Learning with Limited Text Data

Diyi Yang[*] **Ankur P. Parikh**[†] **Colin Raffel**[◇]
[*]Georgia Tech [†]Google Research [◇]University of North Carolina, Chapel Hill

1 Introduction

Natural Language Processing (NLP) has achieved great progress in the past decade on the basis of neural models, which often make use of large amounts of labeled data to achieve state-of-the-art performance. The dependence on labeled data prevents NLP models from being applied to low-resource settings and languages because of the time, money, and expertise that is often required to label massive amounts of textual data. Consequently, the ability to learn with limited labeled data is crucial for deploying neural systems to real-world NLP applications. Recently, numerous approaches have been explored to alleviate the need for labeled data in NLP such as data augmentation and semi-supervised learning.

This tutorial aims to provide a systematic and up-to-date overview of these methods in order to help researchers and practitioners understand the landscape of approaches and the challenges associated with learning from limited labeled data, an emerging topic in the computational linguistics community. We will consider applications to a wide variety of NLP tasks (including text classification, generation, and structured prediction) and will highlight current challenges and future directions.

2 Tutorial Outline

This will be a **three-hour** tutorial devoted to the **cutting-edge** topic of *Learning with Limited Text Data*, divided into three sessions. Each session will be 40 minutes, followed by 10 minutes for Q&A and 10 minutes for a break. Each part includes an overview of the corresponding topic and widely used methods and a deep dive into a set of representative NLP work.

2.1 Data Augmentation

Data augmentation is a common technique used to artificially increase both the size (i.e. the number of datapoints) and the diversity (i.e. the deviation from the true data distribution) of a given training dataset. Small labeled training datasets often lead to overfitting, and data augmentation can help alleviate this issue by creating augmented data automatically or manually. Such techniques have been widely explored in the computer vision (CV) field, with methods like geometric/color space transformations, mixup, and random erasing. Although it is relatively challenging to augment textual data because of its complex syntactic and semantic structures, there exists a wide range of methods designed to augment text data.

Representative data augmentation methods in NLP include: *token-level augmentation* such as randomly deleting or masking tokens (Bowman et al., 2015), replacing words with synonyms or related words (Zhang et al., 2015; Kobayashi, 2018), and inserting or replacing non-important tokens with random tokens (Xie et al., 2017, 2019); *sentence-level augmentation* by paraphrasing (Roy and Grangier, 2019; Edunov et al., 2018) based on back-translation that first translates sentences into certain intermediate languages and then translates them back to generate paraphrases as intermediate languages with different vocabulary and linguistic structures like POS, syntax could introduce certain variance, round-trip translation (Xie et al., 2019; Coulombe, 2018), or generating sentences conditioned on given label; *adversarial data augmentation* that uses perturbed data to dramatically influence the model's predictions and confidence without affecting human judgements (Morris et al., 2020), such as finding neighbors in a model's hidden representations using gradients (Cheng et al., 2019) or concatenating distracting but meaningless sentences as the end of paragraphs (Jia and Liang, 2017); and *hidden-space augmentation* that manipulates the hidden representations through perturbations like adding noise or performing interpolations with other data points (Chen et al., 2020a).

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 28 - 31

We will walk audiences through the recent widely-used data augmentation methods and use example NLP applications such as back-translation for unsupervised translation to demonstrate how to utilize these representative data augmentation techniques in practice.

2.2 Semi-supervised Learning

While data augmentation can be applied in the supervised setting to produce better results when only a small labeled training dataset is available, data augmentation is also commonly used in semi-supervised learning. Semi-supervised learning provides a way to leverage unlabeled data when training a model, which can significantly improve the models when there is only limited labeled data available. This is particularly useful in the common setting where unlabeled data is cheaper and easier to obtain compared to labeled data.

In this tutorial, we will briefly discuss various semi-supervised techniques explored by recent research in NLP using example applications or tasks. We group existing semi-supervised learning methods into different categories based on how they utilize unlabeled data: *Self-training* leverages supervision that inherently exists or can be automatically generated from the dataset (McClosky et al., 2006); *multi-task training* leverages extra auxiliary tasks with labels to further utilize unlabeled data related to the task of interest; and *consistency regularization* trains a model to output the same prediction when the input is perturbed through data augmentation (Sachan et al., 2019; Xie et al., 2019; Chen et al., 2020a,b).

2.3 Limited Data Learning for Low Resourced Languages and Future Work

There are other orthogonal directions for tackling the problem of learning with limited data, such as other methods for semi-supervised learning such as self-training (He et al., 2020), generative models (Cheng et al., 2016), and co-training (Clark et al., 2018). We will briefly discuss these methods, and more specifically, we will walk through audiences on how the aforementioned techniques can be leveraged for improving performance on **low-resource languages** as a case study, including cross-lingual transfer learning which transfers models from resource-rich to resource-poor languages (Schuster et al., 2019), few/zero-shot learning (Pham et al., 2019; Abad et al., 2020) which uses only a few examples from the low-resource domain to adapt models trained in another domain.

