Exploring Domain Adaptation Techniques for Abstractive Text Summarization

Aditi Arora, 15.05.2023, Guided Research Final presentation

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

- **Background**
  - Automatic Text Summarization
  - Domain Adaptation

- **Research Questions**

- **Literature Survey**

- **Results:**
  - Proposed Taxonomy of domain adaptation techniques used.

- **Future Work**

Background: Automatic Text Summarization

- Automatic text summarization is a subfield of NLP which aims at automatically shortening the content of a textual information source in a way that retains its most important information. Some of the key features of Automatic Text summarization are:

Produces a concise and fluent summary

Preserves the meaning of the original text document

Does not contain repetition of the same information

Does not require domain expertise

Awasthi, Ishitva, et al. "Natural language processing (NLP) based text summarization-a survey." *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. IEEE, 2021.

Background: Why Abstractive Text Summarization?

TUM

Abstractive Text Summarization

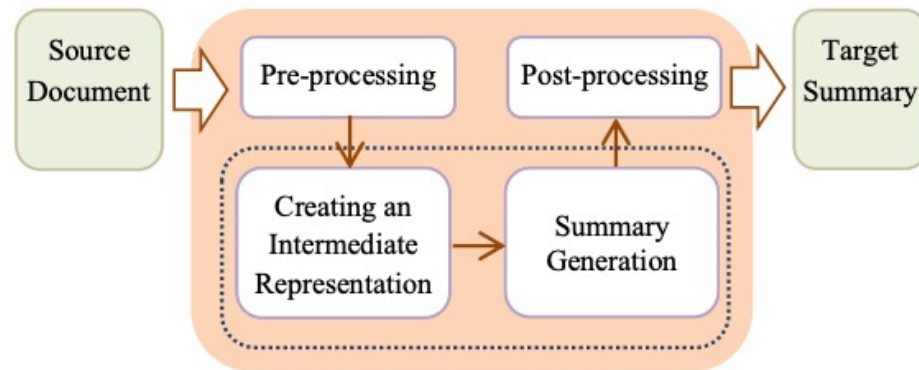Uses advanced NLP to understand the semantics of the text and generate a meaningful summary.

Image: Abstractive Text Summarization Architecture (Wafaa et al.)

Extractive Text Summarization

Selects the most important sentences within a text (without necessarily understanding the meaning), therefore the result summary is just a subset of the full text. There are certain disadvantages associated with extractive text summarization.**
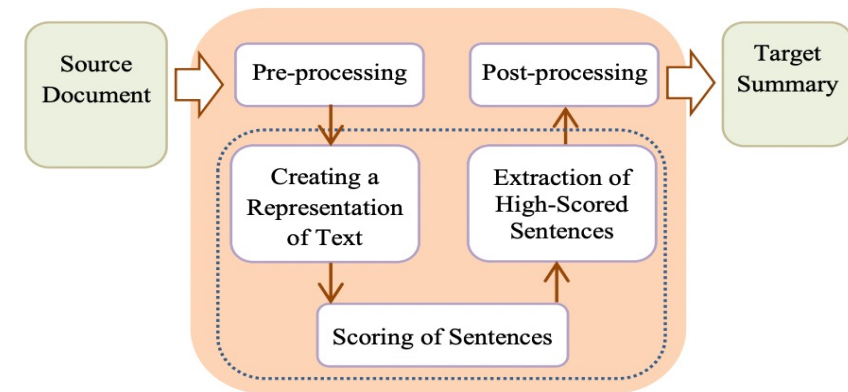

Image: Abstractive Text Summarization Architecture (Wafaa et al.)

Background: Why do we need domain adaptation in NLP?

TUM

Ideal Situation

Training and Test set follows the same underlying data distribution.

Real Situation

In real life, the training and test set do not follow the same distribution.

Most of the large language models are trained on generic datasets.

Performance drop

Due to difference in underlying data distribution, a drop in performance is observed.

To combat this issue, domain adaptation techniques are adopted.

Tank, Meenaxi, and Priyank Thakkar. "Text Summarization Approaches Under Transfer Learning and Domain Adaptation Settings—A Survey." *Computational Intelligence and Data Analytics: Proceedings of ICCIDA 2022*. Singapore: Springer Nature Singapore, 2022. 73-88.

Background: Domain Adaptation

- Text data can be classified into different domains depending on the **context**.

- One of the key features of domain specific text data is the presence of domain specific **cue terms** and **phrases**. These words might have different meaning in different domains.

