Description Boosting for Zero-Shot Entity and Relation Classification

Anonymous ACL submission

Abstract

Zero-shot entity and relation classification models leverage available external information of unseen classes – e.g., textual descriptions – to annotate input text data. Thanks to the minimum data requirement, Zero-Shot Learning (ZSL) methods have high value in practice, especially in applications where labeled data is scarce. Even though recent research in ZSL has demonstrated significant results, our analysis reveals that those methods are sensitive to provided textual descriptions of entities (or relations). Even a minor modification of descriptions can lead to a change in the decision boundary between entity (or relation) classes. In this paper we formally define the problem of identifying effective descriptions for zero shot inference, we propose a strategy for generating variations of an initial description, a heuristic for ranking them and an ensemble method capable of boosting the predictions of zero-shot models through description enhancement. Empirical results on four different entity and relation classification datasets show that our proposed method outperform existing approaches and achieve new SOTA results on these datasets under the ZSL settings. The source code of the proposed solutions and the evaluation framework are open-sourced. [1]

1 Introduction

Zero-shot learning (ZSL) is a classification task in machine learning where – at inference time – samples are classified into one of several classes which were not observed during training. Having a classifier that can generalize to new unseen classes is important for a variety of practical reasons. First, ZSL methods can be used to learn models that are more robust to labeled data shortages and distributional shifts. Moreover, they can be used to extend the reach of models to new domains.

[1]Anonymized for double-blind review

ZSL approaches in the Natural Language Processing (NLP) domain have seen significant improvements in recent years thanks to the availability of large pre-trained Language Models (LMs). For example, it has been shown that models such as GPT-3 (Brown et al., 2020), OPT (Zhang et al., 2022) and FLAN (DBL) achieve strong performances on many NLP tasks, including translation, question-answering, and cloze tests without any gradient updates or fine-tuning.

For entity recognition – including classification and linking – and relation classification problems, recent ZSL methods (Aly et al., 2021; Ledell Wu, 2020; Chen and Li, 2021a) rely on textual descriptions of entities or relations. Descriptions provide the required information about the semantics of entities (or relations), which help the models to identify entity mentions in texts without observing them during training. Works such as (Ledell Wu, 2020; De Cao et al., 2021) and (Aly et al., 2021) show how effective it is to use textual descriptions to perform entity recognition tasks in the zero-shot context. The same mechanism can also be applied in other contexts such as relation classification (Chen and Li, 2021b).

An example of named entity classification with ZSL is demonstrated in Figure 1. At inference time, a zero-shot model is given short textual descriptions of new entity classes such as *Company* or *Fruits*, it then identifies and annotates mentions of those entity classes in an input sentence. Although state-of-the-art ZSL methods such as SMXM (Aly et al., 2021) have demonstrated significant results in recent research works, this toy example shows how the quality of the provided descriptions influences the accuracy of these models. For example, in Figure 1 even with a small modification of the *Company* entity class description, the SMXM model changes its entity prediction. In practice, the sensitivity to entity descriptions is problematic because, for non-expert users, it is not a trivial task to

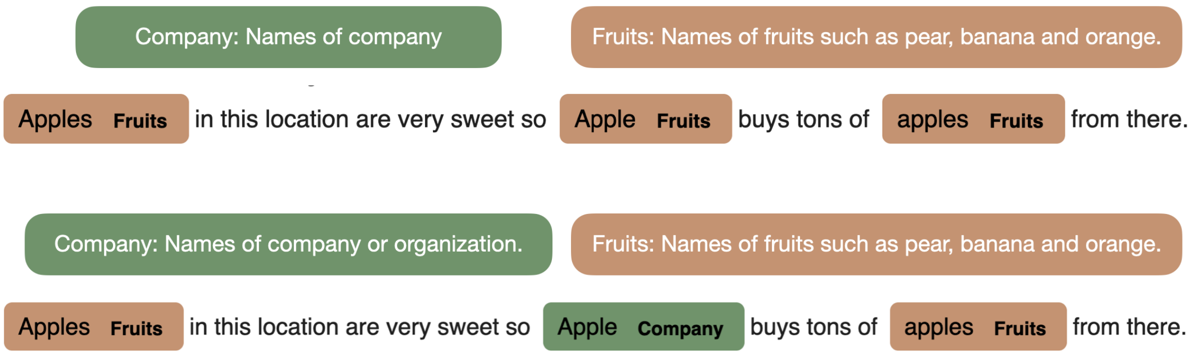

Figure 1: A small modification of the *Company* class description results in different entity predictions.

choose a proper description for black-box zero-shot models, in particular in an unfamiliar domain.

In this paper, we study different methods for boosting model performance with automatic description enhancement. Specifically, we propose UDEBO (for Unsupervised DEscription BOosting), the first unsupervised method capable of automatically modifying/generating description to improve entity (or relation) predictions in the zero-shot settings. We present several strategies to alter descriptions, such as using a generative model, paraphrasing, and summarization combined with description ranking/ensemble methods to reduce model uncertainty and increase overall performance. We empirically evaluate the performance of UDEBO on 4 existing standard zero-shot datasets, spanning two tasks: (i) name entity classification and (ii) relation classification.

