Language Scaling: Applications, Challenges and Approaches

A Lecture-style Tutorial Proposal for KDD 2021

Linjun Shou[†] Ming Gong[†] Jian Pei[‡] Xiubo Geng[†] Xingjie Zhou[†] Daxin Jiang[†]
[†]Microsoft STCA NLP Group, Beijing, China
[‡]School of Computing Science, Simon Fraser University
{lisho,migon,xigeng,xingzhou,djiang}@microsoft.com jpei@cs.sfu.ca

ABSTRACT

Language scaling aims to deploy Natural Language Processing (NLP) applications economically across many countries/regions with different languages all over the world. Although recent deep learning techniques have achieved great performance in NLP, they heavily rely on huge amounts of human labeled data. Unfortunately, most languages are resource-low, that is, they have very limited linguistic resources. Language scaling by transferring knowledge from resource rich languages, such as English, to resource low languages is invaluable to the advance of social welfare. Language scaling has been heavily invested in industry parties that want to deploy their applications/services to global markets. At the same time, scaling out NLP applications to various languages, essentially a data science problem, remains a grand challenge due to the huge differences in the morphology, syntax, and pragmatics among different languages, and thus has attracted intensive interest from researchers in machine learning, data mining, and natural language processing.

Thanks to the recent progress in deep learning, pre-training, and transfer learning, many approaches for language scaling have been proposed in the past years. It is high time we provided a comprehensive survey and tutorial to promote further research and practical applications. In this tutorial, we start with a clear problem description for language scaling and an intuitive discussion on the overall challenges. Then, we outline two major categories of approaches to language scaling, namely, model transfer and data transfer. We present a taxonomy to summarize various methods in literature. A large part of the tutorial will be organized to address various types of NLP applications. For each type, we pick up one representative application to elaborate the training/evaluation data as well as the methods. We also demonstrate how to generalize the methods to similar applications of the same type, and share with the audience our lessons and experience learned in developing the applications for Microsoft products. Finally, we discuss several important challenges in this area and future directions.

1 INTENDED AUDIENCE AND LEVEL

Our tutorial is designed to serve three categories of audience.

First, *researchers* working on cross-lingual models and algorithms will find our tutorial a systematic survey of the state-of-the-art methods as well as a stimulating discussion on the core challenges and promising directions. The tutorial can serve as a fast-track introduction course bringing them to the frontier quickly and equipping them with practical ideas and tools. More importantly, the tutorial connects research, industry best practice and applications.

Second, *general audience* in the areas of Natural Language Processing (NLP), data mining, and machine learning can get an overall picture of the frontier of language scaling and various approaches, which transfer knowledge from resource rich languages to resource low languages. Moreover, researchers in other fields who need to tackle related problems can quickly understand the available techniques that they can borrow to solve their problems.

Third, *industrial NLP practitioners* will find our tutorial a comprehensive and in-depth reference to the advanced techniques and engineering practice for language scaling. This tutorial will serve as a bridge between the research frontier and the industrial best practice.

We do not assume that the audience has any deep background knowledge in Natural Language Processing, Deep Learning, or Reinforcement Learning. We will build on only some basic concepts in these areas and use sufficient examples to explain the ideas and intuitions.

The tutorial will be delivered in lecture style. We will provide slides to attendees, with proper copyright permission granted if necessary. As the tutorial will be delivered online, we will make sure it will fit virtual online only format.

2 TUTORS

2.1 In-person Presenters

Daxin Jiang, Ph.D., Chief Scientist, Microsoft Software Technology Center Asia. Daxin Jiang has years of experience of Research and Engineering in Machine Learning, Data Mining, Natural Language Processing, and Bioinformatics. He received Ph.D. in Computer Science from the Statue University of New York at Buffalo in 2005. He has published extensively in prestigious conferences and journals, and served as a PC member of numerous conferences. He received Best Application Paper Award of SIGKDD'08 and Runner-up for Best Application Paper Award of SIGKDD'04. Daxin is leading an R&D group in Microsoft with 200 applied scientists and engineers to develop NLP algorithms, applications and

KDD '21, August 14–18, 2021, Singapore
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

platforms, which support various Microsoft products, including Bing, Cortana, Teams, Outlook, and Microsoft Cognitive Services.