- For this study, we are focusing on a particular domain.
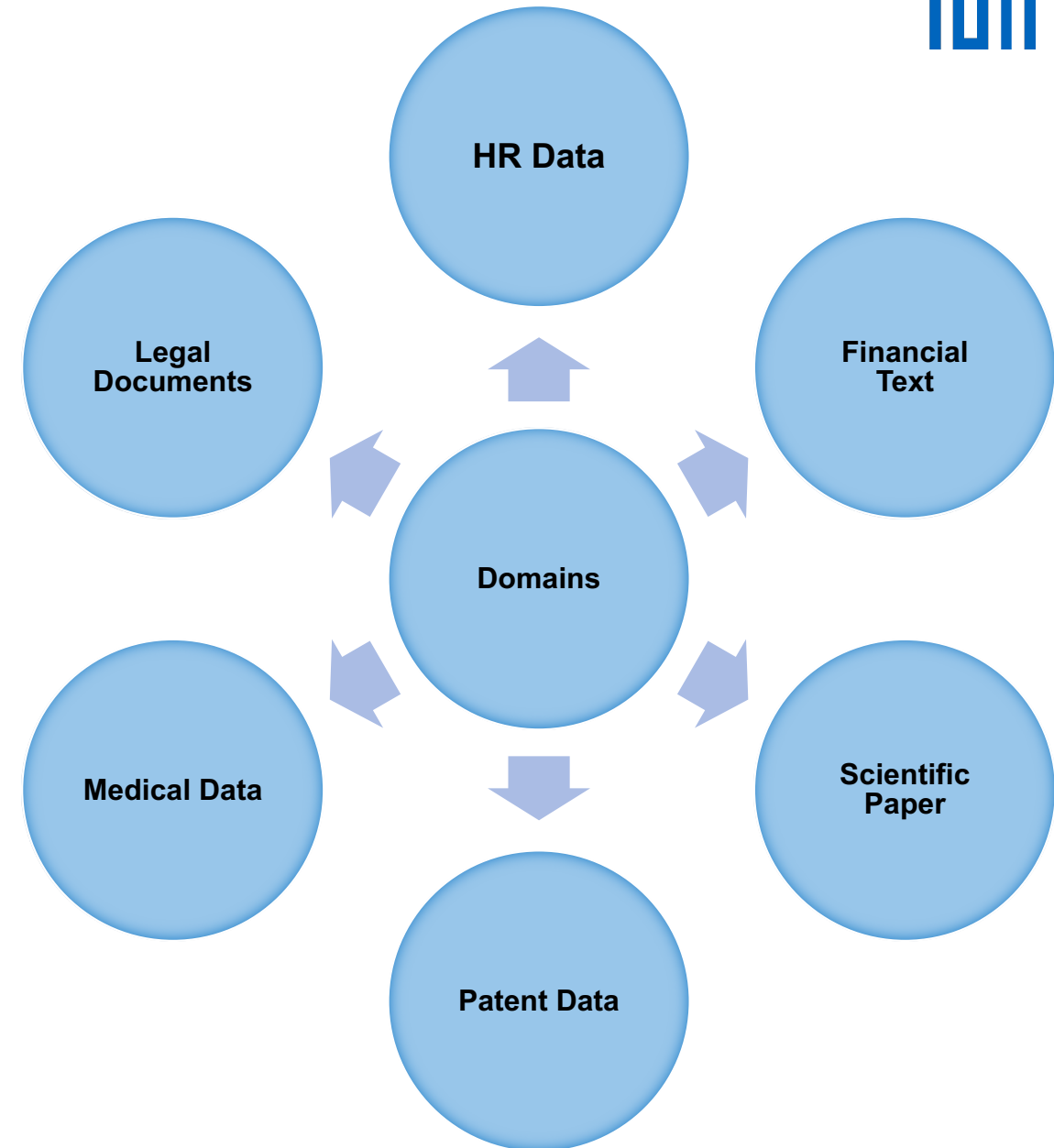
Background: Why do we need domain adaptation techniques in text summarization?

Training a model to generate meaningful and coherent summaries requires large amount of labelled data.

Acquiring large amount of labelled data for different domains is expensive .

Lack of labelled data in new domains affects the quality of generated summaries.

Domain adaptation techniques allow a summarization system to learn content selection from out-of-domain data while **acquiring language generating behaviour** with in-domain(target) data.

