# Multi-Domain Neural Machine Translation with Knowledge Distillation for Low Resource Languages

•••

Supervisor : Dr. Nisansa de Silva
Co-Supervisor : Dr. Surangika Ranathunga

# Introduction

- Neural Machine Translation (NMT)[1] has overtaken Statistical Machine Translation (SMT)[10] in recent years due to the former performing on par with human translators for some high-resource language pairs [2],[3].
- Despite NMT's success in Machine Translation (MT), they still tend perform poorly on low resource and domain adaptation scenarios [11].
- Even though NMTs perform well in general translation problem, there has been a rise in demand for domain specific MT systems [14].
- A Domain has its own unique vocabulary as well as shared vocabulary,
  Eg:

  "benzodiazepine" which is unique to the biomedical domain.
  "Conductor" may indicate different context in Engineering and Transportation.
  Even verbs like "administer" has different meanings in medical domain and goverment/political domain.

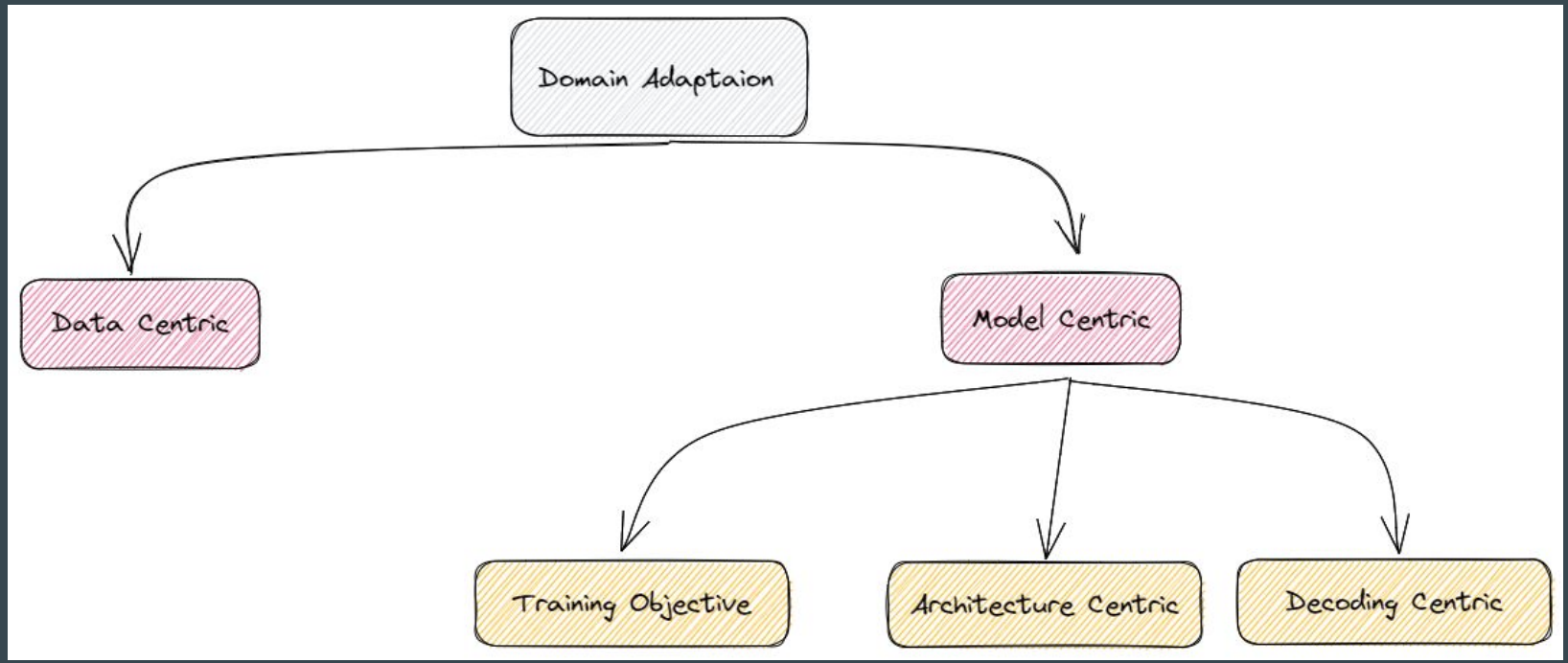
[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
[2] Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. Findings of the 2019 conference on machine translation (wmt19). ACL, 2019.
[3] Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares,et al. Findings of the wmt 2019 biomedical translation shared task: Evaluation for medline abstracts and biomedical terminologies. In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pages 29–53, 201
[10] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pages 177–180, 2007.
[11] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. InProceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics
[14] Danielle Saunders. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. Journal of Artificial Intelligence Research, 75:351–424, 2022.

