent of *"pre-trained language models"*.

The most influential Transformer-based model is BERT [1]. Its defining feature is its ability to pre-train deep bidirectional representations. Such pre-trained language models remain part of the most successful approaches for a many NLP tasks, such as text classification. Despite the wide availability of these language models, many classification experiments require also annotated and balanced training data to make a model properly associate documents with labels, which can be either expensive or not available at all for niche domains.

Zero-Shot Classification

With rising popularity of zero-shot classification methods, there are now more attempts to benchmark and evaluate them on text classification approaches. [9] provides a survey of the recent advances in the field, while proposing *Entail*, a zero-shot classification model based on using language models fine-tuned on the task of Natural Language Inference to classify documents. Some zero-shot classification models also takes advantage of “prompt-based learning” [10], a new paradigm used for many NLP tasks that allows to extract information out of Language Models.

Explainability in NLP

One direction of a growing amount of work interested in explainable methods is to generate explanations and to develop evaluations that measure the extent and likelihood that an explanation and its label are associated with each other in the model that generated them [11]. However, none of these techniques totally compensate for the obscurity associated with language models. This is the main reason why the approach presented in this paper relies on ZesTE (Zero Shot Topic Extraction) [2], which is not based on a pre-trained language model, and provides explainability of its classification results using ConceptNet as a prediction support.

Previous contributions leverage knowledge graphs [12], [13], [14] and common-sense [15] to improve the performance of several classification tasks. To the best of our knowledge, our approach is the first to use a common-sense knowledge graph to not have a learning component, and uses the KG as is, allowing it to retain explainability.

ConceptNet

A central resource for this work is ConceptNet [3], a semantic network "*designed to help computers understand the meanings of words that people use*"[1]. Broadly speaking, ConceptNet is a graph of words (or *concepts*), connected by edges representing semantic relations that go beyond the lexical relations than can be found in a dictionary such as "Synonym" or "Hypernym". Most importantly, ConceptNet contains relations of general "relatedness" (or `/r/RelatedTo` on ConceptNet), which imply an undefined semantic relation between two concepts, such as "Business" and "Outsourcing": while both terms are used in similar contexts, one cannot define such relation as one of containment, usage or typing. It

[1] https://conceptnet.io

it notable that, unlike semantic similarity between two terms via word embeddings, ”relatedness” relations are usually mined for dictionary entries or corresponding Wikipedia articles, thus making them explainable to the user.

Other than the knowledge graph, ConceptNet comes with its set of graph embeddings called ”ConceptNet Numberbatch”. Computed in a special way to reflect both the connectedness of nodes on the ConceptNet graph and the linguistic properties of words via retrofitting to other pretrained word embeddings [3], these embeddings can better capture semantic relatedness between words, as demonstrated by their performance on the SemEval 2017 challenge (https://alt.qcri.org/semeval2017).

We use both the semantic graph for generating explanations and the Numberbatch embeddings to prune out excessive and noisy relations in our method.

Method

Our model can be seen as a pipeline comprising several components. In this section, we explain each step of the process in further details.

Generating Label Neighborhoods

The first step of our approach is to manually create mappings between target class labels and their ConceptNet nodes. For instance, if we want our classifier to recognize documents for the class “sport”, we designate the node `/c/en/sport` as our starting node.[2]

Based on these mappings between target labels and concept nodes, we can then generate a list of candidate words (from ConceptNet) that are related to the respective concept. This list can be called the ”label neighborhood”. Each of the candidate is produced by retrieving every node that is N-hops away from the class label node.

Afterwards, a score can be calculated for each label based on which words are present in the input text or document to classify. To this end, we score every word in the label neighborhood based on its ”similarity” to the class label.

Scoring a Document

Like ZeSTE, we proceed to score each document by first generating a score for each node in a label neighborhood. To do so, multiple approaches exist. In this paper, we present and compare 3 such scoring methods (SM):

