Neural Transfer Learning for Natural Language Processing

Student:	Sarthak Gupta
Matriculation:	03681155
Supervisor:	PD Dr. Georg Groh
Advisor:	Gerhard Hagerer
Chair:	Chair of Social Computing
University:	Technical University of Munich
Department:	Faculty of Informatics
Date:	09.05.2019
Place:	Munich, Germany
Timeline:	15.09.2018 – 15.04.2019

Munich RE

Agenda

1. Fundamentals
   i. Motivation
   ii. Problem Statement
   iii. Taxonomy
   iv. Methodology
   v. Experiments
2. Transfer Learning for NLP
   i. Tasks
   ii. Datasets
   iii. Experiments
3. Output
   i. Result
   ii. Research Questions
   iii. Discussion and Conclusion
   iv. Contribution
   v. Limitation
   vi. Future Work

Munich RE

Section 1

What, Why and How

“Science is not about building a body of known “facts”. It is a method for asking awkward questions and subjecting them to a reality-check, thus avoiding the human tendency to believe whatever makes us feel good.”

- Terry Pratchett, The Science of Discworld

Motivation

1. Learning from a blank state is antithetic to the way humans acquire language.
2. Labelled data is not available in abundance.
3. Current solutions do not generalize
   i. Breaks under adversarial noise [Jia, Robin & Liang, Percy. (2017); Belinkov, Yonatan & Bisk, Yonatan. (2017)]
   ii. Breaks under domain shift
4. Manual Labeling is costly
   i. Domain Expertise
   ii. Annotator's Subjectivity
   iii. Costs with Crowd-Sourcing
5. Semantic Relatedness of multiple NLP tasks
   i. Language Modeling for Sentiment Classification

Problem Statement

Hypothesis

Deep neural networks in natural language processing that leverage existing relevant information from related domains, tasks, and languages outperform models not using this information across a wide range of tasks. [S. Ruder. (2019)]

Idea

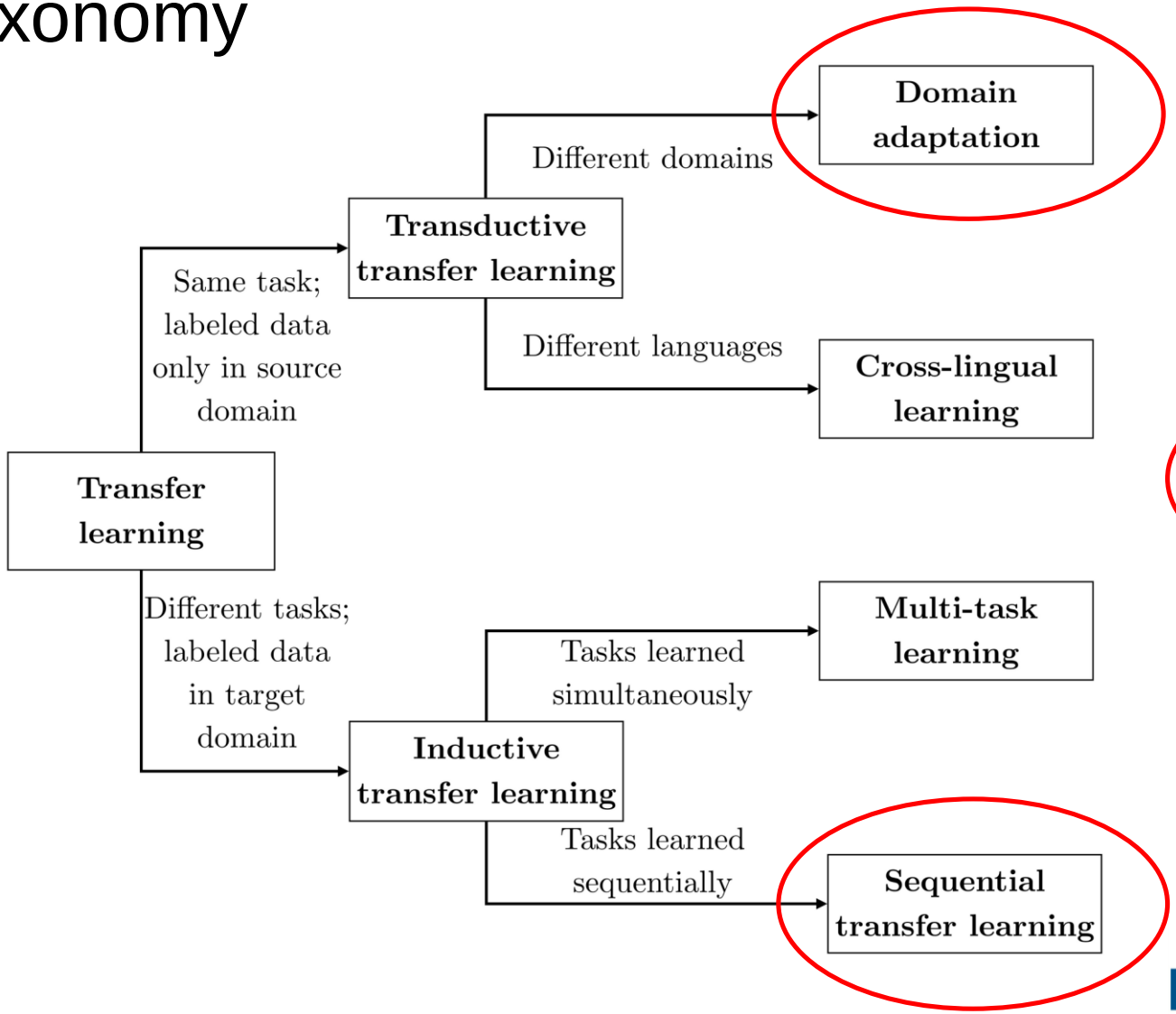
Investigate various techniques of Sequential Transfer Learning for Deep Learning based Natural Language Processing