Background: Domain Adaptation vs Transfer Learning?

Domain Adaptation

- DA is a particular case of transfer learning, namely transductive transfer learning.

- In transductive DA, the source and target tasks T_S and T_T remain the same, but the source and target domains D_S and D_T differ in their underlying probability distributions.

- DA typically addresses the shift in marginal distribution $P_S(X) != P_T(X)$, also known as covariate shift.

Ramponi A, Plank B. Neural unsupervised domain adaptation in NLP---a survey. arXiv preprint arXiv:2006.00632. 2020 May 31. Ramponi A, Plank B. Neural unsupervised domain adaptation in NLP---a survey. arXiv preprint arXiv:2006.00632. 2020 May 31.

Classification of transfer learning
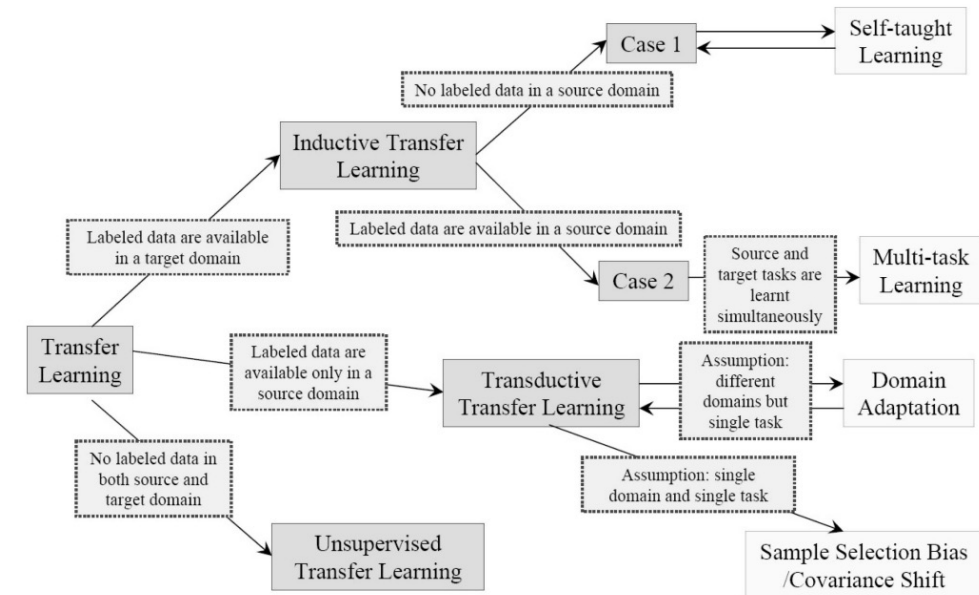
Ramponi A, Plank B. Neural unsupervised domain adaptation in NLP---a survey. arXiv preprint arXiv:2006.00632. 2020 May 31.

Research Questions

Question 1:
What are the different types of models being used for domain adaptation in abstractive text summarization?

Question 2:
What are the different types of domain adaptation techniques used for abstractive text summarization?

Question 3:
What is the criteria for classifying the various domain adaptation techniques found in the literature?
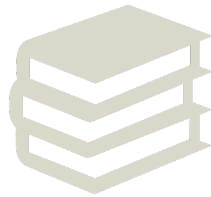
Literature Survey

Keywords Used:
Domain Adaptation, Limited/Low Resource Setting + Abstractive text summarization