Address: 5 Danling Street, Hai Dian, Beijing, China, 100080. . Email: djiang@microsoft.com. Tel: +86 (10) 5917 3321.

Jian Pei, Ph.D., Professor, School of Computing Science, Simon Fraser University. His expertise is in developing effective and efficient data analysis techniques for novel data intensive applications. He is a research leader in the general areas of data science, big data, data mining, and database systems. He is recognized as a fellow of Royal Society of Canada (RSC) (i.e., the national academy of Canada), the Canadian Academy of Engineering (CAE), ACM and IEEE. He is one of the most cited authors in data mining, database systems, and information retrieval. His research has generated remarkable impact substantially beyond academia. As a renowned professional leader, he has played important roles in many academic organizations and activities. He is the Chair of ACM SIGKDD and was the Editor-in-Chief of IEEE TKDE. He received many prestigious awards, including the 2017 ACM SIGKDD Innovation Award and the 2015 ACM SIGKDD Service Award.

Address: 8888 University Drive, Burnaby, BC Canada, V5A 1S6. Email: jpei@cs.sfu.ca. Tel: +1 (778) 782 6851. Fax: +1 (778) 782 3045.

Linjun Shou, Senior Applied Scientist Manager, Microsoft Software Technology Center Asia. He has good publications on several prestigious international conferences such as ACL, EMNLP, COLING, SIGKDD, AAAI, WSDM, etc and served as the program committees on numerous conferences. His research interests include question answering, cross lingual transfer learning, representation learning, etc. Plenty of his research has been transferred to real Microsoft products such as Bing universal question answering, query understanding, document understanding, news/tweets ranking systems. Besides, he is also actively contributing to the academic community through open sourcing projects (e.g. NeuronBlocks) and benchmarks like XGLUE, CodeXGLUE.

Address: 5 Danling Street, Hai Dian, Beijing, China, 100080. Email: lisho@microsoft.com.

Ming Gong, Ph.D., Principal Applied Scientist Manager, Microsoft Software Technology Center Asia. She received Ph.D. on Graphics and Visual Computing in 2013 from Institute of Computing Technology, Chinese Academy of Sciences. She is leading an elite team with 10+ applied scientists and engineers to develop novel NLP technologies for AI applications. Her research interests include question answering, search intelligence, multilingual/cross-modal modeling, representation learning, etc. She published 30+ papers in top conferences and journals (e.g. ACL, COLING, SIGKDD, EMNLP, WSDM, AAAI, PR, CVIU), and also served as PC members of top NLP/AI conferences. Besides, she is actively contributing to the academic community by opensourcing projects (e.g. NeuronBlocks) and benchmarks like XGLUE, CodeXGLUE. Many of the novel technologies have been transferred to Microsoft,Äôs global products and online services including Bing search, multilingual question answering services and document understanding platform.

Address: 5 Danling Street, Hai Dian, Beijing, China, 100080. Email: migon@microsoft.com.

Xiubo Geng, Ph.D., Senior Applied Scientist, Microsoft Software Technology Center Asia. She received Ph.D in Computer Science in 2011 from Institute of Computing Technology, Chinese Academy of Sciences. Her research interests include machine learning, search intelligence, question answering, multilingual modeling, reasoning, etc. She has good publications on top conferences (e.g. SIGIR, NeurIPS, WWW, EMNLP, IJCAI, NAACL, etc.), and served as a PC member of several conferences.

Address: 5 Danling Street, Hai Dian, Beijing, China, 100080. Email: xigeng@microsoft.com.

2.2 Contributors

Xinjie Zhou, Ph.D., Senior Software Engineer Lead, Microsoft Software Technology Center Asia. He received Ph.D. on natural language processing in 2017 from Peking University. His research interests include machine learning, natural language understanding, multilingual modeling, etc. He has good publications on top conferences including ACL, EMNLP, AAAI, etc.

Address: Bldg #25, 328 Xinghu Street, SIP, Suzhou, China, 215000. Email: xinjzhou@microsoft.com.

3 CORRESPONDING TUTOR

Daxin Jiang, Ph.D., Chief Scientist, Microsoft Software Technology Center Asia. Email: djiang@microsoft.com.

