Low-shot Learning in Natural Language Processing

Congying Xia
Department of Computer Science
University of Illinois at Chicago
Chicago, US
cxia8@uic.edu

Chenwei Zhang
Amazon
Seattle, US
cwzhang@amazon.com

Jiawei Zhang
IFM Lab, Department of Computer Science
Florida State University
Tallahassee, USA
jiawei@ifmlab.org

Tingting Liang
School of Computer Science
Hangzhou Dianzi University
Hangzhou, China
liangtt@hdu.edu.cn

Hao Peng
School of Cyber Science and Technology
Beihang University
Beijing, China
penghao@buaa.edu.cn

Philip S. Yu
Department of Computer Science
University of Illinois at Chicago
Chicago, US
psyu@uic.edu

2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI) | 978-1-7281-4144-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/COGMI50398.2020.00031

***Abstract*—This paper study the low-shot learning paradigm in Natural Language Processing (NLP), which aims to provide the ability that can adapt to new tasks or new domains with limited annotation data, like zero or few labeled examples. Specifically, Low-shot learning unifies the zero-shot and few-shot learning paradigm. Diverse low-shot learning approaches, including capsule-based networks, data-augmentation methods, and memory networks, are discussed for different NLP tasks, for example, intent detection and named entity typing. We also provide potential future directions for low-shot learning in NLP.**

***Index Terms*—Zero-shot Learning, Few-shot Learning, Natural Language Processing, Intent Detection**

I. Introduction

Natural language processing (NLP) is a subfield of Artificial Intelligence that applies to text by computers, in order to process and analyze large amounts of natural language data [1], [2], [9], [48]. Diverse tasks are followed by NLP researches, including text classification [3], [6], named entity recognition [4] and machine translation [5]. Significant success has achieved in these NLP tasks with the development of deep learning techniques [49], such as LSTM [7] and Transformers [8]. These models can achieve decent performance when they are optimized with large-scale human-labeled annotation data.

However, they are not intelligent enough to promptly adapt to new tasks or new domains, especially in this ever-changing digital world. For example, adding a new class in text classification tasks means to collect large amount of annotations for the new class and re-train the whole model. It is wasteful to ignore the previously well-trained model and make the whole process labor-intensive and time-consuming again.

Recently, researchers are interested in achieving decent performance with reduced human annotation and extending models' ability for new tasks or new domains. Low-resource learning paradigms like zero-shot learning [47] and few-shot learning [10] have drawn a lot of attention in the field of machine learning. Zero-shot learning is to adapt to new tasks or new domains without additional labeled data for these new tasks, while few-shot learning is to solve this problem with only a few labeled examples. To unify these two learning paradigms in natural language processing, we propose a new learning paradigm named as Low-shot Learning. The goal of low-shot learning is to provide a model that can adapt to new tasks with limited annotation data, including zero or few labeled examples.

In this paper, we discuss and analyze low-shot learning in natural language processing, including zero-shot learning and few-shot learning. We study different approaches, including capsule-based networks [11], data-augmentation methods [12] and memory networks [13]. Several NLP tasks, like intent detection [14] and named entity typing [15], are utilized to evaluate the performance of different low-shot learning approaches. Potential future directions are also provided for the low-shot learning in NLP.

II. Zero-shot Learning in NLP

The research on zero-shot learning in NLP is still in its infancy [23]. We mainly discuss text classification tasks for the zero-shot learning setting in NLP, including intent detection and named entity typing.

A. Zero-shot Intent Detection

Intent detection is a crucial task for intelligent assistants. It aims at identifying user intentions from their spoken language [14]. As more features and skills are being added to devices that expand their capabilities to new programs, it is common for intelligent assistants to encounter the zero-shot intent detection scenario, where no labeled utterances are available for the new intents.