Tasks

1. Transfer across same tasks but different domains with same label space
2. Transfer across same tasks but disparate domains with distinct label space
3. Transfer across multiple tasks

Munich RE

Taxonomy

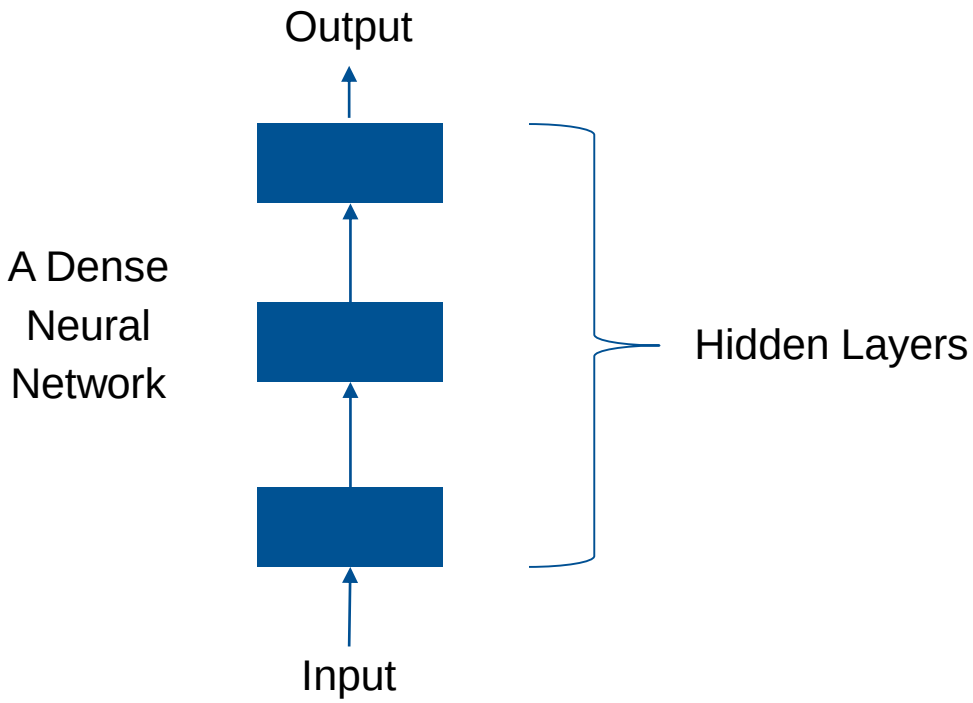


Focus of this Thesis

Image taken from [3]