# This leads to Domain Adaptation

- So what is a domain?
  We adopt the definition mentioned in [14], which characterizes a domain by 3 attributes
  - ***Provenance*** is the source of the text.
  - ***Topic*** holds the subject of the text.
  - ***Genre*** stands orthogonal to topic, consisting of function, register, syntax and style.
- Domain Adaptation is the problem of improving performance of a model trained on general domain data and test instances from a new domain [7].
- Domain Adaptation in MT is thoroughly researched field.

[7] Praveen Dakwale and Christof Monz. Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data. In Proceedings of Machine Translation Summit XVI: Research Track, pages 156–169, 201
[14] Danielle Saunders. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. Journal of Artificial Intelligence Research, 75:351–424, 2022.

# [17] Categorize Domain adaptation in the following manner




[17] Chenhui Chu and Rui Wang. A survey of domain adaptation for machine translation. Journal of information processing, 28:413–426, 2020.

# Existing Line of Thoughts

- Most of the promising results rely on either ensembling, a priori domain clustering in order to add domain tags and introducing a new domain specific gating vector [12].
- The best technique that does not add more complexity or prior classification, either using supervised or unsupervised methods, is based on fine-tuning on the concatenation of all in-domain data.

[12] Idriss Mghabbar and Pirashanth Ratnamogan. Building a multi-domain neural machine translation model using knowledge distillation. arXiv preprint arXiv:2004.07324, 2020.