Previous zero-shot learning methods for intent detection utilize external resources such as label ontologies [17] or manually defined attributes that describe intents [18] to associate existing and emerging intents, which require extra annotation. Compatibility-based methods for zero-shot intent detection [19] assume the capability of learning a high-quality mapping from the utterance to its intent directly, so that such mapping can be further capitalized to measure the compatibility of an

978-1-7281-4144-2/20/$31.00 ©2020 IEEE
DOI 10.1109/CogMI50398.2020.00031

utterance with emerging intents. However, the diverse semantic expressions may impede the learning of such mapping.

Reference [16] is the first work that attempt to tackle the zero-shot intent detection problem with a capsule-based [11] model. A capsule houses a vector representation of a group of neurons, and the orientation of the vector encodes properties of an object (like the shape/color of a face), while the length of the vector reflects its probability of existence (how likely a face with certain properties exists). The capsule model learns a hierarchy of feature detectors via a routing-by-agreement mechanism: capsules for detecting low-level features (like nose/eyes) send their outputs to high-level capsules (such as faces) only when there is a strong agreement of their predictions to high-level capsules.

The aforementioned properties of capsule models could be quite appealing for text modeling, specifically in this case, modeling the user utterance for intent detection: low-level semantic features such as the get_action, time and city_name contribute to a more abstract intent (like get_weather) collectively. A semantic feature, which may be expressed quite differently among users, can contribute more to one intent than others. The dynamic routing-by-agreement mechanism can be used to dynamically assign a proper contribution of each semantic and aggregate them to get an intent representation.

More importantly, [16] discover the potential of zero-shot learning ability on the capsule model, which is not yet widely recognized. It makes the capsule model even more suitable for text modeling when no labeled utterances are available for emerging intents. The ability to neglect the disagreed output of low-level semantics for certain intents during routing-by-agreement encourages the learning of generalizable semantic features that can be adapted to emerging intents.

B. Zero-shot Named Entity Typing

Named entity typing (NET) is to classify the types of the named entity mentions in a given utterance [20], [21]. For example, entity "Amy" in utterance "Amy bought 300 shares of Acme Corp. in 2006." is recognized as the type of Person. However, the number of entity types are diverse and unlimited, it's imperative to develop zero-shot models for NET.

Previous zero-shot NET models only learn a simple mapping function between entity mentions and types. The representations of mentions or types are learned either from handcrafted features [22], or pre-trained word embeddings [24]. These models do not have explicit knowledge transfer from seen types to unseen types. Intuitively, we want to mimic the way how humans learn new concepts. Humans learn new concepts by comparing the similarities and differences between new concepts and old concepts stored in our memory.

Reference [25] is the first work that proposes the memory augmented zero-shot NET model (MZET) to tackle the aforementioned problems. MZET stores the representations of the seen types in the memory as the knowledge we learned from the training data. To detect the zero-shot types, MZET compares the similarities between the seen types and unseen types.

III. Few-shot Learning in NLP

Inspired by humans' ability to adapt existing knowledge to new concepts quickly with only a few examples, few-shot learning [10] has recently drawn a lot of attention. Few-shot learning approaches [26] are expected to discriminate new classes from each other with only a few examples, namely, the few shots. This setting is a challenge problem in NLP for which little attention has been paid by the research community. In this section, we discuss both few-shot learning and generalized few-shot learning setting in NLP. In addition, data augmentation methods are introduced to solve the scarce annotation problem for these low-resource settings.

A. Few-shot Learning

Few-shot learning [10] is to learn classifiers for new classes with only a few training examples per class. Recent few-shot learning approaches either learn one generalizable distance metric to separate the classes [26], [27] or optimize parameters based on the gradients computed from few-shot examples [28].

Recently, some few-shot learning studies are presented with a special focus on text classification [29]–[32]. Reference [29] develops a few-shot text classification model for multi-label text classification where there is a known structure of the label space. Reference [30] proposes Induction Networks that use dynamic routing induction method to encapsulate the abstract class representation from a few examples. Reference [31] proposes an open-world learning model to deal with the unseen classes in the product classification problem. However, they all focused on classification among the few-shot classes without the seen classes.