4 OUTLINE OF THE TUTORIAL AND LENGTH

The tutorial is planned for three hours.

1. Introduction (30 minutes)
    (A) **Motivation for language scaling**: values to the advance of social welfare; examples in Microsoft products such as Bing, Office, and Cognitive Services
    (B) **Challenges of language scaling**: (1) lack of training data; (2) maintenance cost; and (3) connection among different languages.
    (C) **Overview of NLP applications**: set up an application framework including: 1) high level: NLU (Natural Language Understanding) and NLG (Natural Language Generation); and 2) low level: classification, retrieval, single-text and pairwise text sequence labeling, text-to-text generation, text-to-structure generation.
    (D) **Overview of approaches to language scaling**: A taxonomy for language scaling approaches as show in Figure 1. We will briefly describe the major idea behind each approach and delineate the relations among the approaches.
2. Language Scaling for Sequence Labeling (45 minutes)
    (A) **Task Introduction & Dataset**: (1) Cross lingual NER (WIKIANN, CoNLL-2003; (2) Cross-lingual Slot filling (MTOP, SNIPS); (3) Cross-lingual extractive Machine reading comprehension (MLQA [11], TydiQA), etc.
    (B) **Model transfer**: (1) Cross-lingual NER & Slot filling [1, 2, 15, 24, 27]; (2) Cross-lingual MRC [4, 12, 13].
    (C) **Data augmentation**: (1) Machine Translation [7, 13, 17, 28]; (2) Task adaptation [12, 23].
3. Language Scaling for Generation (45 minutes)
    (A) **Tasks**: (1) Cross-lingual Grammar Error Correction (2) Cross-lingual Question Generation, etc.
    (B) **Model transfer**: Cross-lingual generation pretraining based approaches [3, 10, 14, 25, 26].

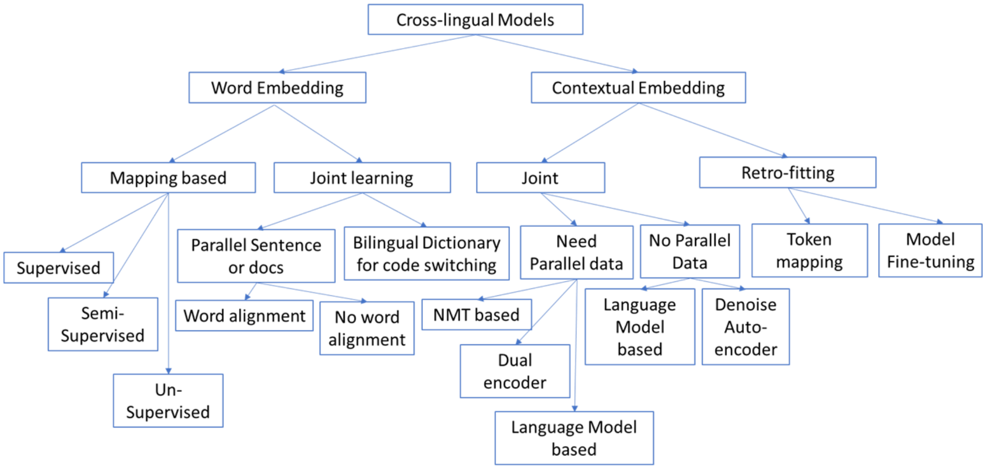

Figure 1: A taxonomy of language scaling approaches.

(C) **Data augmentation**: (1) Machine Translation [30]; (2) semi-supervised learning [8, 29].

4. Language Scaling for syntactic parsing (45 minutes)
   (A) **Tasks**: Cross-lingual dependency parsing
   (B) **Model transfer**: (1) multilingual word embeddings [9], (2) universal POS tags [16], (3) multilingual contextualized representations [22].
   (C) **Data augmentation**: (1) machine Translation [21] (2) annotation projection [19]
5. Summary: Challenges and Future Directions: (15 minutes)
   (A) **Challenges**: (1) crowd-sourcing in small languages (throughput and quality control); (2) trade-off for Return-Over-Investment; (3) language-specific features *vs.* language-agnostic features
   (B) **Future directions**: (1) Common representation of languages (syntax and semantics) by large pre-trained models; (2) Zero-shot and few-shot learning by data generation, transfer learning, and active learning; (3) Automatic language scaling for various NLP tasks with different availability of data and constraints of compliance.