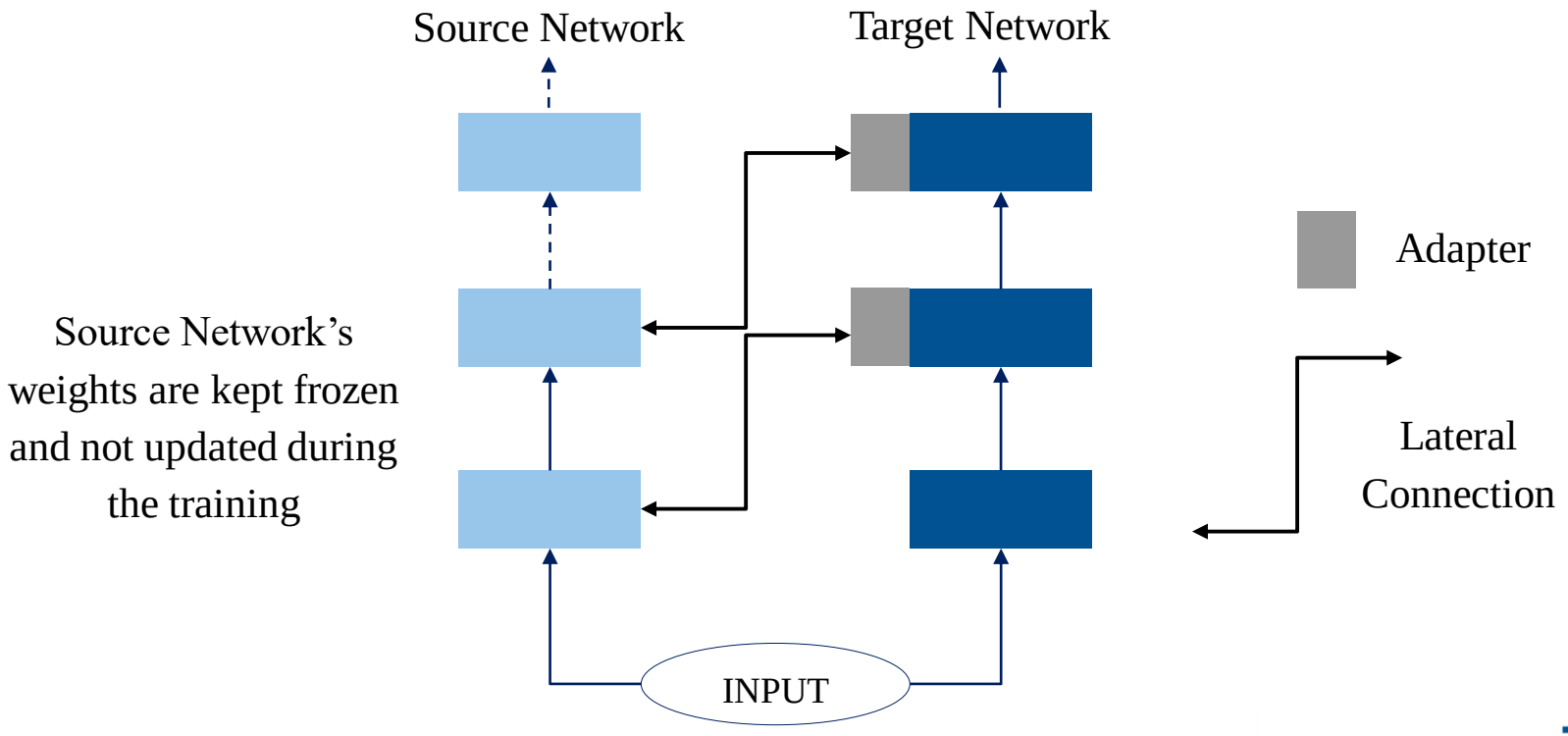
Methodology

Progressive Neural Networks: Adapt the intermediate representations of a source network to a target network [Rusu et. al., (2016)]



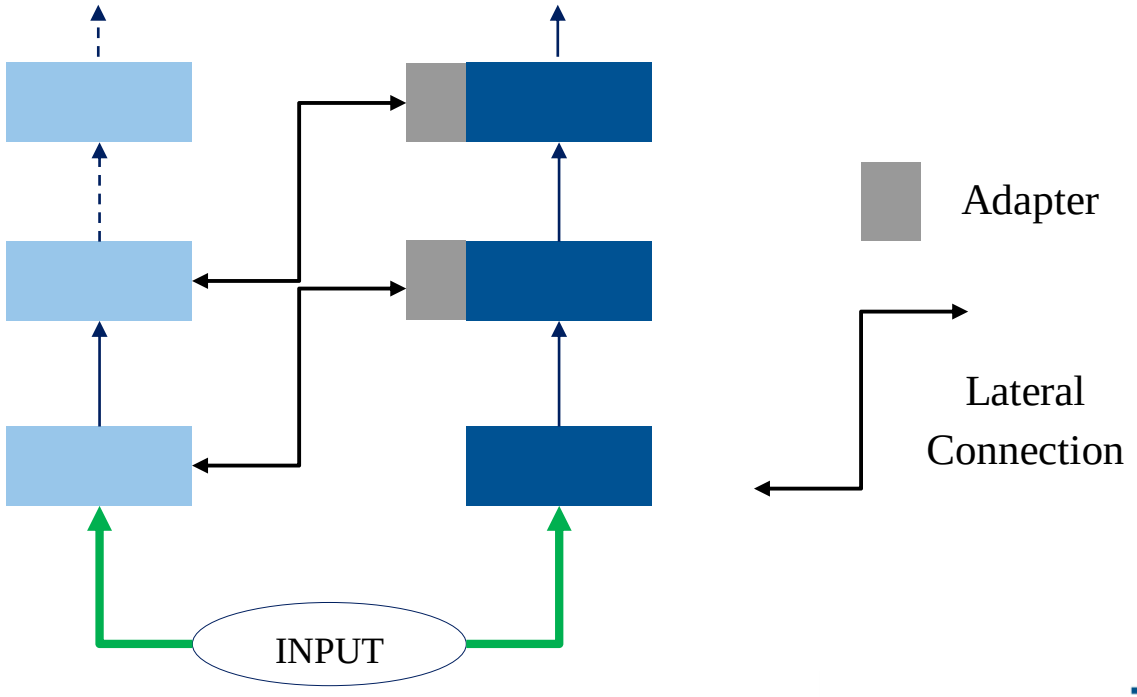
Methodology

Progressive Neural Networks: Adapt the intermediate representations of a source network to a target network