Our results show that for the zero-shot entity classification tasks, UDEBO improved the results of state-of-the-art models by 7 and 1.3 percentage points in terms of Macro F1 Score in the OntoNotes and MedMentions datasets, respectively. For what concerns relation classification, we achieve a performance improvement of 6 and 3 percentage points (Macro F1 Score) on the FewRel and WikiZS datasets over our baseline models, respectively.

We organize the paper as follows. In Section 2 we provide a description of the zero-shot setting for entity recognition and relation classification. We also formally define the problem we aim to solve in this paper, i.e. how to enhance entity or relation descriptions to improve the performance of zero-shot models. In Section 3 we describe the proposed approaches for description boosting while in Section 4 we describe our experimental setup and results. We provide a literature review and draw the conclusions of our work in Sections 5 and 6, respectively.

2 Preliminaries and problem definition

Entity and relation classification are key steps to extract or query knowledge from unstructured documents. Zero-shot approaches can identify which tokens in a text refer to an entity (its mention) and determine its type (entity typing) without the need of observing other instances of the same entity during training.

In ZSL, the sets of training and test entity (or relation) classes are disjoint. Therefore, the strategy employed by zero-shot models is to rely on prior general knowledge that could be transferred to unseen instances at inference time. In particular, novel zero-shot approaches leverage the fact that textual descriptions for entity classes are either available in existing datasets or can be easily provided by users.

Given a textual description of an entity class (or relation) of interest, zero-shot models recognize mentions in a text and predict whether the given mentions belong or not to the entity class (or relation) with a certain probability. One classic paradigm is to embed all entities with their textual description and the input sentence with each mention into one common space and measure the probability of each entity by assessing their distance. Descriptions for model pre-training are typically sourced from Wikipedia by joining an entity page title or label with the first 10 sentences in the respective Wikipedia page (Wu et al., 2020). However, the quality of the descriptions has an impact on how effective the transfer of knowledge from observed to unseen entities (Aly et al., 2021).

Given a set of entity classes E (or relations) of interest with their textual descriptions D and a corpus of sentences S to annotate as input, we describe in Section 3.1 different strategies to generate new entity (or relation) descriptions D' for the input set E, intending to improve the accuracy of the predictions by the ZSL models over that corpus. We can define the problem of description enhancement as follows:

Problem 1 (Description enhancement) *Denote $\phi(D, S)$ as the function estimating the accuracy of ZSL models when using a given entity (or relation) description D for annotating an input text corpus S. Our goal is to generate a set of descriptions D^* such that:*

$$D^* = \arg\max_{D} \phi(D, S) \tag{1}$$

As exemplified in Figure 1, if the labeled data is known, it is possible to select the best descriptions via a brute force search across different description reformulations by measuring the accuracy as a function of D and S. However, given the absence of labeled data in the zero-shot context, an unsupervised approach is needed for ranking the descriptions D that yield the highest classification accuracy. In Section 3.2 and Section 3.3, we will discuss methods for ranking or combining predictions from different description variations to achieve better results.

3 Methods

We begin our discussion with methods for generating description variations before providing details about description ranking and ensemble strategies in the following subsections.

3.1 Generating description variations

Improving the completeness or clarity of entity (or relation) descriptions is a complicated problem without a formal definition of an objective function, as there is a large space of candidates to explore. To enhance entity (or relation) descriptions, in a more controlled way, we propose the following strategies:

Extension with pre-trained LMs. We propose to use large pre-trained LMs for generating text using the given description as context. Large LMs, as shown in (Petroni et al., 2019), capture linguistic and relational knowledge that can be extracted trough generation to extend a given description. In Section 4 we analyse the use of GPT-2 (Radford et al., 2019) for generating descriptions variations.

Extension with a fine-tuned LM. We fine-tune a LM for description generation and expansion. The LM is fine-tuned on a large dataset containing about 5.3 million Wikidata instances, including the name and the first few sentences of the respective articles. The model is fine-tuned on extending a truncated sub-string of the textual description, using a sequence to sequence objective. In Section 4 we analyse the use of a T5 large (Raffel et al., 2020) fine-tuned model for generating descriptions variations.

Summarization. Text summarization can be used to generate a concise description with less noise compared to the original one. In the experimental results we analyse the effect of using a BERT2BERT (small) (Turc et al., 2019) model fine-tuned on CNN/Dailymail for text summarization to enhance entities (or relations) descriptions.

Paraphrasing. Paraphrasing a description can simplify its linguistic form, using more common and general terms. In the experimental results we analyse the effect of using a Pegasus (Zhang et al., 2019) model fine-tuned for paraphrasing.

3.2 Description ranking via entropy

To rank a description for an entity (or relation), we propose to use a zero-shot model to first compute the probabilities of classes for each mention (or relation) in the input text with a candidate description. We then compute the information entropy H from this input. In information theory, entropy is the average level of "information" or "uncertainty" inherent to a variable's possible outcomes. Our assumption is that the lower the entropy is, the higher the confidence of the prediction will be, so Problem 1 can be reformulated as:

$$D^* = \arg\min_{D} H(D, S) \tag{2}$$

Where H is the entropy of a zero-shot model for a corpus S, using the description D to accomplish a certain classification task. This way we can rank different candidate descriptions and choose the best one without requiring any labeled data, which is ideal for the zero-shot setting.

