Imperial College London

Integrating Semantic Knowledge to Tackle Zero-shot Text Classification

Jingqing Zhang*, Piyawat Lertvittayakumjorn*[1], and Yike Guo

Data Science Institute, Imperial College London, UK

Email [1] : pl1515@imperial.ac.uk

* Both authors contributed equally to this work

The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019).

Motivations

- Insufficient or even unavailable training data of emerging classes is a big challenge in real-world text classification.
- **Zero-shot text classification** – recognising text documents of classes that have never been seen in the learning stage
- In this paper, we propose **a two-phase framework together with data augmentation and feature augmentation** to solve this problem.

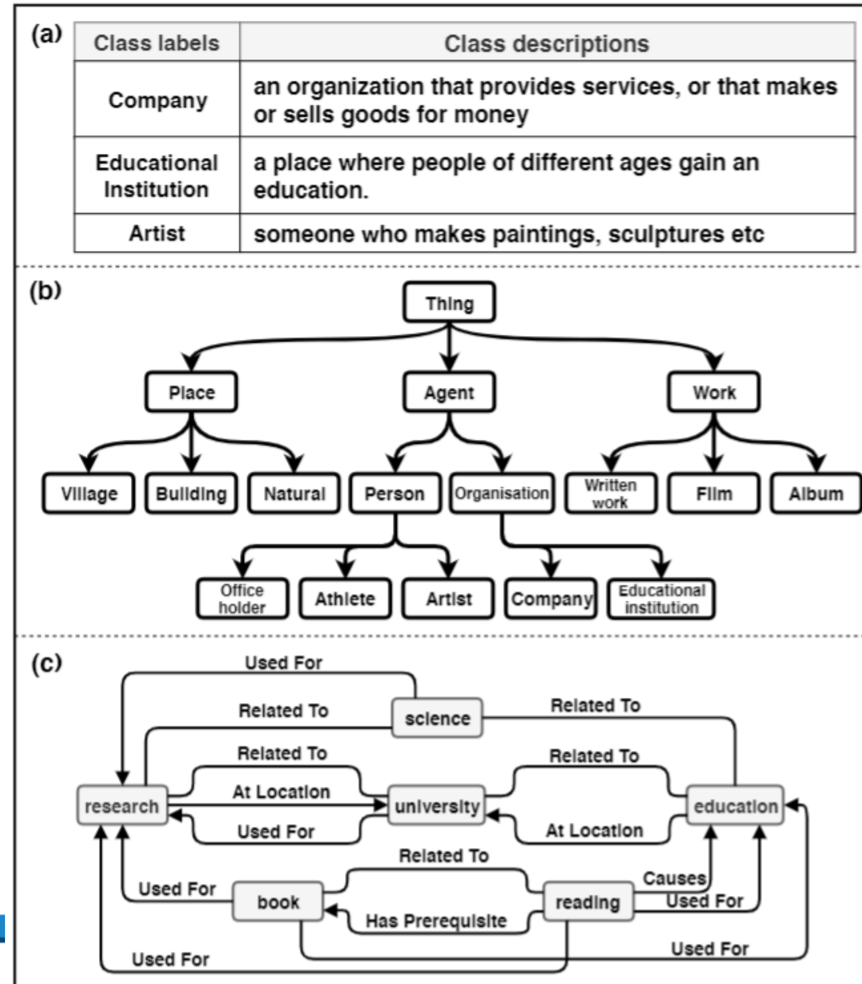
Contents

Zero-shot Text Classification

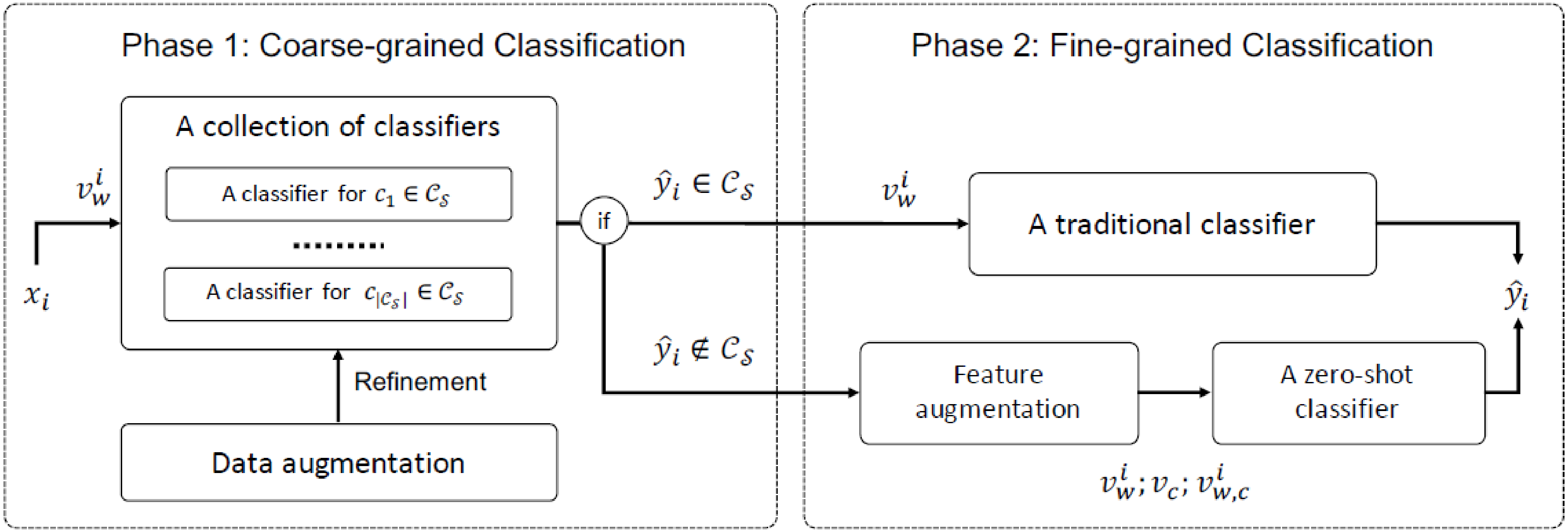
- Let C_S and C_U be disjoint sets of seen and unseen classes of the classification respectively.
- In the learning stage, a training set $\{(x_1, y_1), \dots, (x_n, y_n)\}$ is given where
  - x_i is the i^{th} document containing a sequence of words $[w_1^i, w_2^i, \dots, w_t^i]$
  - $y_i \in C_S$ is the class of x_i
- In the inference stage, the goal is to predict the class of each document, $\hat{y}_i$, in a testing set
  - y_i comes from $C_S \cup C_U$
- Supportive semantic knowledge is needed to generally infer the features of unseen classes using patterns learned from seen classes.

Our Proposed Framework: Overview

- We integrate four kinds of semantic knowledge into our framework:
  - Word embeddings
  - Class descriptions
  - Class hierarchy
  - General knowledge graph



Our Proposed Framework: Overview



- Data augmentation technique helps the classifiers be aware of the existence of unseen classes without accessing their real data.
- Feature augmentation provides additional information which relates the document and the unseen classes to generalise the zero-shot reasoning.

Phase 1: Coarse-grained Classification

- Each seen class c_s has its own CNN text classifier to predict $p(\hat{y}_i = c_s | x_i)$
  - The classifier is trained with all documents of its class in the training set as positive examples and the rest as negative examples.

- For a test document x_i, this phase computes $p(\hat{y}_i = c_s | x_i)$ for every seen class $c_s \in C_S$.
  - If there exists a class c_s such that $p(\hat{y}_i = c_s | x_i) > \tau_s$, it predicts $\hat{y}_i \in C_S$
  - Otherwise, $\hat{y}_i \notin C_S$.
  - τ_s is a classification threshold for the class c_s, calculated based on the threshold adaptation method from (Shu et al., 2017)

Phase 1: Data Augmentation

- We use the idea of **"Topic translation"** – translating an original document from a seen class into an augmented document of an unseen class.