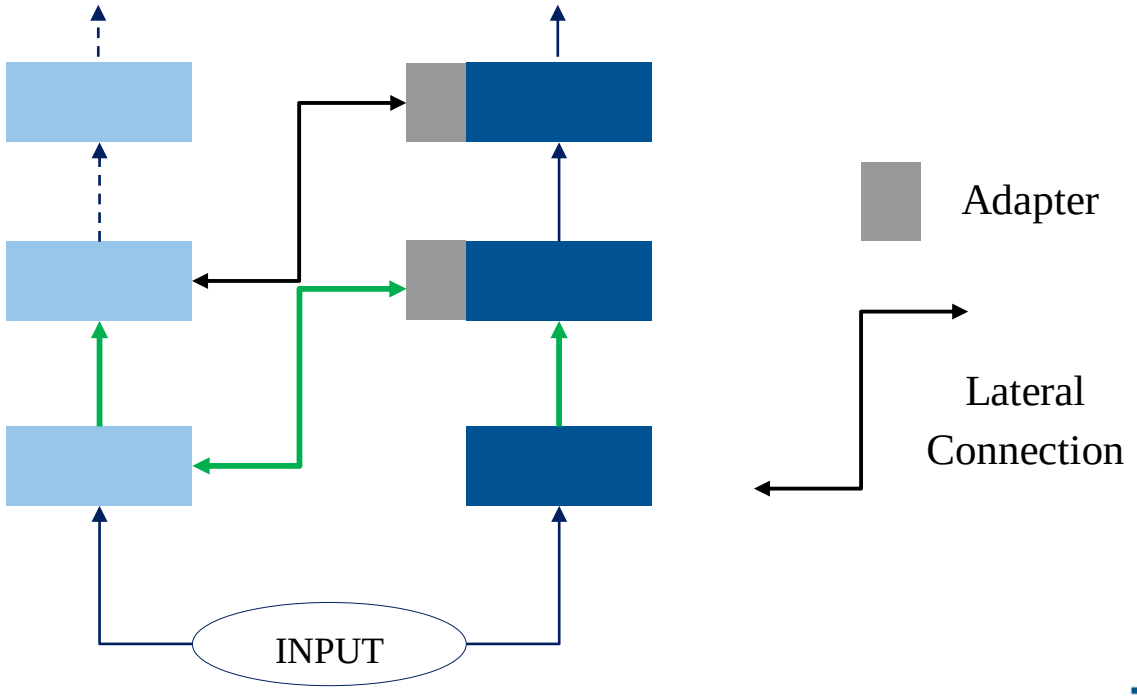
Methodology

Progressive Neural Networks: Adapt the intermediate representations of a source network to a target network



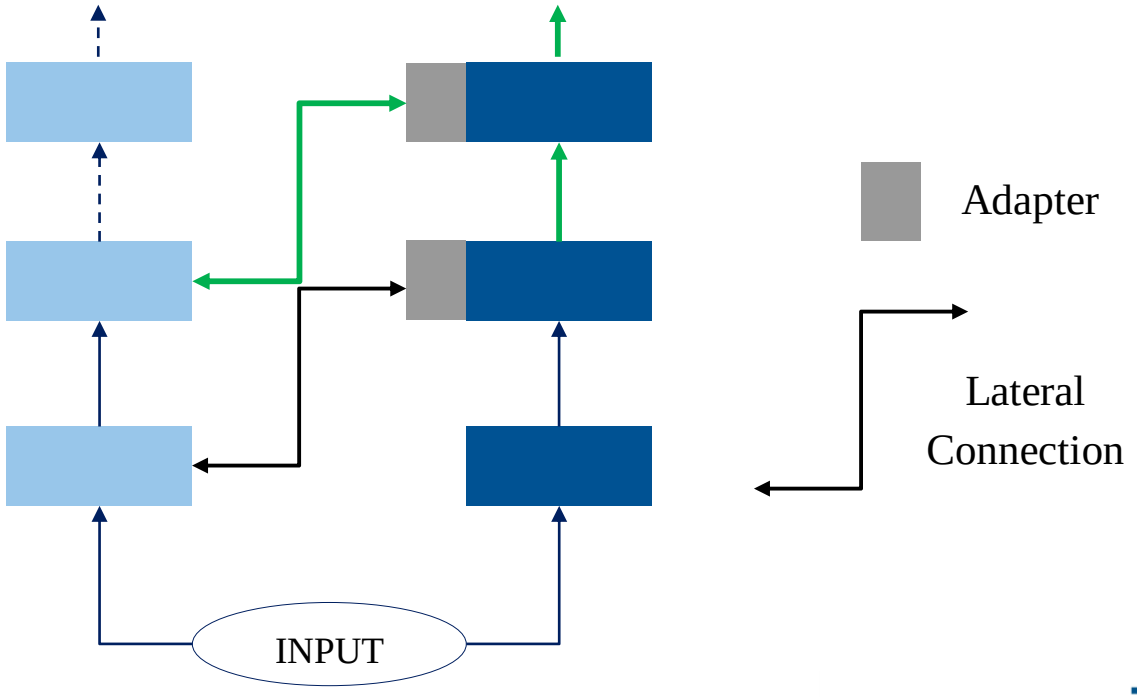
Methodology

Progressive Neural Networks: Adapt the intermediate representations of a source network to a target network