Despite the success of learning with limited data in recent years, there are still certain challenges that need to be tackled for better learning. To this end, we will conclude our tutorial by highlighting some of these challenges, including but not limited to the data distribution shift, quantify the diversity and efficiency of augmentation, dealing with out-of-domain unlabeled data, learning data augmentation strategies that are specific to text, and discussing future directions that may help advance the field.

2.4 Breadth

While we will give pointers to dozens of relevant papers over the course of the tutorial, we plan to cover around 7-8 research papers in close detail. Only 1-2 of the "deep dive" papers will come from the presenter team.

3 Diversity Considerations

This tutorial will cover techniques and topics beyond English as an application domain. We will also cover content around how learning with limited text data can be applicable to low-resourced language, dialects, and other related tasks. Our presenter team has a diverse background from both academia (a junior female faculty from Georgia Institute of Technology, and an assistant professor from University of North Carolina, Chapel Hill) and industry (a research scientist from Google). Our presenter team will share our tutorial with a worldwide audience by promoting it on social media. We will work with ACL/NAACL D&I teams, and consult resources such as the BIG directory to diversify our audience participation. Furthermore, we will engage with NLP initiatives like Masakhane that our team has connections to.

4 Prerequisites

The prerequisite includes familiarity with basic machine learning and deep learning models, especially those typically used in modern NLP, including attention mechanisms (Bahdanau et al., 2014), the Transformer architecture (Vaswani et al., 2017), sequence-to-sequence learning (Sutskever et al., 2014), etc. Furthermore, this tutorial assumes background in basic probability, linear algebra, and calculus. We will also provide a more paced introduction to the material with additional readings.

4.1 Reading List

1. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP (Chen et al., 2021)[1];

2. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification (Chen et al., 2020a)[2];

3. Understanding Back-Translation at Scale (Edunov et al., 2018);

4. Cross-lingual Language Model Pretraining (Conneau and Lample, 2019);

5. Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank (Chau et al., 2020);

6. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP (Morris et al., 2020);

7. Self-training Improves Pre-training for Natural Language Understanding (Du et al.)

5 Tutorial Presenters

Diyi Yang is an assistant professor at the School of Interactive Computing, Georgia Tech. Her research focuses on learning with limited and noisy text data, user-centric language generation, and computational social science. Diyi has organized four workshops at NLP conferences: Widening NLP Workshops at NAACL 2018 and ACL 2019, Casual Inference workshop at EMNLP 2021, and NLG Evaluation workshop at EMNLP 2021. She also gave a tutorial at the 2020 Chinese CSCW Summer School. She has taught courses on natural language processing at Georgia Tech since 2019.

Ankur Parikh is a senior research scientist at Google NYC and adjunct assistant professor at NYU. His research interests are in natural language processing and machine learning with a recent focus on high precision text generation. Ankur received his PhD from Carnegie Mellon in 2015 and has received a best paper runner up award at EMNLP 2014 and a best paper in translational bioinformatics at ISMB 2011. He has taught natural language processing at NYU since 2017.

Colin Raffel is an assistant professor of Computer Science at the University of North Carolina, Chapel Hill. His research is focused on machine learning algorithms for learning from limited labeled data, including semi-supervised, unsupervised, and transfer learning methods. His best-known work on the topics related to this tutorial include the T5 model and the MixMatch/ReMixMatch/FixMatch series of semi-supervised learning algorithms. He gave a tutorial at the 2017 International Society for Music Information Retrieval Conference[3] and has taught machine learning courses at UNC, Columbia University, and Google's TechExchange program.

6 Ethics Statement

We do not anticipate any ethical issues related to the topics of the tutorial.

References

Alberto Abad, Peter Bell, Andrea Carmantini, and Steve Renais. 2020. Cross lingual transfer learning for zero-resource domain adaptation. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Ethan C Chau, Lucy H Lin, and Noah A Smith. 2020. Parsing with multilingual bert, a small corpus, and a small treebank. *arXiv preprint arXiv:2009.14124*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in nlp. *arXiv preprint arXiv:2106.07499*.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020a. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *ACL*.

Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. 2020b. SeqVAT: Virtual adversarial training for semi-supervised sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8801–8811, Online. Association for Computational Linguistics.

[1] Collaboration from two of our tutorial presenters.

[2] Work from one of our tutorial presenters.

[3] https://colinraffel.com/talks/ismir2017leveraging.pdf

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Claude Coulombe. 2018. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718*.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. Self-training improves pre-training for natural language understanding.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159.

John X Morris, Eli Lifland, Jin Yong Yoo, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks in natural language processing. *arXiv preprint arXiv:2005.05909*.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Aurko Roy and David Grangier. 2019. Unsupervised paraphrasing without translation. *arXiv preprint arXiv:1905.12752*.

Devendra Singh Sachan, Manzil Zaheer, and Ruslan Salakhutdinov. 2019. Revisiting lstm networks for semi-supervised text classification via mixed objective function. In *AAAI*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.