Search Engine:
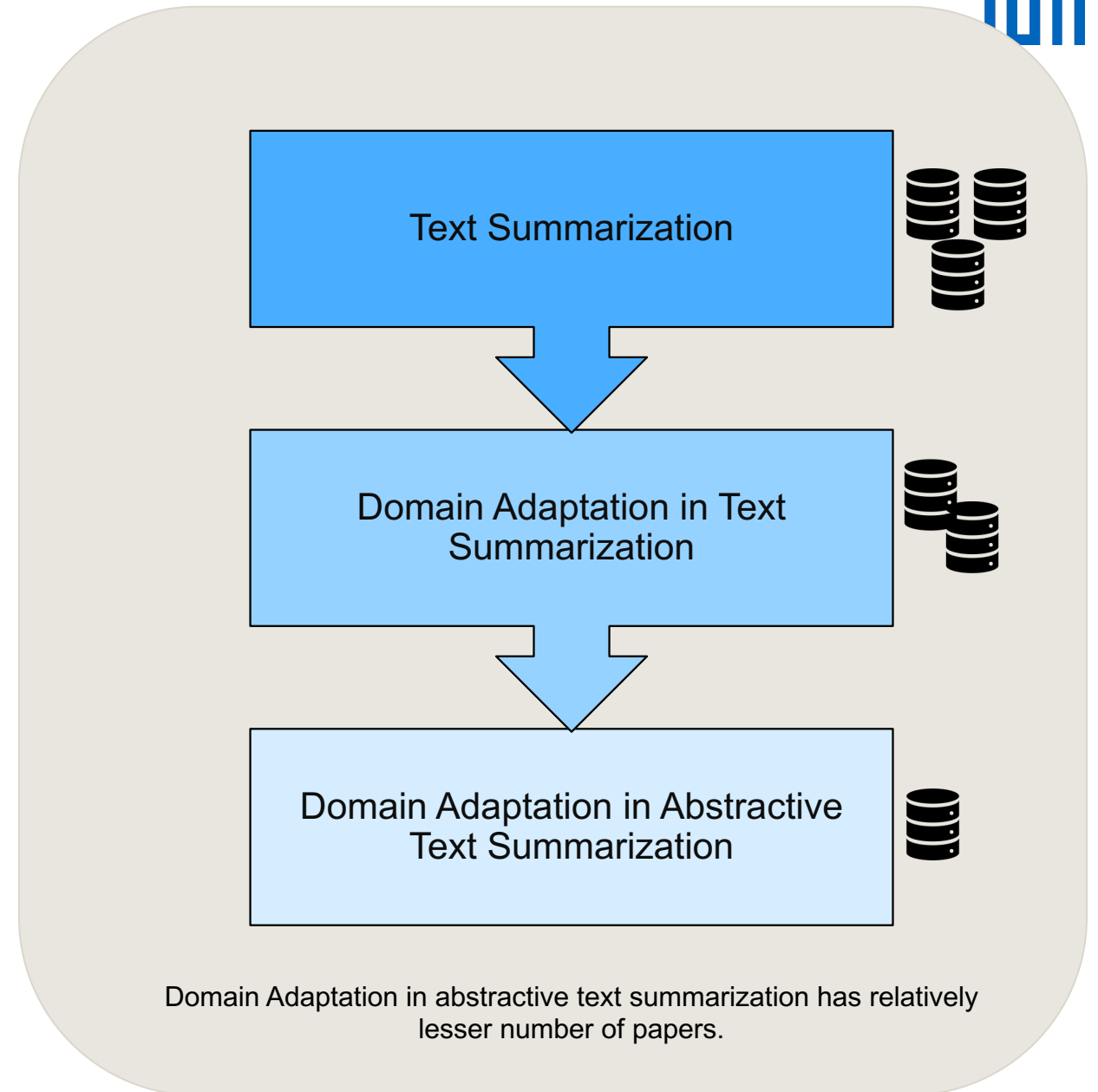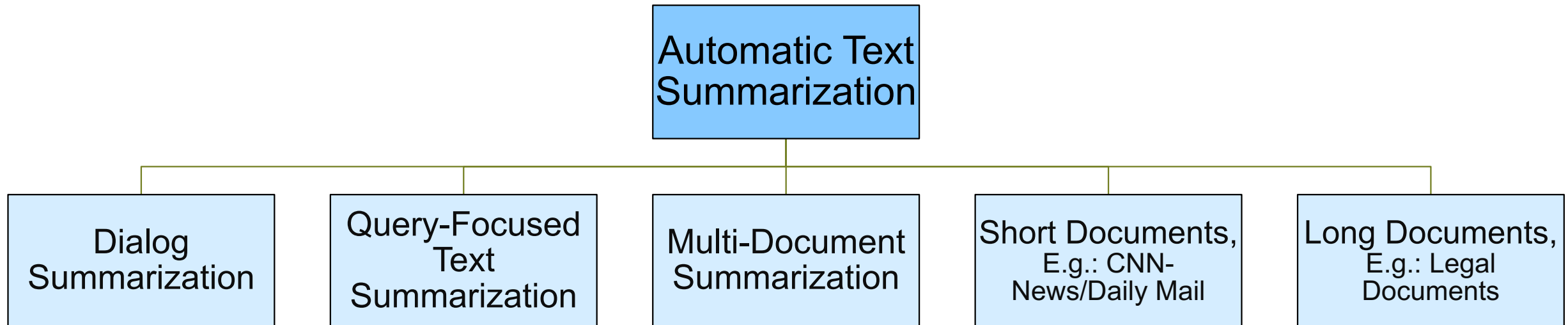Google Scholar

Total Papers Retrieved:
21

Text Summarization

Domain Adaptation in Text Summarization

Domain Adaptation in Abstractive Text Summarization

Domain Adaptation in abstractive text summarization has relatively lesser number of papers.