B. Generalized Few-shot Learning

The formulation of few-shot learning only focuses on discriminating new classes and ignores existing classes. It fails to maintain a globally consistent label space that contains both existing classes and new classes. From a practical point of view, a good text classification model is expected to detect any class, no matter if it is an existing class or a new class.

A more realistic yet challenging problem setup is considered in the few-shot scenarios for NLP, which is named as generalized few-Shot learning. Generalized few-shot learning aims to correctly classify utterances that might belong to both existing and new classes. Compared to few-shot learning that only needs to discriminate new classes, Generalized few-shot learning is a much more challenging task. Due to the lack of annotations for new classes, the model has a bias on existing classes over new classes and tends to predict the test samples as existing classes.

C. Data Augmentation Methods for few-shots

Since the bottleneck in few-shot learning is the lack of annotations, the performance can be easily improved if we can generate labeled utterances for few-shot classes. Several works [12], [33], [40] have proposed to utilize variational autoencoders (VAE) [34] to augment the training data for low-resource spoken language understanding. However, these

model's ability is limited by encoders built with simple LSTMs [7], and they can only generate utterances with simple modifications.

To generate high quality and diverse utterances for new classes, [35] propose a Conditional Text Generation with BERT (CG-BERT) to transfer expressions learned in existing classes to new classes. CG-BERT is a conditional variational autoencoder (CVAE) [36] that incorporates BERT [37] naturally with specific attention masks. It utilizes the label as the condition in CVAE and provides a tractable method to generate text by controlling the latent variable. The distribution for utterances with the same label is modeled in the latent space. During training, expressions associated with existing classes are learned through regularizing the latent representations with specific prior distributions. And through sampling from the learned latent distribution, we can generate new utterances conditioned on new labels to augment the training data.

In order to provide a flexible framework that can use pre-trained weights from different kinds of transformer-based language models, including BERT [37] and GPT [38], [35] utilizes the Unified Language model [39] as the backbone. The encoder and decoder are built with multiple transformer layers. And a hidden layer is added between the encoder and the decoder. The latent variable contained in the latent layer learns the representations for utterances in a low-dimensional space. To incorporate CVAE into the unified language model, specific attention masks are proposed for these transformer layers.

Although CG-BERT has shown the ability to generate high-quality utterances that improve the performance for generalized few-shot learning, the model structure is not designed for the low-resource text generation setting. Existing classes and few-shot classes are learned separately in the model without explicit connections. Therefore the ability to generalize to few-shot intents is implicit and limited.

Reference [41] focuses on the natural language generation for few-shot intents. Firstly, they define the intent in a way that benefits the few-shot generalization ability. When users interact with intelligent assistants, their goal is to query some information or execute a command in a certain domain [42]. For instance, the intent of the input "wake me up at 7 am" is to set an alarm. The intent consists of an action "Set" in the domain of "Alarm". These actions or domains are very likely to be shared among different intents including the few-shot ones [31]. For example, there are a lot of actions ("query", "set", "remove") can be combined with the domain of "alarm". The action "query" also exists in multiple domains like "weather", "calendar" and "movie". Therefore, [41] define the intent as a pair of two parts: a domain and an action.

Unlike CG-BERT in which the input utterance is modeled as a whole, [41] want the model to learn which parts of the utterance are related to the domain and what kind of expressions contribute to the action. A composed variationaL natural language generator (CLANG) is proposed to model the local features corresponding to the domain and the action in each utterance. CLANG is a transformer-based [8] conditional variational autoencoder (CVAE) [36] with a bi-latent component. In the bi-latent component, two independent latent variables are utilized to model the distribution of the action and the domain separately, thus improving the model flexibility. Special attention masks are designed to guide the model to focus on different parts of the utterance and learn the local features. Through decomposing the utterances for the existing intents, the model learns to express the utterances for the few-shot intents as a composition of the learned local features. The generalization ability for low-resource text generation is enhanced with this local-aware model.