3.3 Boosting performances with descriptions variations ensembling

Besides description ranking via entropy, we propose an ensemble method that combines predictions from multiple pipelines executed with different entity (or relation) descriptions. The main idea

Dataset	Split	Instances	Entities / Relations
MedMentionsZS	train	26770	11
	val	1289	5
	test	1048	5
OntoNotesZS	train	41475	4
	val	1358	4
	test	426	3
Fewrel	train	44800	64
	test	11200	16
WikiZS	train	70952	83
	val	12982	15
	test	9494	15

Table 1: Number of sentences and entities (or relations) for each split of the considered datasets.

behind this approach is to leverage the complementary information provided by the different definitions to make a more accurate prediction, reducing the variance and bias of an individual pipeline. Furthermore, using the methods described in section 3.1, the descriptions variations can provide additional information useful for correctly discriminating between unseen classes.

Entity description ensemble. Given a sentence, for each span s and an entity label $e \in E$, denote $v(s,e)$ as the number of pipelines that predict s or a sub-sequence of s with entity label e. For instance, given a span $s =$ *London Bridge*, assume that among ten pipelines, four pipelines predict the label of s as $e_1 =$ Facility, the other four pipelines predict the label of *London* as $e_2 =$ Location and the rest of the pipelines predict *Bridge* as Facility. Therefore, the accumulated number of votes for the span *London Bridge* are $v(s, e_1) = 6$ and $v(s, e_2) = 4$. Considering the majority of the votes, the final predicted label for the span *London Bridge* is Facility. Once the span *London Bridge* has been assigned a label, all of its sub-spans become redundant and thus are removed from consideration.

Relation description ensemble. For each set of descriptions generated using the strategies discussed in Subsection 3.1, we run a pipeline to obtain the predicted relation for each provided pair of entities. The votes of all the relations are aggregated across different pipelines. We use the majority voting rule to select the relation with the highest aggregated number of votes from different pipelines. That relation is considered as the output relation for the given pair of entities.

4 Experiments and Results

This section discusses experimental settings, baseline methods, and empirical results for both entity and relation classification tasks.

4.1 Datasets and experimental settings

We use two different settings: one for the Entity Classification (EC) task and one for the Relation Classification (RC) one.

Entity Classification setting. We use the pre-trained SMXM model (Aly et al., 2021) with the checkpoints available in the official GitHub repository. [2] We refer the reader to the original paper (Aly et al., 2021) to see the details of the implementation, the training parameters, and the datasets used for fine-tuning the model. There are two different checkpoints, one for each one of the datasets used, OntoNotes (Pradhan et al., 2013) and MedMentions (Mohan and Li, 2019). Both datasets have been processed as in the respective official GitHub repositories. Table 1 shows the number of rows and the entities of each dataset. Note that the number of rows reported in Table 1 refers to the zero-shot version of the dataset, containing only sentences with entities. See Appendix A for more information on this process and the datasets. The results reported are all based on the *test* split of the datasets.

Relation Classification setting. For RC, we use ZS-BERT [3] (Chen and Li, 2021b), a multitask learning model, based on BERT, to directly predict unseen relations. We trained our checkpoint using the official implementation of the model and following the steps of the official repository. [3] The datasets we use are FewRel (Han et al., 2018) and WikiZS (Sorokin and Gurevych, 2017). The results reported are all based on the *test* split of the datasets.

Description alteration settings. The language models used for the description alteration strategies: summarization, paraphrasing and pre-trained were obtained from the checkpoints available on Huggingface, while for the latter strategy we have fine-tuned a pre-trained T5-large model. We report detailed hyper-parameters of description alteration methods in section B of the appendix.

[2] https://github.com/Raldir/Zero-shot-NERC/

[3] https://github.com/dinobby/ZS-BERT

Datasets	Methods	Precision	Recall	Micro F1	Macro F1	Accuracy
OntoNotesZS	SMXM	20.96	48.15	30.76	29.12	86.36
	SMXM (Pre-trained)	24.05	**51.40**	32.77	32.78	87.69
	SMXM (Finetuned)	17.97	42.21	25.21	23.90	85.76
	SMXM (Summarization)	18.93	35.45	24.68	19.47	85.93
	SMXM (Paraphrased)	18.49	40.90	25.46	23.41	85.14
	SMXM (Combined)	18.86	42.58	26.15	23.74	84.83
	UDEBO	**31.14**	46.51	**36.78**	**36.15**	**88.29**
MedMentionsZS	SMXM	16.79	**40.55**	20.38	21.70	83.05
	SMXM (Pre-trained)	13.25	37.98	19.64	18.26	81.88
	SMXM (Finetuned)	13.67	36.05	19.82	19.13	83.18
	SMXM (Summarization)	10.96	26.68	15.37	17.92	83.02
	SMXM (Paraphrased)	14.77	26.51	18.97	19.41	86.74
	SMXM (Combined)	12.80	37.15	19.04	17.92	81.63
	UDEBO	**19.51**	32.73	**23.86**	**22.97**	**85.70**

Table 2: UDEBO, i.e. the ensemble of predictions with description variations, compared to the SMXM baseline.