5 RELATED TUTORIALS

A half-day tutorial titled "Scaling out NLP Applications to 100+ Languages" has been accepted as a Lecture-style tutorial for The Web Conference 2021, April 19–23, 2021, Ljubljana, Slovenia. Due to the very broad scope of NLP applications and the large number of related works, we only selected several NLP tasks (e.g., Natural Language Inference (NLI), Information retrieval (IR), and Machine Reading Comprehension (MRC)) in that three-hour tutorial.

In this tutorial, we follow the similar application framework and the approach taxonomy as proposed in the WWW tutorial, but discuss many different tasks, including Topic Modeling (TM), Machine Reading Comprehension (MRC), Syntactic and Semantic Parsing. These tasks not only have wide applications in practice, but represent different task types. For example, compared with Natural Language Inference, Topic modeling target at a much larger label set. Different from NER which accept single text input, MRC conducts sequence labeling on one text (e.g., the document) with respect to another text (e.g., the query). Syntactic and Semantic Parsing, which are both fundamental tasks in NLP, were not covered in the WWW tutorial either. The overlap between the two tutorials is about 3 5%. Due to the difference in the nature of tasks, the approaches to language scaling for the new tasks are very different from those in WWW tutorial. Therefore, we believe the addition of new tasks will be a substantial complement to the previous tutorial not only in application types, but also from the methodology perspective.

There are a small number of related tutorials by other authors. Ruder *et al.* [18] provided a tutorial on unsupervised cross-lingual representation in ACL'19. The tutorial mainly focuses on weakly-supervised and unsupervised cross lingual word embedding in low resource settings where bilingual supervision may not be available. Various approaches, training conditions, robustness for distant language pairs, and applications are discussed. Our tutorial conducts a comprehensive survey on various approaches to language scaling for NLP applications, where cross-lingual word embedding is one of the approaches discussed in Section 1.D. We will not cover details but refer the audience to Ruder et al.'s tutorial.

Other related tutorials [5, 6, 20] target at one specific cross-lingual task each, such as machine translation, entity linking, or cross-lingual parallel data mining. Our tutorial covers a broad range of NLP applications. A unique feature is that, to connect research and industry best practice, we will use as examples the research, techniques, and engineering in Microsoft products and services that need to be scaled out to 100+ languages, including word breaking, spelling correction, information extraction, search relevance, question answering, language understanding, etc.

6 TUTORIAL EXPERIENCE

We have rich and successful experience in delivering well accepted tutorials in premier conferences. For example, Pei gave 27 tutorials in conferences such as KDD (12 times) [1], WWW and SIGIR.

[1] https://www.youtube.com/playlist?list=PL8n-erTbIhTMvdXs657kBOp2pXFJVyAnB

He also presented many keynote speeches in some conferences and workshops. Jiang gave tutorials at SIGIR (twice), KDD and WWW. He also delivered keynote speeches and invited talks at many conferences and workshops.

7 TUTORIAL STRATEGIES

In order to attract and encourage audience participation and interactivity throughout the tutorial presentation, we plan to interleave technical presentations and demos so that the audience can connect the technical materials with the real applications. Moreover, we plan to bring to the tutorials real data sets and encourage the audience to explore the data sets during sections.

8 EQUIPMENT

Since this will be online tutorial, we will coordinate with organizers to prepare the equipment for presenters. There is no requirement to the equipment for attendees other than a device to access the online tutorial.

REFERENCES

[1] Zuyi Bao, Rui Huang, C. Li, and Kenny Zhu. 2019. Low-Resource Sequence Labeling via Unsupervised Multilingual Contextualized Representations. *ArXiv* abs/1910.10893 (2019).

[2] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-Source Cross-Lingual Model Transfer: Learning What to Share. In *ACL*.

[3] Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and He yan Huang. 2020. Cross-Lingual Natural Language Generation via Pre-Training. *ArXiv* abs/1909.10481 (2020).