1) **ConceptNet embeddings similarity (SM1)**: ConceptNet Numberbatch[3] are graph embeddings computed for ConceptNet nodes. To quantify their similarity, we compute cosine similarity between the embedding of each node on the label neighborhood and the label node itself.
2) **Scoring through Inference (SM2)**: for this scoring method, we use a model that is pretrained on the task of Natural Language Inference. In a similar setting to the previous method, we prompt the model with a sentence related to the label or its domain, and then we ask it to score all the words from its neighborhood based on the logical entailment between the prompt (premise) and a template containing the word (hypothesis).
3) **Language Modeling Probability (SM3)**: for this scoring method, we combine the predictive power of language models with the explicit relations that we can find on the label neighborhood. For each label, we supply the language model with a *prompt*, or a sentence that is likely to guide it towards a specific meaning of the label we target (for example, the definition of the label), and then, we ask it to predict the next word in a Cloze statement (a sentence where one word is removed and replaced by a blank). For example, to score words related to the label ”sport”, we can give the model a definition of the word, and then ask it to predict the blank word in the following Cloze statement: *”Sport is related to [blank].”*. Given that language models, are pre-trained on predicting such blanks, we can use the scores they attribute to that blank to measure the similarity between our label and the candidate words from its neighborhood. For instance, when we give the dictionary definition of sport to the language model, the top predicted words are ’recreation’, ’fitness’ and ’exer-

[2] In the remainder of this paper, we will omit the prefix `/c/en/` as alllabels in our datasets are in English.

[3] https://github.com/commonsense/conceptnet-numberbatch

cise'. Because the language model outputs a probability for every word in its vocabulary, we score only the words that are originally on the label neighborhood. If a word in the neighborhood does not appear among the predictions of the model (i.e. out of the model's vocabulary), the score from SM1 is used.

Once the scores are computed by one of these methods, we can proceed to score any document given as input to the model. To score such document, we first tokenize it into separate words. We then take all the nodes from the neighborhood of a label that appear in the tokenized document, and we add up their scores to produce a score for the label. We do so for each label we are targeting, and the final prediction of the model corresponds to the label with the highest score. Because all the nodes in the neighborhood are linked to the label node with explicit relations on ConceptNet, we can explain in the end how each word in the document contributed to the score and how it is related to our label.

Prompting Language Models

In this section, we explain how we leverage language models to score the label neighbors extracted from ConcepNet, as per the scoring methods SM2 and SM3 described above.

Both SM2 and SM3 methods rely on prompting the language model, i.e. to feed it a sentence that would function as a context to ”query” its content (also known as *probing* [16]). As expressed in the related work, prompting language models is an open problem in the literature. In this work, we explore some potential ideas for prompting to serve our objective of measuring word-label relatedness.

The prompting follows the same scheme for both scoring methods. We vary both the premise and hypothesis templates and report the results for some proposals in the Evaluation section. For the premise, we experiment with two approaches:

1) Domain description: where we prime the model with the name or description of the domain of the datasets, i.e. ”Silk Textile”, ”Crime series”, etc.
2) Label definition: where we prime the model with the definition of the label, with the assumption that this will help it disambiguate the meaning of the label and thus come up with better related words. For instance, for the label ”space”, we provide the language model with the sentence ”Space is the expanse that exists beyond Earth and between celestial bodies”. We take the definitions from Wikipedia or a dictionary, we generate it using a NLG model etc.

We observed experimentally that using just the description of the domain as a prompts gives better overall performance. Therefore, we only report results on these prompts in the following sections. As for the hypothesis, we provide the model with a sentence like *”[blank] is similar to **space**”* or *”**Space** is about [blank]”* which we use in our reported results.

We note that, while the combination of premise and hypothesis can impact the overall performance of the model, the search space for a good prompt is quite wide. Thus, we only report the performance on some combinations, as we intend this paper to only point out the use of such mechanism for this task rather than fully optimize the process.

Tool Demonstrator

To explain the decisions of the model, we follow the same method as ZeSTE [2], i.e. we highlight the words which contribute to the decision of the classification as shown in a graph that links them with semantic relations to the label node. The difference is that the scores in ProZe take also into account the scoring from the language model. To illustrate the contribution of the language model, we developed an interactive demonstrator enabling a user to test the effect of prompting the language model to improve the results of zero-shot classification (Figure 1). This demonstrator is available at http://proze.tools.eurecom.fr/.

After choosing a label to study, the user is asked to enter a prompt that can help the model to identify words related to the label (e.g. definition or domain). The user is then shown an abridged version of the prompt-enhanced label neighborhood: the connection between any node and the label node is omitted for clarity but it can be trivially retrieved from ConceptNet, and only the top 50 (based on the used scoring) words are shown to represent the new label neighborhood, with the intensity of the color reflecting higher scores.