Datasets	Methods	Precision	Recall	Micro F1	Macro F1	Accuracy
Fewrel	ZS-BERT	25.08	21.59	21.59	17.89	21.59
	ZS-BERT (Pre-trained)	18.25	25.29	25.29	19.10	25.29
	ZS-BERT (Finetuned)	19.39	16.09	16.09	14.59	16.09
	ZS-BERT (Summarization)	19.83	19.81	19.81	15.21	19.81
	ZS-BERT (Paraphrased)	25.89	21.76	21.76	19.90	21.76
	ZS-BERT (Combined)	17.09	16.53	16.53	16.53	16.53
	UDEBO	**28.38**	**25.68**	**25.68**	**22.12**	**25.68**
WikiZS	ZS-BERT	34.18	33.90	37.14	30.97	37.14
	ZS-BERT (Pre-trained)	14.73	15.80	14.29	11.72	14.29
	ZS-BERT (Finetuned)	16.23	16.26	16.62	13.65	16.62
	ZS-BERT (Summarization)	19.07	19.57	19.62	16.87	19.62
	ZS-BERT (Paraphrased)	25.50	27.60	27.60	24.56	27.60
	ZS-BERT (Combined)	17.34	19.62	18.43	16.27	18.43
	UDEBO	**34.79**	**37.11**	**40.17**	**34.25**	**40.17**

Table 3: UDEBO, i.e. the ensemble of predictions with description variations, compared to the ZS-BERT baseline.

4.2 Empirical results

This section discusses the results of entity (or relation) classification using methods for description enhancement.

4.2.1 Entity classification

Table 2 shows the results of the ensemble method (UDEBO) with ten descriptions generated by each of the description enhancing strategies, including pre-trained, finetuning, summarization and paraphrasing. For each enhancing strategy, we report the results when the descriptions with the lowest entropy are chosen for each class. The *Combined* strategy shows the results with the lowest entropy among all description-enhancing strategies.

We can see that the ensemble method (UDEBO) outperforms the SMXM baseline using the original descriptions provided on the OntoNotesZS dataset with a significant margin of 7 percentage points in terms of Macro F1 Score. On the MedMentionZS dataset, the improvement is 1.3 percentage points on the same reference performance measure (Macro F1 Score). Description ranking based on entropy works well with the pre-trained strategy on OntoNotesZS. However, the entropy does not seem to be a reliable score of model uncertainty on the MedMentionsZS dataset. Finding an alternative uncertainty score to entropy could be considered as future work. Overall, these results confirm our hypothesis – discussed in Section 1 – that zero-shot methods are sensitive to provided descriptions and that an ensemble of description enhancement methods is needed to obtain more robust results.

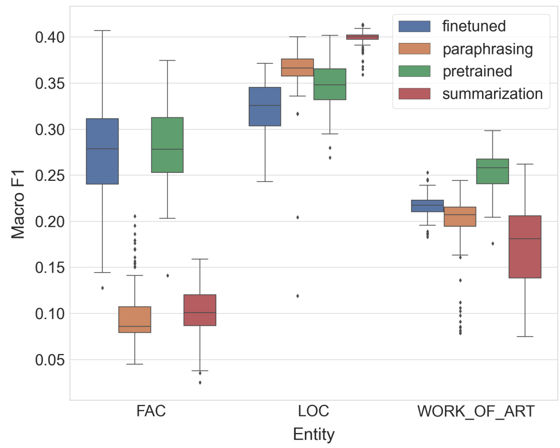

Figure 2: The figure shows the distributions of Macro F1 Score values on the test split of the OntoNotesZS dataset for each class, using the strategies described in Section 3.1 to generate 100 description variations for each class.

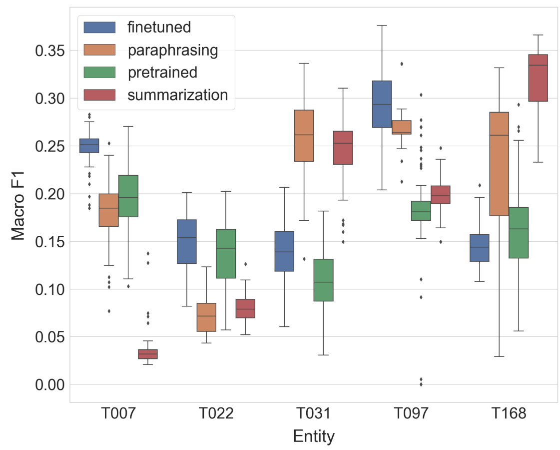

Figure 3: The figure shows the distributions of Macro F1 Score values on the test split of the MedMentions dataset for each class, using the strategies described in Section 3.1 to generate 100 description variations for each class.

4.2.2 Relation classification

In Table 3, we report our evaluation of the proposed approaches on the RC task. The results we observe here are similar to what we described for entity classification where the proposed ensembling method (UDEBO) achieves a higher performance across different measures compared to the baseline ZS-BERT model that does not rely on any relation description reformulation approach. We also observe on the FewRel dataset a higher Macro F1 Score associated with most of the description enhancement variants when employed independently from each other. These results further validate the strength of the proposed approach to enhance relation descriptions employed by ZSL models to improve their performance.