[4] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-Lingual Machine Reading Comprehension. In *EMNLP-IJCNLP*. 1586–1595.

[5] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. Multilingual Neural Machine Translation. In *COLING*.

[6] Ahmed El-Kishky, Philipp Koehn, and Holger Schwenk. 2020. Mining the Web for Cross-lingual Parallel Data. In *SigIR*.

[7] Y. Fang, S. Wang, Z. Gan, S. Sun, and JJ. Liu. 2020. FILTER: An Enhanced Fusion Method for Cross-lingual Language Understanding. *ArXiv* abs/2009.05166 (2020).

[8] Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. 2019. Neural Grammatical Error Correction Systems with Unsupervised Pre-training on Synthetic Data. In *BEA@ACL*.

[9] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1234–1244.

[10] Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction. *ArXiv* abs/2005.00987 (2020).

[11] Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475* (2019).

[12] Shining Liang, Linjun Shou, Jian Pei, Ming Gong, WanLi Zuo, and Daxin Jiang. 2020. CalibreNet: Calibration Networks for Multilingual Sequence Labeling. *ArXiv* abs/2011.05723 (2020).

[13] Junhao Liu, Linjun Shou, Jian Pei, Ming Gong, Min Yang, and Daxin Jiang. 2020. Cross-lingual Machine Reading Comprehension with Language Branch Knowledge Distillation. *ArXiv* abs/2010.14271 (2020).

[14] Yinhan Liu and Jiatao Gu et al. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *ArXiv* abs/2001.08210 (2020).

[15] Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-Informed Mixed-Language Training for Zero-shot Cross-lingual Task-oriented Dialogue Systems. *ArXiv* abs/1911.09273 (2020).

[16] Ryan McDonald, Slav Petrov, and Keith B Hall. 2011. Multi-source transfer of delexicalized dependency parsers. (2011).

[17] L. Qin, Minheng Ni, Y. Zhang, and W. Che. 2020. CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP. In *IJCAI*.

[18] Sebastian Ruder, Anders Søgaard, and I. Vulic. 2019. Unsupervised Cross-Lingual Representation Learning. In *ACL*.

[19] Michael Sejr Schlichtkrull and Anders Søgaard. 2017. Cross-lingual dependency parsing with late decoding for truly low-resource languages. *arXiv preprint arXiv:1701.01623* (2017).

[20] A. Sil, Heng Ji, D. Roth, and S. Cucerzan. 2018. Multi-lingual Entity Discovery and Linking. In *ACL*.

[21] Jörg Tiedemann and Zeljko Agić. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research* 55 (2016), 209–248.

[22] Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. *arXiv preprint arXiv:1909.06775* (2019).

[23] Q. Wu, Zijia Lin, Börje F. Karlsson, B. Huang, and Jianguang Lou. 2020. UniTrans : Unifying Model Transfer and Data Transfer for Cross-Lingual Named Entity Recognition with Unlabeled Data. In *IJCAI*.

[24] Q. Wu, Zijia Lin, Börje F. Karlsson, Jian-Guang Lou, and B. Huang. 2020. Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language. *ArXiv* abs/2004.12440 (2020).

[25] Linting Xue and Noah Constant et al. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *ArXiv* abs/2010.11934 (2020).

[26] Ikumi Yamashita, Satoru Katsumata, Masahiro Kaneko, Aizhan Imankulova, and Mamoru Komachi. 2020. Cross-lingual Transfer Learning for Grammatical Error Correction. In *COLING*.

[27] Z. Yang, R. Salakhutdinov, and William W. Cohen. 2016. Multi-Task Cross-Lingual Sequence Tagging from Scratch. *ArXiv* abs/1603.06270 (2016).

[28] Fei Yuan, Linjun Shou, X. Bai, Ming Gong, Yaobo Liang, N. Duan, Y. Fu, and Daxin Jiang. 2020. Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension. In *ACL*.

[29] Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data. In *NAACL-HLT*.

[30] Wangchunshu Zhou, Tao Ge, C. Mu, Ke Xu, Furu Wei, and M. Zhou. 2020. Improving Grammatical Error Correction with Machine Translation Pairs. *ArXiv* abs/1911.02825 (2020).