The user can view in detail the updates happening before and after introducing the new scoring from the Language Model. For this demonstration, we use the SM3 method to score the nodes as it requires only one pass through the Language Model to generate a score for all words in its vocabulary, whereas the SM2 method requires an inference for every word in the label neighborhood. As a consequence, while the SM2 methods takes up to 7 minutes per label on our hardware, the SM3 method takes less than a second while still delivering good performance.

Datasets

In this section, we present three widely used topic categorization datasets in the news domain, as well as three other very different and domain-specific datasets making used of fine-grained labels.

News Topics Datasets

Used to benchmark multiple text classification approaches, news datasets are often categorized by topic and are written in simple and common language. In our experiments, we report results on three such commonly-used datasets: AG News, BBC News and 20NG.

- **20 Newsgroups** [17]: a collection of 18000 user-generated forum posts arranged into 20 groups seen as topics such as *"Baseball", "Space", "Cryptography", and "Middle East"*.
- **AG News** [18]: a news dataset containing 127600 English news articles from various sources. Articles are fairly distributed among 4 categories: *"World", "Sports", "Business" and "Sci/Tech"*.
- **BBC News** [19]: a news dataset from BBC containing 2225 English news articles classified in 5 categories: *"Politics", "Business", "Entertainment", "Sports" and "Tech"*.

Crisis Situations

The first low-resource classification dataset we use is the Situation Typing dataset [6]. The goal is to predict the type of need (such as the need for water or medical care) required in a specific situation or to identify issues such as violence. Therefore, this dataset constitutes a real world, high-consequence domain for which explainability is particularly important. The entire dataset contains 5,956 labeled texts and 11 types of situations: "food supply", "infrastructure", "medical assistance", "search/rescue", "shelter", "utilities, energy, or sanitation", "water supply", "evacuation", "regime change", "terrorism", "crime violence" and a "none" category. In our experiment, we use the test set (2343 texts), where we only select texts that represent at least one of the situations and we consider it a success if the model predicts at least one correct label.

Crime Aspects

The Crime Scene Investigation (CSI) dataset contains 39 CSI video episodes together with their screenplays segmented into 1544 scenes [4]. An episode scene contains on average 21 sentences and 335 tokens. Originally, this dataset is used for screenplay summarization as each scene is annotated with a binary label denoting whether it should be part of a summary episode or not. Additionally, the three annotators had to justify their choice of their selected summary scenes with regards to it being about one/more or none of the following six aspects: i) victim, ii) the cause of death, iii) an autopsy report, iv) crucial evidence, v) the perpetrator, and vi) the motive/relation between perpetrator and victim.

We define the following labels to evaluate the ProZe system: victim, cause of death, crime scene, evidence, perpetrator, motive. For our classification task, we kept only the scenes which were associated to at least one aspect (449 scenes). In the case where one scene is associated to multiple labels, if the model predicts one of the labels, we consider it a success.

Silk Fabric Properties

This dataset is an excerpt from the multilingual knowledge graph of the European H2020 SILKNOW research project[5] aiming at improving the understanding, conservation and dissemination of European silk heritage. The SILKNOW knowledge graph consists of metadata about 39,274 unique objects integrated from 19 museums and represented through a CIDOC-CRM-based set of classes and properties. This metadata about silk fabrics contains usually both explicit categorical information, like specific weaving techniques or