4.2.3 Descriptions enhancement strategies comparison and limitations

Generating variations of descriptions is relatively simple, as described in Subsection 3.1, several strategies allow to generate plausible extensions or variations of a text. Considering the results of ranking the descriptions using entropy in Section 4, we analyze and discuss here the correlation between Macro F1 Score and entropy measures and the limitations of the proposed approach.

Figure 2 and Figure 3 show the distributions of the Macro F1 Score on the test split of the OntoNotesZS and the MedmentionsZS dataset for each class, using the strategies described in Section 3.1 to generate 100 description variations for each class. None of the strategies is a clear champion over all the classes. The high variance of the performance explains the fact that the ensemble method makes a better prediction as observed in Table 2 and Table 3 thanks to successfully combining the strength of individual description alteration strategies. Figure 4 shows the correlations between Macro F1 Score and entropy for each unseen class on the OntoNotesZS test split with 100 description variations. Although there appears to be a significant statistical correlation using a sign test with ($p\text{-}value = 0.03$) between Macro F1 Score and entropy measures on the OntoNotesZS test set, the correlation does not appear to be statistically significant in the MedMentionsZS dataset. Also, as evidenced by the results in Table 2 and 3, using the descriptions with minimum entropy does not seem like a good strategy for selecting descriptions.

This phenomenon may be due to several factors like the change in the style of generated descriptions compared to the ones observed during training. Although a new description might seem more relevant, it could make the model more uncertain. See an example in Appendix C.2. The importance of this problem motivates the future study of alternative heuristics with more significant correlations, indirectly unveiling the mechanism behind zero-shot predictions.

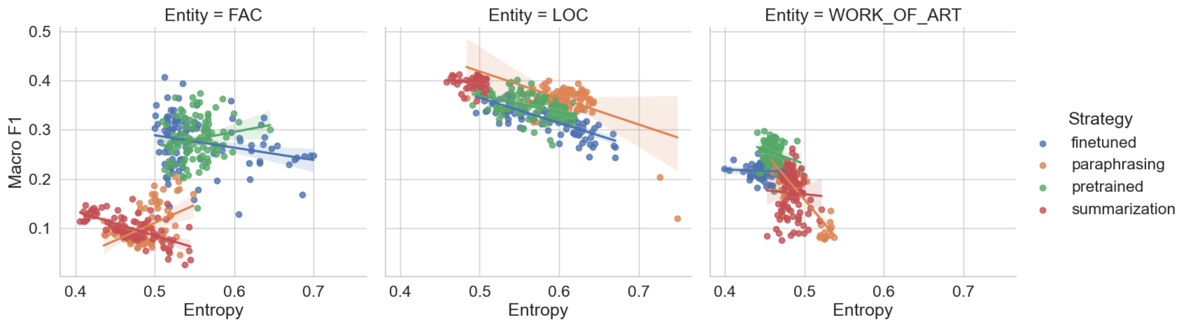

Figure 4: Analysis of the correlation between entropy and Macro F1 Score on unseen classes on the OntoNotesZS test split. Entropy can be calculated without the need for labeled data, therefore, if a correlation exists it can be used as an unsupervised heuristic to select descriptions that improve model performance.

5 Related work

Zero-shot entity recognition and linking. Zero-shot end-to-end entity linking refers to the task of detecting and disambiguating entity mentions by linking them to an entity in a Knowledge Base (KB), without requiring new labeled data. KBs are inherently incomplete and evolve over time with the addition of new entities and relations. Zero-shot entity linking usually relies on available textual information, or other set of relations in the KB, to generalise to entity sets unseen in the training data.

BLINK (Wu et al., 2020) is a BERT-based solution for Zero-shot linking of textual mentions – extracted for example using FLAIR (Akbik et al., 2018) – to entities in Wikipedia. It follows a bi-encoder architecture, each mention is encoded in a dense space, together with its context (left and right part of the input sentence). Independently, each entity in the KB is encoded in the same dense space together with its context e.g., entity description. Mentions are linked to entities in the dense space using a nearest neighbour search. To improve accuracy, candidate entities are ranked by passing each concatenated mention, its context and entity description to a more expensive cross-encoder.

GENRE (De Cao et al., 2021) is a BART based model fine-tuned using a sequence to sequence objective, which claims to outperform BLINK. It is an autoregressive end-to-end entity linker, it detects and retrieves mentions and the respective entities in a KB by generating their unique textual name – left to right, token-by-token. To do so, it uses a constrained decoding strategy that forces the generated name to be in a predefined candidate set. Compared to multi-class classification models such as BLINK, GENRE has a lower memory footprint to store dense vectors for large KBs, scaling linearly with vocabulary size, not entity count, and does not need to subsample negative data during training.

Zero-shot entity classification. Entity classification consists in predicting a probability for each semantic type of an entity mention, given a set of types (e.g, organisation, organic compound). The most straightforward feature used to generalise to unseen types is the textual descriptions. For example, SMXM (Aly et al., 2021) uses a cross-attention encoder to generate a vector representation for each type description and token in the input sentence and recognizes as entity types those representations that are closer to each other, including rarer classes unseen in training. It is evaluated using zero-shot adaptations of *OntoNotes* (Pradhan et al., 2013) and the domain specific biomedical dataset *MedMentions* (Mohan and Li, 2019), it also considers *out-of-KB* predictions i.e., *nil* predictions for mentions that do not have a valid gold entity.