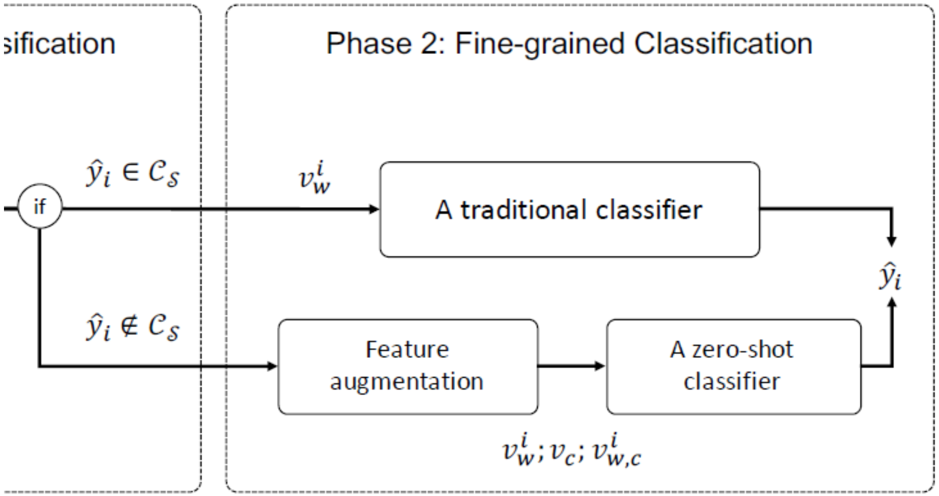
Animal → Athlete

Mitra perdulca is a species of sea snail a marine gastropod mollusk in the family Mitridae the miters or miter snails.

Mira perdulca is a swimmer of sailing sprinter an Olympian limpets gastropod in the basketball Middy the miters or miter skater.

- Using analogy questions, e.g., animal:species :: athlete:? → ? = swimmer
  - Solved by the **3CosMul** method by Levy and Goldberg (2014)

Phase 2: Fine-grained Classification

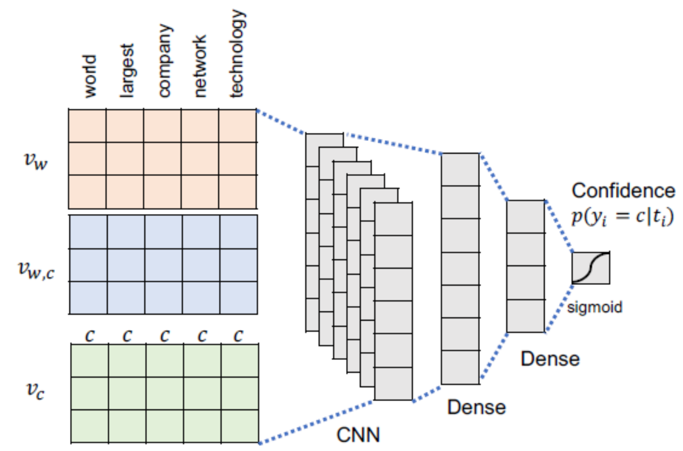


- The traditional classifier is a multi-class classifier ($|C_S|$ classes) with a softmax output, so it requires only the word embeddings v_w^i as an input.
- The zero-shot classifier is a binary classifier with a sigmoid output. It takes a text document x_i and a class c as inputs and predicts the confidence $p(\hat{y}_i = c|x_i)$.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo