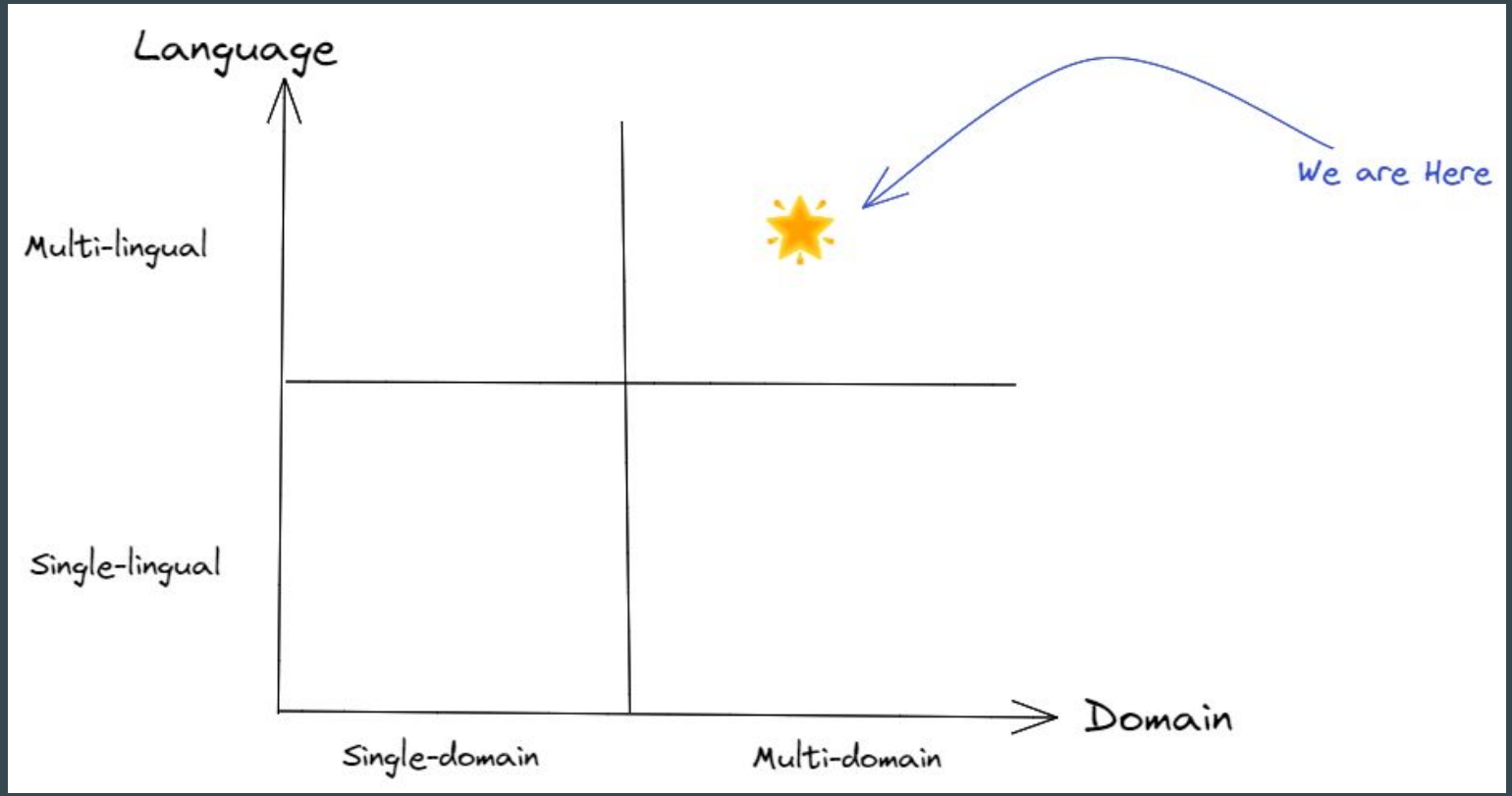
# Pain Points in Domain Adaptation

- Although Domain adaptation for NMT is a thoroughly researched field, but it is not a finished field (meaning there is not a single solution without tradeoffs)
- Having dedicated models for different domain and different languages will provide a scalability issue.
- As new domain gets added, we may have to train models from scratch, which requires data storages of previous domains and will incur high computational cost during training.
- Most real world datasets (web crawled datasets) may not be one hundred percent pure, there will be presence of other domains as well.

# Research Problem

"How to design a **Multi-Domain** NMT system for **Low Resource Languages** which has the ability to scale well towards new domains while having low/no degradation of  in performance on the existing domains"

# Problem Setting



# Fine Tuning (Continued Learning)

- Fine tuning has proven effectiveness in transferring between similar tasks [13], [15], [16].
- We can perform domain adaptation by fine tuning a model trained on large out-of-domain data with in-domain data.
- Two key issues fine tuning adheres to are,
  a. **Over-fitting** when the in-domain data is small.
  b. **Catastrophic forgetting** happens when out-of-domain translation is degraded.
- Authors of [5] show a simple yet effective way of overcoming the above mentioned issues by “domain mixing” (here the out-of-domain data is mixed with in-domain data during fine tuning).