ReFinED (Ayoola et al., 2022) is an end-to-end entity linking model optimised to perform mention detection, fine-grained entity typing (classification), and entity disambiguation in a single pass. Similar to BLINK, ReFinED uses a bi-encoder architecture modified to encode all mentions in a document simultaneously, which improves efficiency relatively to zero-shot models such as (Wu et al., 2020) that requires a forward-pass for each mention. Mention embeddings and entity description embeddings are projected into a shared vector space to calculate their dot product as the entity score. A fast bi-encoder combined with a score for unseen entities, computed based on the scores for entity types and description, is enough for ReFinED to obtain state-of-the-art performance on entity linking and to scale the approach from Wikipedia (5.9M entities) to Wikidata (90M entities).

The analyses in (Aly et al., 2021) show that while Wikipedia descriptions work well on general entity types, they perform poorly on domain specific data, e.g. *MedMentions*. They also show the impact of using annotation guidelines for descriptions to improve the transfer of knowledge from observed to unseen entities. The adoption of this approach led to a better performance compared to using a class name itself or Wikipedia passages. In particular, description vagueness, noise and negations had a negative effect, while annotation guidelines, including explicit examples and syntactic and morphological cues, improved the performance.

Zero-shot relation classification. Textual descriptions have also been employed in the relation classification task to predict new relations that could not be observed at training time. For example, ZS-BERT (Chen and Li, 2021c) learns two functions – one to project sentences and the other to project relation descriptions into an embedding space. The objective is first to jointly minimise the distance between the embedding vectors for an input sentence and the relation description for positive entity pairs and then to classify the relation (using a softmax layer to produce a classification probability). At inference time, the prediction of unseen relation classes can be achieved through nearest neighbor search. Overall, using descriptions seems to improve existent zero-shot methods and expand their domains of application. Still, descriptions are not always good enough to get good predictions. Improving the accuracy of these approaches remains an open challenge. The better the separation between embedding of different relations, the more accurate the model predictions, however, as the number of unseen relations increases, it becomes more difficult to predict the right one (Chen and Li, 2021c).

Existent ZSL methods usually rely on external knowledge from KGs, ranging from textual information, class attributes, hierarchy, domain and range constrains and relations to logic rules. There are relatively few studies evaluating their performance for unseen relations, a comparison using different external knowledge settings for zero-shot relation classification and KG completion can be seen in (Geng et al., 2021). To the best of our knowledge, we present the first approach to automatically predict and generate entity (or relation) descriptions to improve the accuracy of entity recognition and relation classification models.

Query auto completion in information retrieval systems. in relation to our work, query auto completion is the problem where a computer extends the initial parts of user queries to a search engine to save users time and enhance search performance (Cai et al., 2016). Most query auto completion approaches are based on mining query logs (Whiting and Jose, 2014). The most related approach to our work is based on personalised LMs fine-tuned on users' historical data (Jaech and Ostendorf, 2018). The key difference between our work and the query auto completion setting is that in the context of named entity recognition, we don't have historical data to learn from. Moreover, the objective of query extension is to maximise the retrieval documents accuracy while named entity recognition looks at the descriptions that maximise the entity annotation accuracy.

6 Conclusion and future work

In this paper, we formally defined the problem of selecting descriptions to make predictions about unseen classes in the ZSL context. We empirically evaluated the sensitivity of two ZSL methods to description changes, and proposed 4 different strategies to enhance them using the implicit knowledge of pre-trained language models. We also studied in detail the efficacy of the proposed entropy-based heuristic to rank different description formulations, analyzing its correlation with the performance (in terms of Macro F1 Score) of the model. We observed a negative correlation between the proposed heuristic and Macro F1 Score on two out of four of the considered datasets (OntoNotesZS and FewRel). The same assumption however was not valid for the other datasets (MedMentionsZS and WikiZS), thus motivating the need to develop more effective heuristics in future. Finally, we described the UDEBO method, which combines the predictions obtained by the same model using different automatically generated variants of entity and relation descriptions. Our experimental results, on 4 different datasets, spanning across two different NLP tasks (Entity Classification and Relation Classification) showed how UDEBO outperforms the baselines by a significant margin and achieves new state-of-the-art results on these benchmarks under the zero-shot setting.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Rami Aly, Andreas Vlachos, and Ryan McDonald. 2021. Leveraging type descriptions for zero-shot named entity recognition and classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1516–1528, Online. Association for Computational Linguistics.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Fei Cai, Maarten De Rijke, et al. 2016. A survey of query auto completion in information retrieval. *Foundations and Trends® in Information Retrieval*, 10(4):273–363.

Chih-Yao Chen and Cheng-Te Li. 2021a. Zs-bert: Towards zero-shot relation extraction with attribute representation learning. *arXiv preprint arXiv:2104.04697*.

Chih-Yao Chen and Cheng-Te Li. 2021b. ZS-BERT: towards zero-shot relation extraction with attribute representation learning. *CoRR*, abs/2104.04697.