Literature Survey

Types of Text Summarization Papers

```
                        ┌─────────────────────┐
                        │  Automatic Text     │
                        │  Summarization      │
                        └─────────────────────┘
```

Dialog Summarization	Query-Focused Text Summarization	Multi-Document Summarization	Short Documents, E.g.: CNN-News/Daily Mail	Long Documents, E.g.: Legal Documents

Question 1:

What are the different types of models used for domain adaptation in abstractive text summarization?

RQ1: Models Used

Word Cloud

Models used in the surveyed papers.

These models have been used in the architecture either as a whole or in encoder/decoder parts.

BART is one of the popular choices.

Question 2:

What are the different types of domain adaptation techniques used for abstractive text summarization?

Question 3:

What is the criteria for classifying the various domain adaptation techniques found in the literature?

Literature Review: Data Centric Methods

The data-centric approaches focus on augmenting or selecting the training data to represent the target domain better.

Details

- **Muti-Phase Pretraining:**
  - Multi-phase pretraining is a machine learning technique where a model is pre-trained in multiple phases, each phase building on the knowledge learned in the previous phase.

- **Mixed-Domain Pretraining:**
  - It involves training a language model on diverse text data.
  - Improves the language model's performance by learning the transferable knowledge between different domains

- **Data Augmentation:**
  - Data augmentation (DA) denotes a set of techniques employed to enhance the variety of training instances without the need for procuring new data explicitly. Eg: Segmantic self segmentation.

Literature Review: Model-Centric Methods

These approaches focus on modifying the architecture or parameters of the summarization model to improve its performance on the target domain.

Details

- **Fine-Tuning:**
  - During the process of fine-tuning, the models are first pre-trained on large corpora, and then one task-specific layer is added to the model, which is trained from scratch while re-using the rest of the model parameters as a starting point.

- **Multi-Task Learning:**
  - It involves training a model to perform more than one NLP task at the same time with an aim to learn the shared knowledge of these tasks. For example: The model can learn the domain style by text reconstruction and text classification tasks on unlabelled domain-related target data

- **Architectural Modifications:**
  - Adapters : Adapters are lightweight plug-in modules that are added to the pre-trained models as a small number of new parameters.
  - Reinforcement Learning: Their model uses reinforcement learning reward metrics based on biomedical expert tools and can learn domain-specific information, thus producing domain-aware, abstractive summaries.

Literature Review: Hybrid Methods

These approaches combine elements of both model-centric and data-centric approaches to achieve better performance on the target domain.

Details

- **Multi-Stage Adaptation:**
  - Multi-stage adaptation technique uses different training stages to adapt the model to a specific task. Example: Classification of salient sentences followed by abstractive summarization.
- **Adversarial Training:**
  - Adversarial training is a process where a generator model and a discriminator are trained together. The model generates informative summaries for the target domain while the discriminator learns to distinguish between the source and target domains with an aim to learn domain-invariant features.
- **Meta-Learning:**
  - This method involves training a meta-learner to adapt the summarization model to new domains quickly and efficiently based on a few examples from each domain along with use of adapters.

Literature Review: Key-take aways for further research

- Model-centric approaches are one of the most prevalent domain adaptation techniques in abstractive text summarization.

- Adapters (Houlsby et al., 2019) have become a popular choice for low-resource settings. These are small plug-in modules that are added to the pre-trained models.

- Data-centric approaches are widely used to mitigate the issue of low-resource domain-related data. Data augmentation is one of the prevalent domain adaptation technique in NLP and hence has been studied for domain adaptation in abstractive text summarization as well.

- Hybrid approaches can be used as well to improve the overall performance of the model.

Future Work

Perform a comprehensive qualitative and quantitative evaluation of these techniques using a standardized datasets to compare and contrast their effectiveness.

This benchmarking process will aid in making well-informed decisions regarding the selection of appropriate domain adaptation techniques.

Aditi Arora

M.Sc. Student Informatics

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel   +49.89.289.   17132
Fax   +49.89.289.17136

aditi.arora@tum.de
wwwmatthes.in.tum.de