Methodology

Progressive Neural Networks: Adapt the intermediate representations of a source network to a target network



Experiments

Aims:

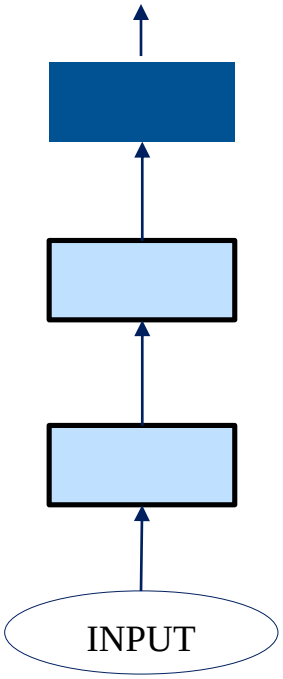
1. To compare Progressive Nets with the standard technique and find out the appropriate scenarios for each technique.
2. To study the effect of various hyperparameters and techniques on the performance of the Progressive Nets.

Categories

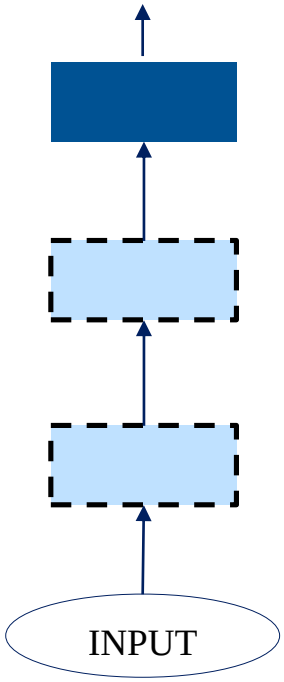
Category	Experiment
A	establish conclusions regarding Transfer Learning for NLP
B	establish conclusions regarding Progressive Net Architecture
C	establish conclusions regarding the single task networks.

Munich RE

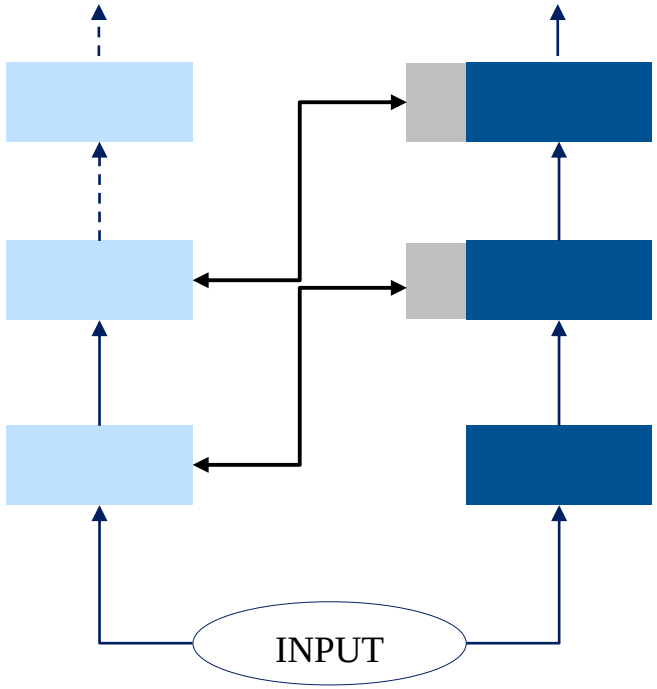
Experiments: Category A



Fine-Tune last layer (TL-L)

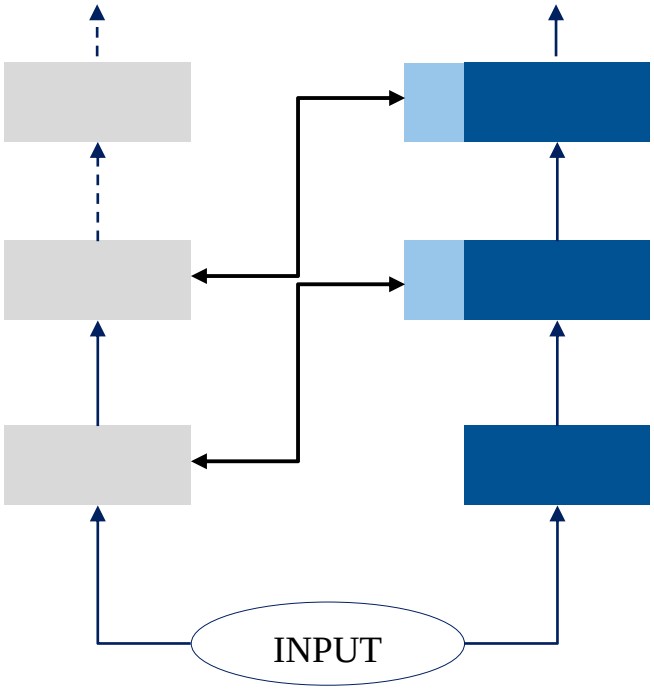


Fine-Tune all layers (TL-A)