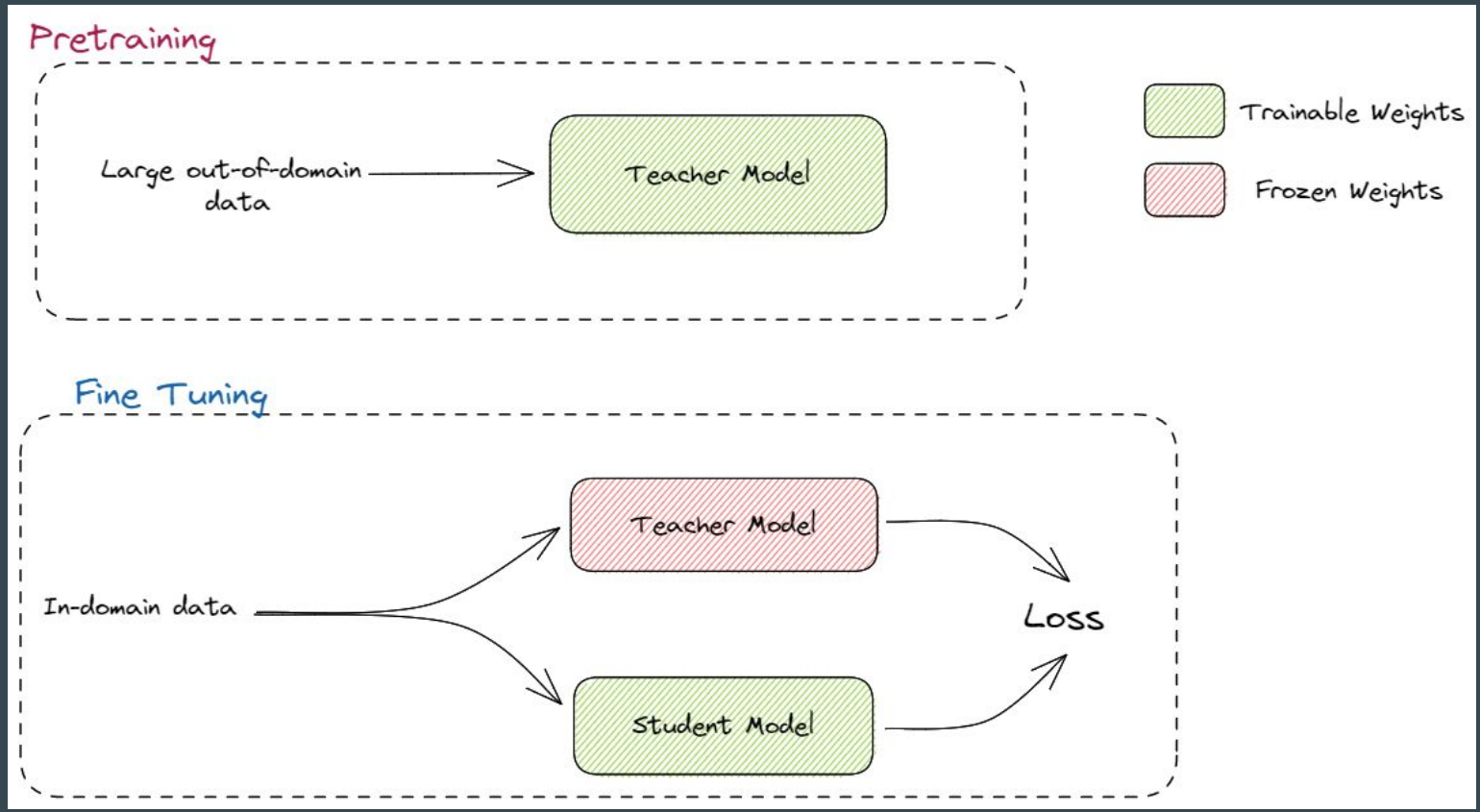
[5] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 385–391, 2017.
[13] Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. Question answering through transfer learning from large fine-grained supervision data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 510–517, 2017.
[15] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709, 2015.
[16] Aliaksei Severyn and Alessandro Moschitti. Unitn: Training deep convolutional neural network for twitter sentiment classification. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pages 464–469, 2015.

# Knowledge distillation

- Knowledge distillation (KD) was initially introduced by [9] as a model compression method.
- But KD can be utilized for Multi-task learning use cases [12]. This allows KD to used for fine tuning NMT systems to adapt to new domains.
- In [6] a single model was adapted to perform successfully on multiple domains.
- In KD the divergence of the distribution of the teacher model and the student model is minimized, and there by transferring the knowledge of the teacher to the student.
- BUT, trying to align the global statistics between the source and the target domain has a disadvantage in cases where the domains or tasks are inherently divergent, this will lead to smoothing. As result the expected performance of the model will be sub-optimal [8].