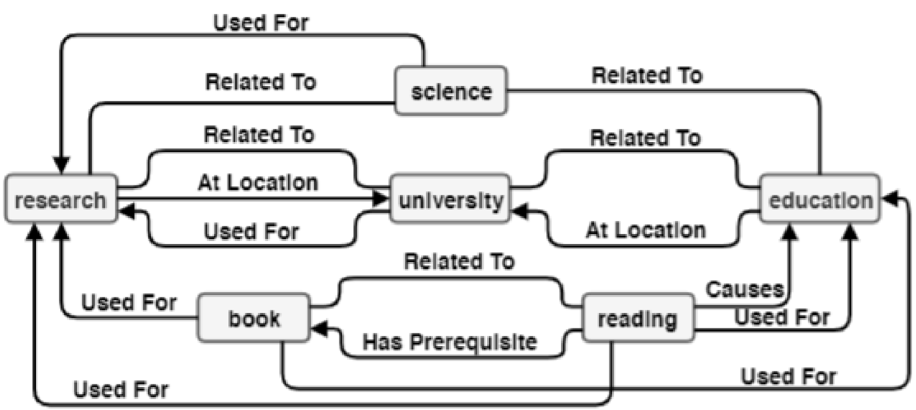
Phase 2: Zero-shot Classifier

- The zero-shot classifier predicts $p(\hat{y}_i = c|x_i)$,
  - Input features: v_w^i, v_c
  - Augmented features: $v_{w,c}^i$



- $v_{w_j,c}^i$ shows how the word w_j and the class c are related considering the relations in a general knowledge graph – ConceptNet

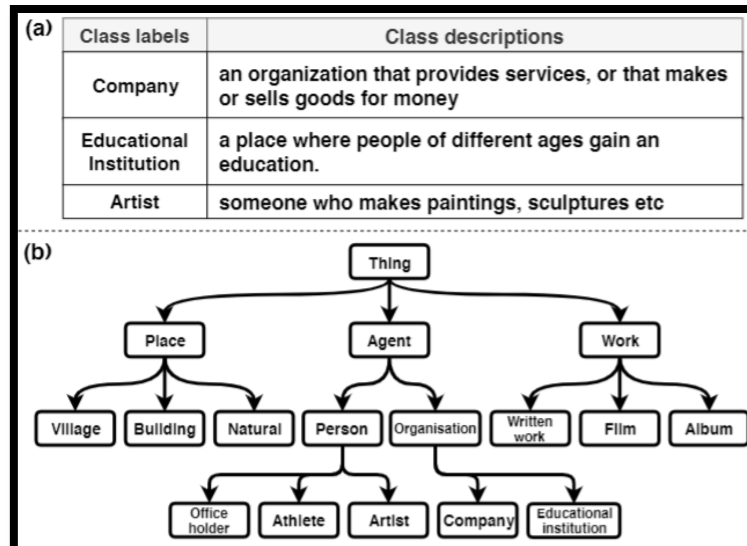
- This classifier is trained with a training data from seen classes only.



Phase 2: Feature Augmentation

- Step 1: represent a class c as three sets of nodes in ConceptNet
  - (1) ***the_class_nodes***
  - (2) ***superclass_nodes***
  - (3) ***description_nodes***

- If c is the class "Educational Institution"
  - (1) educational_institution, educational, institution
  - (2) organization, agent
  - (3) place, people, ages, education.



- Step 2: To construct $v^i_{w_j,c}$, we consider whether the word w_j is connected to the members of the three sets within K hops.

Experiments

- Datasets:
  - **DBpedia** ontology : 14 classes
  - **20newsgroups** : 20 classes

Table 1: The rates of unseen classes and the numbers of augmented documents (per unseen class) in the experiments

Dataset	Unseen rate	$\mid \mathcal{C}_S \mid$	$\mid \mathcal{C}_U \mid$	#Augmented docs per c_u
DBpedia (14 classes)	25%	11	3	12,000
	50%	7	7	8,000
20news (20 classes)	25%	15	5	4,000
	50%	10	10	3,000

An Experiment for Phase 1

Table 3: The accuracy of Phase 1 with and without augmented data compared with DOC .