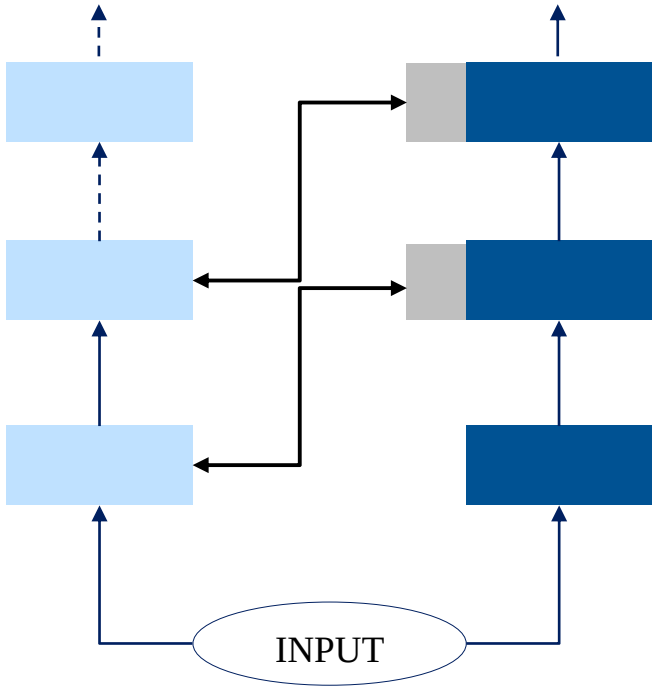


Progressive Net with best source (PNN-B)

Experiments: Category B

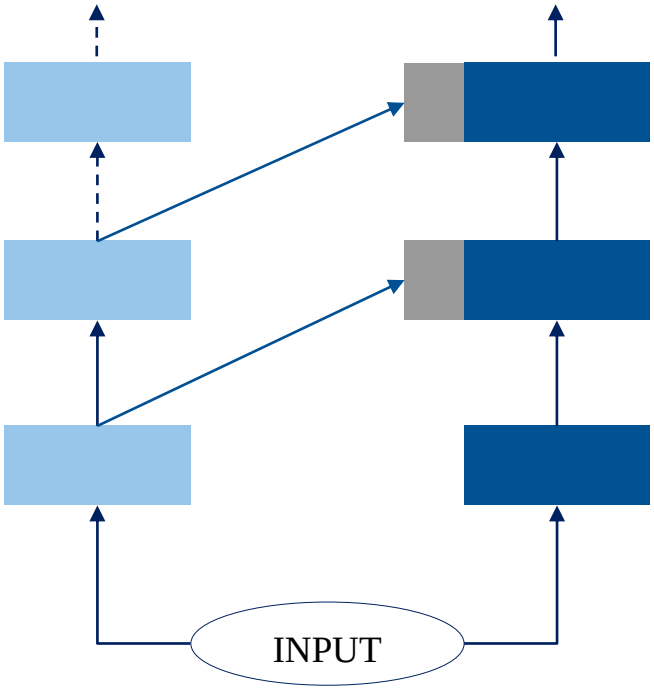


Random Source Column (PNN-R)

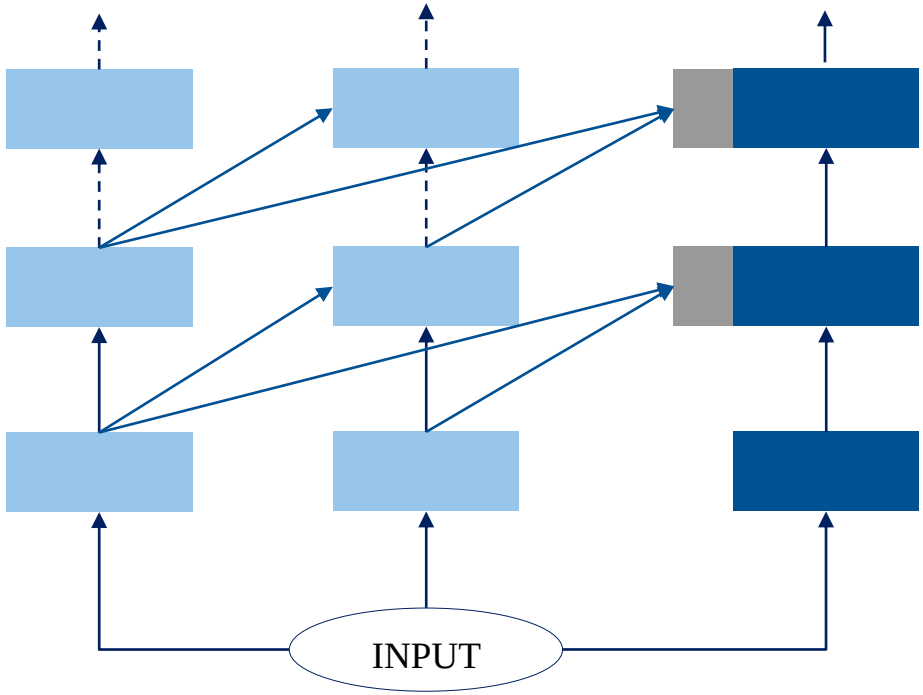


Optimized Source Column (PNN-R)