Reference [43] propose a domain-independent data augmentation technique, Mixup, that linearly interpolates image inputs on the pixel-based feature space. [44] combine Mixup with CNN [45] and LSTM [7] for text applications. They only conduct mixup on the fixed word embedding level like [43] did in image classification. Reference [46] add a mixup layer over the final hidden layer of the pre-trained transformer-based model. This mixup layer is dynamic and trained together within the whole text classification model.

IV. Conclusions and Future Work

This paper studies the low-shot learning paradigm in NLP, which is aim to provide the ability that can adapt to new tasks or new domains with extremely low annotation data, like zero or few labeled examples. Specifically, Low-shot learning unifies the zero-shot learning, few-shot learning and generalized few-shot learning paradigms. Diverse low-shot learning approaches, including capsule-based networks, memory networks and data-augmentation methods are discussed for different NLP tasks, like intent detection and named entity typing.

Although this paper has explore different directions for low-shot learning in NLP, these aforementioned models lack the ability to add the knowledge in the new examples and update itself continuously. Continuous learning in the low-shot setting is a interesting problem for NLP and worth to explore.

Acknowledgment

The corresponding author is Hao Peng. This work is supported by Key Research and Development Project of Hebei Province No. 20310101D, NSFC No.62002007, and in part by NSF under grants III-1763325, III-1909323, IIS-1763365 and SaTC-1930941.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009.

[2] Christopher Manning, and Hinrich Schutze. Foundations of statistical natural language processing. MIT press, 1999.

[3] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. A Text Classification Survey: From Shallow to Deep Learning. arXiv preprint arXiv:2008.00364 (2020).

[4] Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and S. Yu Philip. Multi-grained Named Entity Recognition. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1430-1440. 2019.

[5] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.

[6] Qianren Mao, Jianxin Li, Senzhang Wang, Yuanning Zhang, Hao Peng, Min He, and Lihong Wang. Aspect-Based Sentiment Classification with Attentive Neural Turing Machines. In IJCAI, pp. 5139-5145. 2019.

[7] Sepp Hochreiter, and Jürgen Schmidhuber. Long short-term memory. Neural computation 9, no. 8 (1997): 1735-1780.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pp. 5998-6008. 2017.

[9] Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In Proceedings of the 2018 World Wide Web Conference, pp. 1063-1072. 2018.

[10] Fei-Fei Li, Rob Fergus, and Pietro Perona. One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence 28, no. 4 (2006): 594-611.

[11] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In Advances in neural information processing systems, pp. 3856-3866. 2017.

[12] Jason Wei, and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196 (2019).

[13] Sainbayar Sukhbaatar, Jason Weston, and Rob Fergus. End-to-end memory networks. In Advances in neural information processing systems, pp. 2440-2448. 2015.

[14] Jian Hu, Gang Wang, Fred Lochovsky, Jian-tao Sun, and Zheng Chen. Understanding user's query intent with wikipedia. In Proceedings of the 18th international conference on World wide web, pp. 471-480. 2009.

[15] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260-270. 2016.

[16] Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and S. Yu Philip. Zero-shot User Intent Detection via Capsule Neural Networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3090-3099. 2018.

[17] Emmanuel Ferreira, Bassam Jabaian, and Fabrice Lefevre. Zero-shot semantic parser for spoken language understanding. In Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[18] Yazdani, Majid, and James Henderson. "A model of zero-shot learning of spoken language understanding." In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 244-249. 2015.

[19] Yun-Nung Chen, Dilek Hakkani-Tür, and Xiaodong He. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6045-6049. IEEE, 2016.

[20] Nancy Chinchor, and Patricia Robinson. Appendix e: Muc-7 named entity task definition (version 3.5). In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. 1998.

[21] EF Tjong Kim Sang, and F. De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of CoNLL-2003, Edmonton, Canada, pp. 142-145. Morgan Kaufman Publishers, 2003.