[4] https://github.com/EdinburghNLP/csi-corpus

[5] https://silknow.eu/

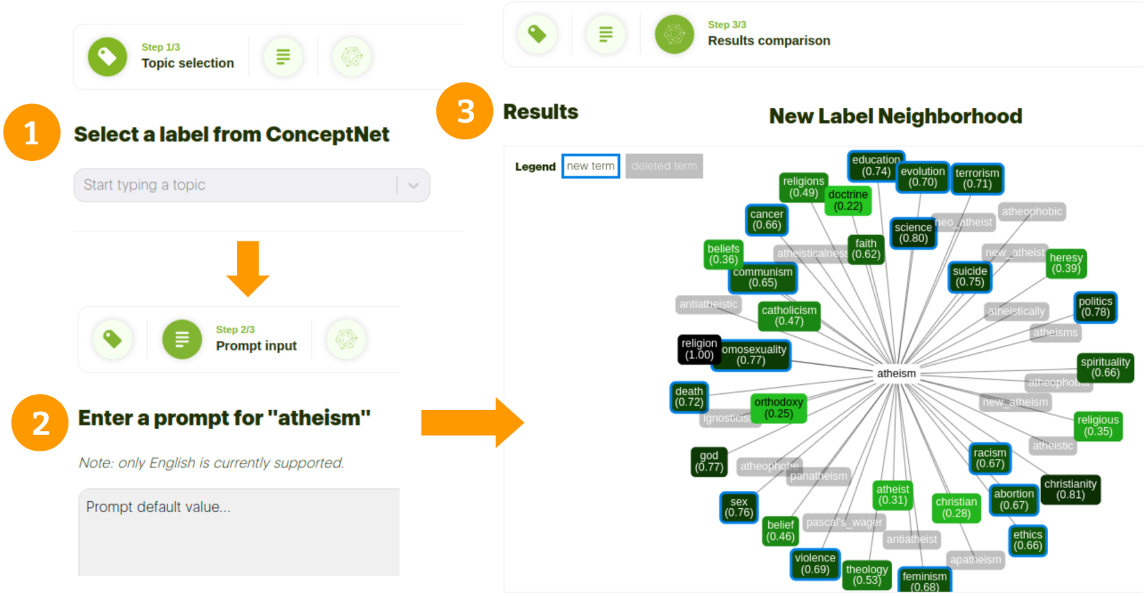

Figure 1: ProZe neighborhoods demo. (1) The user is asked to select a label (2) The user can input a text to prompt and guide the language model. (3) The user can visualize the label neighborhood, with added and removed nodes highlighted, and is shown a detailed list of all the changes resulting from the prompt.

their production years, but also rich and detailed textual descriptions. Our goal is to try to predict categorical values based on these text descriptions.

The SILKNOW Knowledge Graph dataset can be divided into using ”material” and ”weaving technique” subsets. More precisely, we slightly extend the dataset used in [20], and after removing objects with more than one value per property, we obtain 1429 object descriptions making use of 7 different labels for silk materials, and 833 object descriptions with 6 unique labels for silk techniques. The chosen labels have also to be mapped to ConceptNet entries to work with this approach. Table 1 shows the final selection of thesaurus concepts and their mapping to ConceptNet nodes.

Property	SILKNOW Concept	ConceptNet
Material	Cotton	`/c/en/cotton`
Material	Wool	`/c/en/wool`
Material	Textile	`/c/en/textile`
Material	Metal thread	`/c/en/metal`
Material	Metal silver thread	`/c/en/silver`
Material	Silver thread	`/c/en/silver`
Material	Gold thread	`/c/en/gold`
Technique	Damask	`/c/en/damask`
Technique	Embroidery	`/c/en/embroidery`
Technique	Velvet	`/c/en/velvet`
Technique	Voided Velvet	`/c/en/velvet`
Technique	Tabby (silk weave)	`/c/en/tabby`
Technique	Muslin	`/c/en/tabby`
Technique	Satin (Fabric)	`/c/en/satin`
Technique	Brocaded	`/c/en/brocaded`

Table 1: Mapping between the concepts used in the SILKNOW knowledge graph and ConceptNet (ProZe and ZeSTE)

Evaluation

We evaluate ProZe on these 6 datasets. In this section, we present the results of this evaluation.

Baselines

We compare our model with:

- *ZeSTE*: this approach solely relies on ConceptNet to perform Zero-Shot classification;
- *Entail*: this model was originally proposed in [9]. We use `bart-large-mnli` as the backend Transformer model, which it is a version of **BART** [21] that was been fine-tuned on the Multi-genre Natural Language Inference (MNLI) task, as per the implementation we use for our experiments (can be tested at

https://huggingface.co/zero-shot/). Given a text acting as a *premise*, the task of Natural Language Inference (NLI) aims at predicting the relation it holds with an *hypothesis* sentence, labelling it either as false (contradiction), true (entailment), or undetermined (neutral). Generally, the labels are injected in a sentence such as "This text is about" + label, to form an *hypothesis*. The confidence score for the relation between the text to be labelled and the premise to be 'entail' is the confidence of the label to be correct. We use the implementation provided at https://github.com/katanaml/sample-apps/tree/master/01)

Quantitative Analysis

We limit the size of the label neighborhoods to 20k per label for each experiment, except in cases where querying ConceptNet returns less nodes than that. Then, we resize all the other neighborhoods to be all equal in size to the smallest one (by eliminating the nodes with the lowest similarity), as we found that having neighborhoods of different sizes skews the predictions towards the larger ones (by virtue of having more nodes to contribute to the score). This can be circumvented by increasing the number of hops (thus boosting the size of smaller neighborhoods before filtering), but according to our observations, this hurts the quality of the kept nodes as they get less semantically relevant as we hop further. Resizing the neighborhoods eliminate the bias against the in-domain labels that may not have so many related words in the first place.