Experiments: Category B



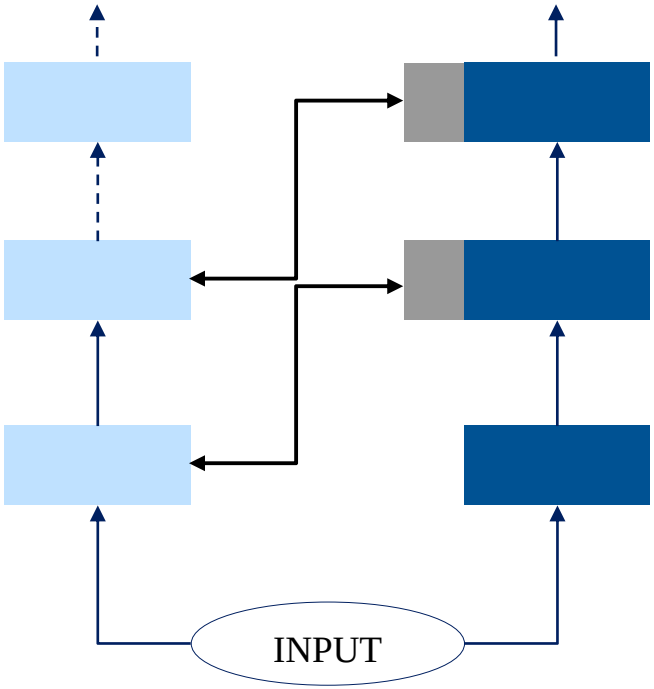
1 Source Column
(1PNN-B)