[22] Yukun Ma, Erik Cambria, and Sa Gao. Label embedding for zero-shot fine-grained named entity typing. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 171-180. 2016.

[23] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3905-3914. 2019.

[24] Xiang Ren, Wenqi He, Meng Qu, Lifu Huang, Heng Ji, and Jiawei Han. Afet: Automatic fine-grained entity typing by hierarchical partial-label embedding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1369-1378. 2016.

[25] Tao Zhang, Congying Xia, Chun-Ta Lu, and Philip Yu. MZET: Memory Augmented Zero-Shot Fine-grained Named Entity Typing. arXiv preprint arXiv:2004.01267 (2020).

[26] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. Matching networks for one shot learning. In Advances in neural information processing systems, pp. 3630-3638. 2016.

[27] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In Advances in neural information processing systems, pp. 4077-4087. 2017.

[28] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1126-1135. 2017.

[29] Anthony Rios, and Ramakanth Kavuluru. Few-shot and zero-shot multi-label learning for structured label spaces. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2018, p. 3132. NIH Public Access, 2018.

[30] Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. Induction Networks for Few-Shot Text Classification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3895-3904. 2019.

[31] Hu Xu, Bing Liu, Lei Shu, and P. Yu. Open-world learning and application to product classification. In The World Wide Web Conference, pp. 3413-3419. 2019.

[32] Hoang Nguyen, Chenwei Zhang, Congying Xia, and Philip S. Yu. Dynamic Semantic Matching and Aggregation Network for Few-shot Intent Detection. arXiv preprint arXiv:2010.02481 (2020).

[33] Nikolaos Malandrakis, Minmin Shen, Anuj Goyal, Shuyang Gao, Abhishek Sethi, and Angeliki Metallinou. Controlled Text Generation for Data Augmentation in Intelligent Artificial Agents. In Proceedings of the 3rd Workshop on Neural Generation and Translation, pp. 90-98. 2019.

[34] Kingma, Diederik P., and Max Welling. Auto-Encoding Variational Bayes. stat 1050 (2014): 1.

[35] Congying Xia, Chenwei Zhang, Hoang Nguyen, Jiawei Zhang, and Philip Yu. CG-BERT: Conditional Text Generation with BERT for Generalized Few-shot Intent Detection. arXiv preprint arXiv:2004.01881 (2020).

[36] Durk P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In Advances in neural information processing systems, pp. 3581-3589. 2014.

[37] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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), pp. 4171-4186. 2019.

[38] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. (2018): 12.

[39] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In Advances in Neural Information Processing Systems, pp. 13063-13075. 2019.

[40] Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. Data augmentation for spoken language understanding via joint variational generation. In Proceedings of the AAAI conference on artificial intelligence, vol. 33, pp. 7402-7409. 2019.

[41] Congying Xia, Caiming Xiong, Philip Yu, and Richard Socher. Composed Variational Natural Language Generation for Few-shot Intents. arXiv preprint arXiv:2009.10056 (2020).

[42] Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. Towards Open Intent Discovery for Conversational Text. arXiv preprint arXiv:1904.08524 (2019).

[43] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In International Conference on Learning Representations. 2018.

[44] Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study. arXiv preprint arXiv:1905.08941 (2019).
[45] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, no. 11 (1998): 2278-2324.
[46] Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip S. Yu, and Lifang He. "Mixup-Transfomer: Dynamic Data Augmentation for NLP Tasks." arXiv preprint arXiv:2010.02394 (2020).
[47] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4582-4591. 2017.
[48] Hao Peng, Jianxin Li, Senzhang Wang, Lihong Wang, Qiran Gong, Renyu Yang, Bo Li, Philip Yu, and Lifang He. Hierarchical taxonomy-aware and attentional graph capsule RCNNs for large-scale multi-label text classification. IEEE Transactions on Knowledge and Data Engineering (2019).
[49] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning. Vol. 1. Cambridge: MIT press, 2016.