Table 3 and Table 2 show a score comparison of the ProZe approaches to the baselines of ZeSTE and the Entail approach. **ProZe-A** refers to scoring the nodes using a combination of SM1 and SM2, whereas **ProZe-B** uses a combination of SM1 and SM3. We tested several ways to combine the scores from ConceptNet (SM1) and language models (SM2 and SM3), including taking the sum of the two scoring methods, their product, their max, or a weighted average. Empirically, we obtain the best empirical results by multiplying the two scores (both normalized to be between 0 and 1). The main advantage of multiplication is that it penalizes disagreement between the language model and the KG over how close two terms are. This also means that the explainability layer reflects accurately the decisions of the model, as words that are not scored well by the language model will not contribute significantly to the classification score.

Table 2 contains the accuracy and weighted average scores for the 3 news datasets that consist of general knowledge texts. ProZe has similar performance, but not beating ZeSTE, which is in line with our expectations: both approaches are based on the ConceptNet commonsense knowledge graph, and the vocabulary does not need or cannot be guided into a more fitting direction with the prompts. For all three news datasets, however, ProZe performs better than Entail.

Table 3 shows the results for the 3 domain-specific datasets. We observe that ProZe is consistently outperforming ZeSTE, which we take as a confirmation that the guidance through the prompt is effective for specific domains. For two datasets, silk material and situations, ProZe even beats the non-explainable baseline scores of the Entail approach. This is not the case for the silk technique and the CSI screenplay datasets as some labels from these datasets have very limited neighborhoods in ConceptNet. Nevertheless, our approach is still close and retains in all cases its higher degree of explainability.

Qualitative Analysis

To illustrate why a re-ranking of related words induced by a domain prompt improves the score, we analyse a concrete example. Taken from the silk technique dataset, the top 10 candidate terms of the ConceptNet label neighborhood for the weaving technique ”embroidery” are as follows: ”Embroidery, overstitch, running stitch, picot, stumpwork, arresene, couture, fancywork, embroider, berlin work”. While these words are clearly related to the concept of embroidery, they are not necessarily relevant in the context of silk textile. For example, ”picot” is a dimensional embroidery related to crochet. The intuition is then that this neighborhood can be improved by specifying the domain.

In comparison, the top 10 candidate terms of the pre-trained BART language model, guided by a prompt that included the term ”silk textile” are: ”Craft artifact sewn, fabric, embroidery stitch, embroidery, detail, embroider, mending, embellishment, elaboration, filoselle”. These terms are

Datasets	20 Newsgroup		AG News		BBC News	
	Accuracy	Weighted Avg	Accuracy	Weighted Avg	Accuracy	Weighted Avg
ZeSTE	63.1%	63.0%	**69.9%**	**70.3%**	84.0%	84.6%
Entail	46.0%	43.3%	66.0%	64.4%	71.1%	71.5%
ProZe-A	62.7%	62.8%	68.5%	69.1%	83.2%	83.7%
ProZe-B	**64.6%**	**64.6%**	69.0%	69.6%	**84.2%**	**84.8%**

Table 2: Prediction scores for the news datasets (the top score in each metric is emboldened).

Datasets	Silk Material		Silk Technique		Crime aspects		Crisis situations	
	Accuracy	Weighted Avg	Accuracy	Weighted Avg	Accuracy	Weighted Avg	Accuracy	Weighted Avg
ZeSTE	34.3%	39.0%	46.9%	47.2%	31.2%	32.3%	46.3%	45.8%
Entail	29.0%	33.3%	**64.0%**	**65.8%**	**43.7%**	**43.7%**	46.7%	48.1%
ProZe-A	**39.0%**	40.1%	50.8%	57.6%	36.3%	37.6%	**50.1%**	49.7%
ProZe-B	37.4%	**41.7%**	48.5%	48.7%	29.8%	31.1%	**50.1%**	**49.8%**

Table 3: Prediction scores for the domain-specific datasets (the top score in each metric is emboldened).

more general even if also related to silk textile. Words such as ”detail”, ”mending”, ”elaboration” or ”embellishment” seem useful for classifying texts that are not only consisting of details about different types of embroidery. When combining the scores from ConceptNet and the language model, the ProZe method increases its F1 score of circa 8%, from 61% to 69%.