Chih-Yao Chen and Cheng-Te Li. 2021c. ZS-BERT: towards zero-shot relation extraction with attribute representation learning. *CoRR*, abs/2104.04697.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuxia Geng, Jiaoyan Chen, Xiang Zhuang, Zhuo Chen, Jeff Z. Pan, Juan Li, Zonggang Yuan, and Huajun Chen. 2021. Benchmarking knowledge-driven zero-shot learning. *Journal of Web Semantics*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Aaron Jaech and Mari Ostendorf. 2018. Personalized language model for query auto-completion. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 700–705, Melbourne, Australia. Association for Computational Linguistics.

Martin Josifoski Sebastian Riedel Luke Zettlemoyer Ledell Wu, Fabio Petroni. 2020. Zero-shot entity linking with dense entity retrieval. In *EMNLP*.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with {umls} concepts. In *Automated Knowledge Base Construction (AKBC)*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Stewart Whiting and Joemon M Jose. 2014. Recent and robust query auto-completion. In *Proceedings of the 23rd international conference on World wide web*, pages 971–982.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. 2019. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Appendix

A Datasets

As mentioned in Section 4.1, we evaluate our approach on four different datasets, two for EC and two for RC. For EC, we use OntoNotes (Pradhan et al., 2013) and MedMentions (Mohan and Li, 2019). OntoNotes is a dataset that comprises various genres of text (news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows). We use the version available in Huggingface [4] and adapt it to perform zero-shot as explained in (Aly et al., 2021), removing all the entities that are out of the split – i.e., each split has a unique set of entities, so all the entities labeled with entities out of that set are removed – removing sentences without any entity labelled and using the same train/test/dev splits, so the pre-trained model has not seen the entities in the test set neither. The entity descriptions used for OntoNotesZS (the zero-shot version of OntoNotes) were provided by the authors of (Aly et al., 2021).

MedMentions is a corpus of Biomedical papers annotated with mentions of UMLS entities. We apply the same preprocessing steps we used for the MedMentions dataset, with the descriptions available in the official GitHub repository of (Aly et al., 2021). [2] The version of the MedMentionsZS dataset we use is also available on Huggingface. Both of them in their zero-shot version, as proposed in (Aly et al., 2021). To convert them to the zero-shot version, we follow the following steps:

1. Get the train/test/dev splits of the datasets;
2. Collect the entities in each split;
3. Remove entities out of the split i.e., if one entity e belongs to the train split, all mentions labelled as e in the test and dev splits will be replaced with the O label.
4. Remove sentences without labels. As the previous processing step (3) may remove all the entities of one sentence, the result dataset will have a lot of empty sentences. These sentences are removed in the final dataset.

Table 4 and Table 5 report the entities for each split in the dataset and the number of entities for MedMentionsZS and OntoNotesZS, respectively.

[4] https://huggingface.co/datasets/conll2012_ontonotesv5

Split	Entity	Count
Train	O	515420
	T103	22360
	T038	25007
	T033	9824
	T062	5445
	T098	3574
	T017	12575
	T074	1165
	T082	7511
	T058	14779
	T170	5996
	T204	4922
Test	O	27433
	T031	212
	T097	360
	T007	448
	T168	321
	T022	89
Validation	O	34400
	T201	404
	T091	196
	T037	434
	T005	224
	T092	452

Table 4: Number of entities labelled in each split in MedMentionsZS.

As we can observe, both datasets are highly imbalanced, with some entities appearing 25007 times and some others only 89 in the case of MedMentionsZS, and 24163 and 65 times for OntoNotesZS. However, the most common entities are used only for training and the ones with fewer examples are used for validation and testing. As pointed in (Xian et al., 2019), real-world scenarios annotated data is likely to be available for the more common ones.

In Table 6 we report some statistics concerning the length of sentences on both MedMentionsZS and OntoNotesZS. In both datasets, there are sentences containing only 1 token and 1 entity. The maximum number of tokens also varies across datasets and splits, with a maximum of 179 for MedMentionsZS and 210 for OntoNotesZS.

For RC, we use the FewRel(Han et al., 2018) and WikiZS (Sorokin and Gurevych, 2017) datasets. FewRel is a dataset for RC compiled by collecting entity-relation triplets with sentences from Wikipedia articles, and manually filtered to ensure the data quality and class balance. We use different

Split	Entity	Count
Train	O	909142
	ORG	24163
	GPE	21938
	DATE	18791
	PERSON	22035
Test	O	11299
	FAC	149
	LOC	215
	WORK_OF_ART	169
Validation	O	36790
	NORP	1277
	LAW	65
	EVENT	179
	PRODUCT	214

Table 5: Number of entities labelled in each split in OntoNotesZS.

relations for the train and the test split to ensure the zero-shot version of the dataset. The dataset is available in the Huggingface hub. [5] We use the *train_wiki* split in Huggingface as training split for the ZS-BERT model and the *wiki_val* as test split. Table 1 shows the total number of sentences in FewRel, and the number of different relations for each split. There are 700 samples for each relation in each split, thus the number of sentences reported in Table 1 is equal to the number of relations times the number of samples for each of them (e.g. train split: $44800 = 64 * 700$). Differently from FewRel, WikiZS was constructed using the Wikidata knowledge base. The dataset contains a total of 93431 sentences, each with an entity pair and a labelled relation between them. In this case, the number of instances per relation class is not balanced and we employ our own random splits containing different distinct sets of relations for the training (83 relations), validation (15 relations) and testing (15 relations) of the ZS-BERT model. More information on the dataset is contained in Table 1.