[6] Anna Currey, Prashant Mathur, and Georgiana Dinu. Distilling multiple domains for neural machine translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4500–4511, 2020.
[8] Zi-Yi Dou, Xinyi Wang, Junjie Hu, and Graham Neubig. Domain differential adaptation for neural machine translation. arXiv preprint arXiv:1910.02555, 2019.
[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.
[12] Idriss Mghabbar and Pirashanth Ratnamogan. Building a multi-domain neural machine translation model using knowledge distillation. arXiv preprint arXiv:2004.07324, 2020.

# Typical use of KD in Domain adaptation for fine tuning



# Knowledge Distillation + Curriculum learning

- To address the above issue, our domain adaptation methodology will be governed by Curriculum learning [4] paradigm.
- This method of learning mimics how we humans learn, we learn task by first experiencing the easy version of the task and gradually increasing complexity of the task.
- Similarly in domain adaptation, we will make the model experience harder sentence pairs gradually, there by performing better than standard data random shuffling.
- To the best of knowledge no has combined KD with CL for Domain Adaptation in NMT

[4] Yoshua Bengio, J ́erˆome Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48, 2009.

# Research Objectives

- Design and Implement a NMT system which is capable of performing in multi-lingual and multi-domain scenarios using knowledge distillation governed by curriculum learning.
- Quantify divergence between domains, and investigate the effects of domain divergence on domain adaptation by knowledge distillation.
- Investigate the performance of curriculum training for domain adaptation by knowledge distillation and report it's impact on the performance of the proposed method.

# Research Methodology outline

Our proposed approach for domain adaptation in machine translation involves utilizing KD governed by curriculum learning, with the goal of minimizing the effects of domain divergence. By employing a training curriculum, we aim to ensure that the resulting model is effective in adapting to various domains and achieving high-quality translations. An outline of our proposed approach is as follows,

- Perform a thorough investigation and select the best large language model (LLM) for our task.
- Identify the best sentence selection method, and form a training curriculum.
- Adapt our model to multiple domains by KD using the training curriculum prepared as stated above.
- Study and quantify the domain divergence between any given two domains and investigate whether adopted training curriculum fits our task.

# Resources

- **Models:** Initially models will be taken from huggingface platform.
- **Data:** We will utilize the data we are preparing for the Google-funded project and other benchmark datasets.
- **Computational Resources:** GPU cluster available in CSE department, UOM.

# References

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

[2] Lo ̈ic Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. Findings of the 2019 conference on machine translation (wmt19). ACL, 2019.

[3] Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares,et al. Findings of the wmt 2019 biomedical translation shared task: Evaluation for medline abstracts and biomedical terminologies. In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pages 29–53, 201

4] Yoshua Bengio, J´erôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48, 2009.

[5] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 385–391, 2017.

[6] Anna Currey, Prashant Mathur, and Georgiana Dinu. Distilling multiple domains for neural machine translation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4500–4511, 2020.

[7] Praveen Dakwale and Christof Monz. Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data. In Proceedings of Machine Translation Summit XVI: Research Track, pages 156–169, 201

[8] Zi-Yi Dou, Xinyi Wang, Junjie Hu, and Graham Neubig. Domain differential adaptation for neural machine translation. arXiv preprint arXiv:1910.02555, 2019.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[10] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pages 177–180, 2007.

[11] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. InProceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics.

[12] Idriss Mghabbar and Pirashanth Ratnamogan. Building a multi-domain neural machine translation model using knowledge distillation. arXiv preprint arXiv:2004.07324, 2020.

[13] Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. Question answering through transfer learning from large fine-grained supervision data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 510–517, 2017.

[14] Danielle Saunders. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. Journal of Artificial Intelligence Research, 75:351–424, 2022.

[15] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709, 2015.

[16] Aliaksei Severyn and Alessandro Moschitti. Unitn: Training deep convolutional neural network for twitter sentiment classification. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pages 464–469, 2015.

[17] Chenhui Chu and Rui Wang. A survey of domain adaptation for machine translation. Journal of information processing, 28:413–426, 2020.

# Thank you!