Dataset Unseen rate	y_i	DOC	Ours w/o aug.	Ours w/ aug.
DBpedia 25%	seen	0.980	**0.982**	**0.982**
	unseen	0.471	0.388	**0.536**
	overall	0.871	0.855	**0.886**
DBpedia 50%	seen	0.983	0.986	**0.987**
	unseen	0.384	0.345	**0.512**
	overall	0.684	0.666	**0.749**
20news 25%	seen	0.800	**0.838**	0.831
	unseen	0.573	0.431	**0.577**
	overall	0.745	0.754	**0.770**
20news 50%	seen	0.824	**0.856**	0.843
	unseen	0.562	0.419	**0.603**
	overall	0.694	0.639	**0.724**

- Compare with DOC – a state-of-the-art open-world text classification
- For seen classes, our framework outperformed DOC on both datasets.
- The augmented data improved the accuracy of detecting documents from unseen classes clearly and led to higher overall accuracy in every setting.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo

An Experiment for Phase 2

Table 6: The accuracy of the zero-shot classifier in Phase 2 given documents from unseen classes only.

Dataset	DBpedia		20news	
Inputs \ Unseen rate	50%	25%	50%	25%
Random guess	0.143	0.333	0.100	0.200
$v_{w,c}$	0.154	0.443	0.104	0.210
$[v_c; v_{w,c}]$	0.163	0.400	0.099	0.215
$[v_w; v_{w,c}]$	0.266	0.460	0.122	0.307
$[v_w; v_c]$	0.381	0.711	0.274	0.431
$[v_w; v_c; v_{w,c}]$	**0.418**	**0.754**	**0.302**	**0.500**

- Using $[v^i_{w_j,c}]$ only could not find out the correct unseen class and neither $[v^i_{w_j}; v^i_{w_j,c}]$ and $[v_c; v^i_{w_j,c}]$ could do.
- $[v^i_{w_j}; v_c]$ increased the accuracy of predicting unseen classes clearly
- $[v^i_{w_j}; v_c; v^i_{w_j,c}]$ achieved the highest accuracy in all settings.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo

An Experiment for the Whole Framework

Table 2: The accuracy of the whole framework compared with the baselines.

Dataset	Unseen rate	y_i	Count-based	Label Similarity (Sappadla et al., 2016)	RNN Autoencoder	RNN + FC (Pushp and Srivastava, 2017)	CNN + FC	Ours
DBpedia	25%	seen	0.322	0.377	0.250	0.895	**0.985**	0.975
		unseen	0.372	**0.426**	0.267	0.046	0.204	0.402
		overall	0.334	0.386	0.254	0.713	0.818	**0.852**
	50%	seen	0.358	0.401	0.202	0.960	**0.991**	0.982
		unseen	0.304	**0.369**	0.259	0.044	0.069	0.197
		overall	0.333	0.386	0.230	0.502	0.530	**0.590**
20news	25%	seen	0.205	0.279	0.263	0.614	**0.792**	0.745
		unseen	0.201	**0.287**	0.149	0.065	0.134	0.280
		overall	0.204	0.280	0.236	0.482	**0.633**	**0.633**
	50%	seen	0.219	0.293	0.275	0.709	0.684	**0.767**
		unseen	0.196	**0.266**	0.126	0.052	0.126	0.168
		overall	0.207	0.280	0.200	0.381	0.405	**0.469**

Conclusions

- To tackle zero-shot text classification, we proposed a novel CNN-based two-phase framework together with data augmentation and feature augmentation.
- The experiments show that
  - data augmentation improved the accuracy in detecting instances from unseen classes
  - feature augmentation enabled knowledge transfer from seen to unseen classes
  - our work achieved the highest overall accuracy compared with all the baselines and recent approaches in all settings.
- Possible future works:
  - multi-label classification with a larger amount of data
  - utilise semantic units defined by linguists in the zero-shot scenario

Thank you

Q&A

Jingqing Zhang*, Piyawat Lertvittayakumjorn*[1], and Yike Guo

Data Science Institute, Imperial College London, UK

Email [1] : pl1515@imperial.ac.uk