B Additional details on the models used for generating description variations

In this section, we report additional details on the methods used to generate description variations described in Section 3.1.

Extension with pre-trained LMs. An off-the-shelf GPT-2 pre-trained model was used for generating the variations, using the checkpoint from the Huggingface Hub. [6] We used $min\_length = 80$, $max\_length = 120$, $num\_beams = 8$, $temperature = 1$ and $no\_repeat\_ngram\_size = 2$ for the generation.

Extension with a fine-tuned LM. A model based on T5 large (Raffel et al., 2020) and fine-tuned on the task of description generation and extension was used for generating the variations. As a starting point for the fine-tuning, the checkpoint from Huggingface Hub [7] was used. The Wikidata dataset, containing the name and the first few sentences of included Wikipedia articles where the model was fine-tuned on, was taken from Facebook Research's BLINK project. [8] After cleaning the data i.e., removing instances with no or too short (less than 10 words) descriptions, about 5,310,000 samples were available for training the model to perform a new sequence to sequence task using $learning\_rate = 3e-05$ and $epochs = 1$. The objective was to complete the input description, starting from a sub-string containing the first ten words of it. For the generation task, just the name of the description was used. In the latter case, we set $min\_length = 80$, $max\_length = 120$, $num\_beams = 8$, $temperature = 1$ and $no\_repeat\_ngram\_size = 2$.

Summarization. A warm-started BERT2BERT (small) model fine-tuned on the CNN/Dailymail for document summarization was used for generating the descriptions variations, using the checkpoint from the Huggingface Hub. [9] We used $min\_length = 80$, $max\_length = 512$, $num\_beams = 8$, $temperature = 1$ and $no\_repeat\_ngram\_size = 2$ for this set of experiments.

Paraphrasing. A PEGASUS model fine-tuned for paraphrasing was used for generating the description variations, using the checkpoint from the Huggingface Hub. [10] We used $min\_length =$

[5] https://huggingface.co/datasets/few_rel

[6] https://huggingface.co/gpt2

[7] https://huggingface.co/t5-large

[8] http://dl.fbaipublicfiles.com/BLINK/entity.jsonl

[9] https://huggingface.co/mrm8488/bert-small2bert-small-finetuned-cnn_daily_mail-summarization

[10] https://huggingface.co/tuner007/pegasus_paraphrase

Dataset	Split	Mean #Tokens	Max #Tokens	Min #Tokens	Mean #Entities	Max #Entities	Min #Entities
MedMentionsZS	train	26	179	1	6	78	1
	test	28	102	2	2	33	1
	validation	28	119	4	2	12	1
OntoNotesZS	train	25	210	1	3	99	1
	test	29	108	2	3	39	1
	validation	28	186	3	1	27	1

Table 6: Entity classification datasets details.

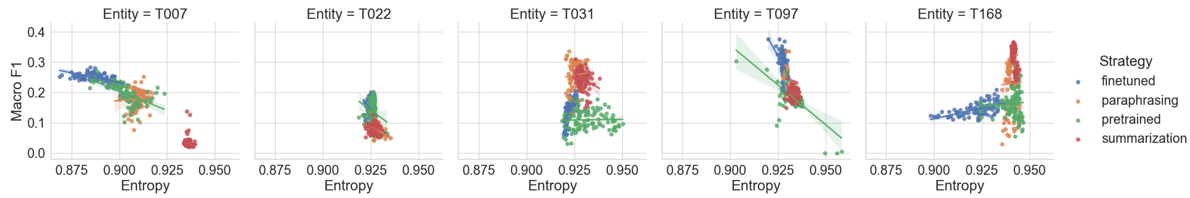

Figure 5: Analysis of the correlation between entropy and Macro F1 Score on unseen classes on the MedmentionsZS test split.

10, $max\_length = 60$, $num\_beams = 8$, $temperature = 1$ and $no\_repeat\_ngram\_size = 2$ for the generation of text.

C Additional experiments on Entropy and Macro F1 Score correlations

In this section we report additional insights on the correlation analysis discussed in the paper.

C.1 Correlations analysis of Macro F1 Score and entropy on MedmentionsZS

Figure 5 reports the correlations between Macro F1 Score and entropy on MedmentionsZS test-set. As discussed in the paper we did not observe any statistically significant correlations, with $p\text{-}value = 0.50$.

C.2 Example of generated descriptions and entropy values

Given the relation **Film Director** described as:

"director(s) of film, TV-series, stageplay, video game or similar".

The fine-tuned approach for generating variations produces the alternative description:

The director(s) of a film, TV-series, stageplay, video game or similar is the person who directs the production of the film or television series. The term "director" is also used to describe an individual or group of people who are responsible for the creation, production, and/or directing of video games, films, television shows, or other forms of media..

Although the generated description seems more complete and containing relevant additional information, the entropy calculated with ZS-BERT is higher in this case than when using the original description. This means that the model is more uncertain of its prediction.