Conclusion and Future Work

In this paper, we demonstrated the potential of fusing knowledge about the world from two sources: First, a common-sense knowledge graph (ConceptNet), which explicitly encodes knowledge about words and their meaning. Second, pre-trained language models, which contain a lot of knowledge about language and word usage that is latently encoded into them. We explored several methods to extract this knowledge and leverage it for the use case of zero-shot classification. We also empirically demonstrated the efficiency of such combination on several diverse datasets from different domains.

This work is experimental and does not fully explore all possibilities of this setup. As future work, we want to study the effect of prompt choice in more detail, and seeing how such choice impacts not only the quality of the predictions but also that of the explanations. Different language models can also be tried to measure how such choice can improve the overall classification, especially for specific domains such as e.g. medical documents.

Another potential improvement over this method is to filter out words unrelated to the label using the slot-filling predictions from the language model. From early experiments, this method seems to give good results by restricting the neighborhood nodes to ones that almost exclusively relate to the label in some way.

A natural direction of work is to involve the user in the creation of the label neighborhood (human-in-the-loop) by asking whether some words that only the Language Model and not ConceptNet suggests pertain to the target label. This allows to inject the extracted knowledge from the language model back into the zero-shot classifier, and fill in the gaps of knowledge from ConceptNet.

Finally, some existing limitations of the original work can be still improved upon such as letting the language model inform the label selection and expansion, handling multi-word labels, and integrating more informative concepts from ConceptNet beyond word tokenization (e.g. *’crime_scene’*, *’tear_gaz’*).

Acknowledgment

This work has been partially supported by European Union's Horizon 2020 research and innovation programme within the Odeuropa project (grant agreement No. 101004469), by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003) and by raisin.ai within the MyLittleEngine project.

REFERENCES

1. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL*. Association for Computational Linguistics, 2019, pp. 4171–4186.
2. I. Harrando and R. Troncy, "Explainable Zero-Shot Topic Extraction Using a Common-Sense Knowledge Graph," in *3^{rd} Conference on Language, Data and Knowledge (LDK)*, Zaragoza, Spain, 2021.
3. R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *31^{st} AAAI Conference on Artificial Intelligence*, 2017.
4. L. Frermann, S. B. Cohen, and M. Lapata, "Whodunnit? crime drama as a case for natural language understanding," *TACL*, vol. 6, pp. 1–15, 2018.
5. T. Schleider, T. Ehrhart, P. Lisena, and R. Troncy, "Silknow knowledge graph," Nov. 2021.
6. S. Mayhew, T. Tsygankova, F. Marini, Z. Wang, J. Lee, X. Yu, X. Fu, W. Shi, Z. Zhao, and W. Yin, "University of pennsylvania lorehlt 2019 submission," Technical report, Tech. Rep.
7. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.
8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, vol. 30. Curran Associates, Inc., 2017.
9. W. Yin, J. Hay, and D. Roth, "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach," in *EMNLP*, 2019, pp. 3914–3923.
10. P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.
11. B. Paranjape, J. Michael, M. Ghazvininejad, H. Hajishirzi, and L. Zettlemoyer, "Prompting contrastive explanations for commonsense reasoning tasks," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: ACL, Aug. 2021, pp. 4179–4192.
12. Q. Chen, W. Wang, K. Huang, and F. Coenen, "Zero-shot text classification via knowledge graph embedding for social media data," *IEEE Internet of Things Journal*, pp. 1–1, 2021.
13. T. Liu, Y. Hu, J. Gao, Y. Sun, and B. Yin, "Zero-shot text classification with semantically extended graph convolutional network," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 8352–8359.
14. J. Zhang, P. Lertvittayakumjorn, and 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1031–1040. [Online]. Available: https://aclanthology.org/N19-1108
15. N. Nayak and S. Bach, "Zero-shot learning with common sense knowledge graphs," 2021. [Online]. Available: https://openreview.net/forum?id=jYkO_0z2TAr
16. A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties," in *ACL*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2126–2136.
17. K. Lang, "Newsweeder: Learning to filter netnews," in *ICML*, 1995, pp. 331–339.
18. A. Gulli, *AG's corpus of news articles*, 2005. [Online]. Available: http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
19. D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *23^{rd} International Conference on Machine learning (ICML)*, 2006, pp. 377–384.
20. T. Schleider and R. Troncy, "Zero-shot information extraction to enhance a knowledge graph describing silk textiles," in *Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, Nov. 2021, pp. 138–146.
21. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, 2020.