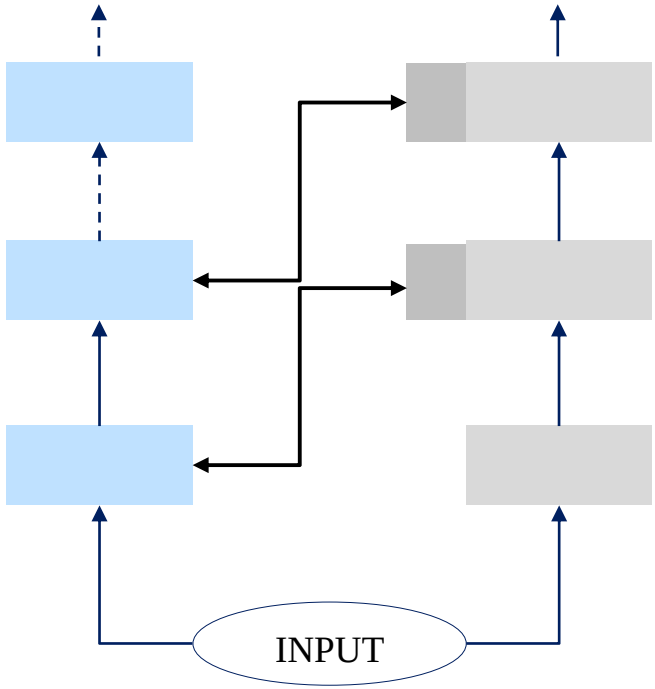
2 Source Column
(2PNN-

Munich RE

Experiments: Category B



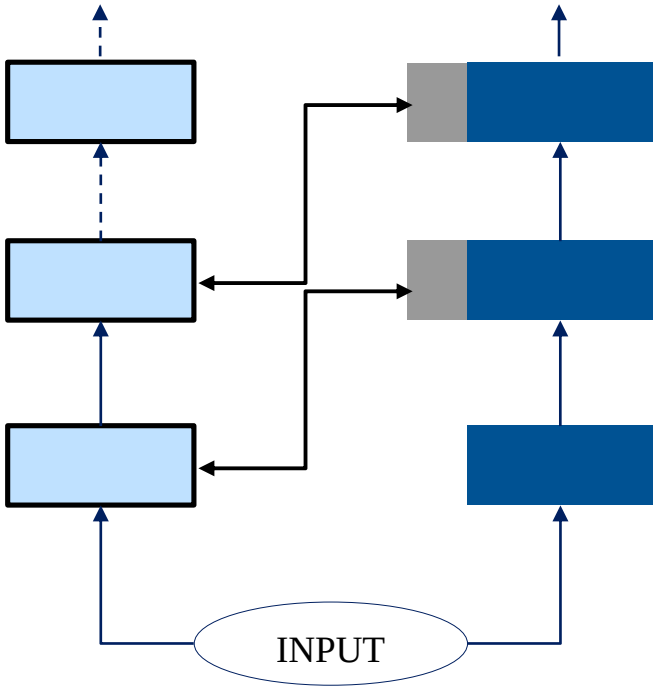
Optimized Target Column



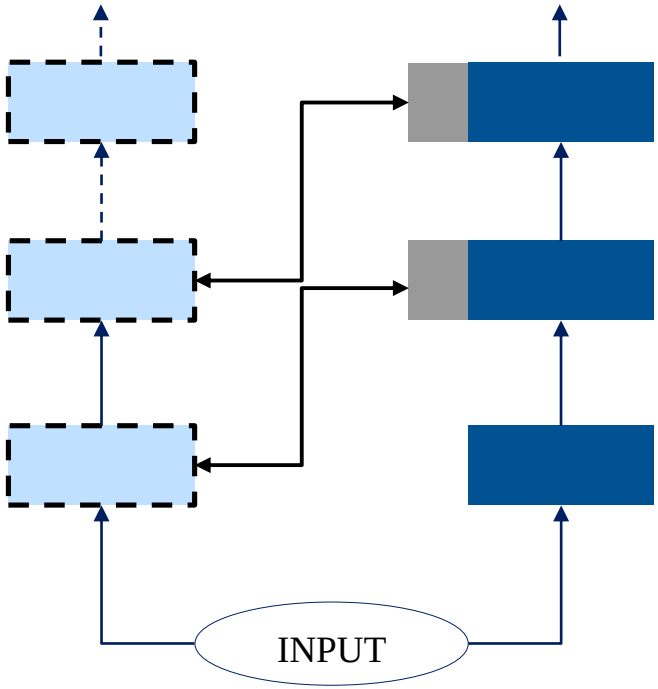
Random Target Column

Munich RE

Experiments: Category B



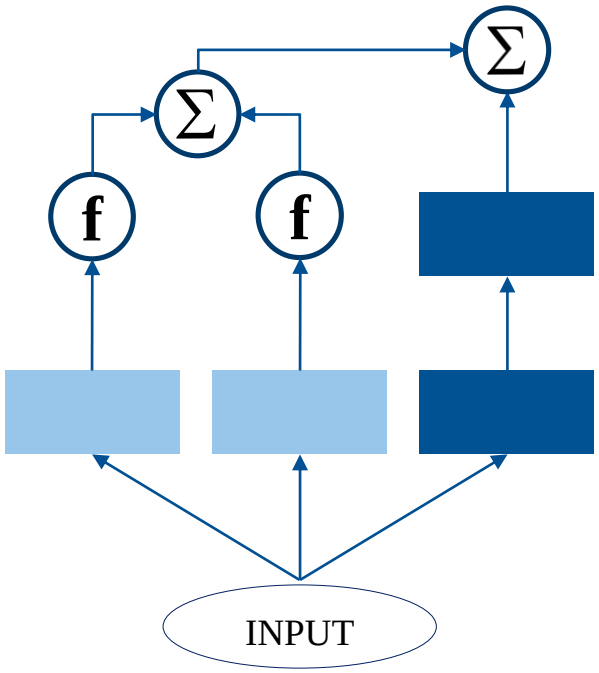
Frozen Source Column



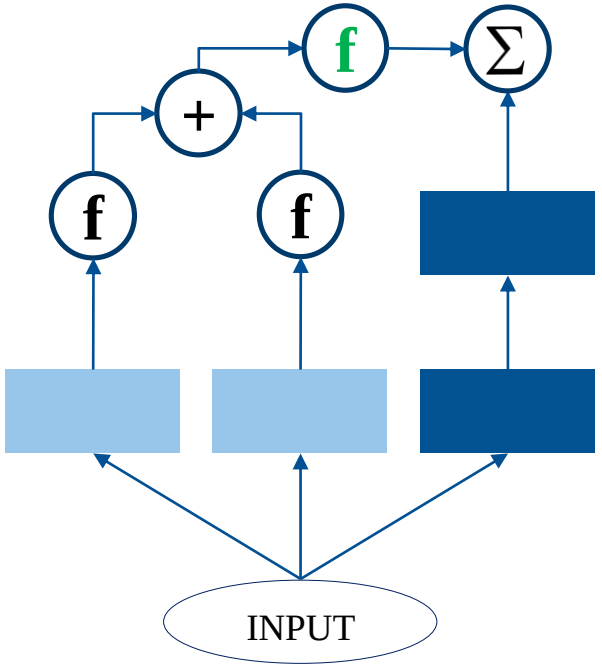
Trainable Source Column

Munich RE

Experiments: Category B



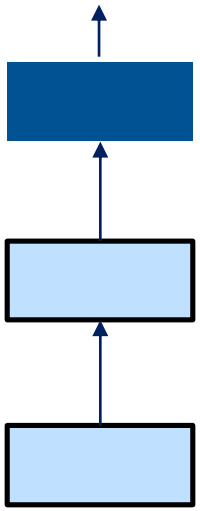
Linear Adapter



Non-Linear Adapter

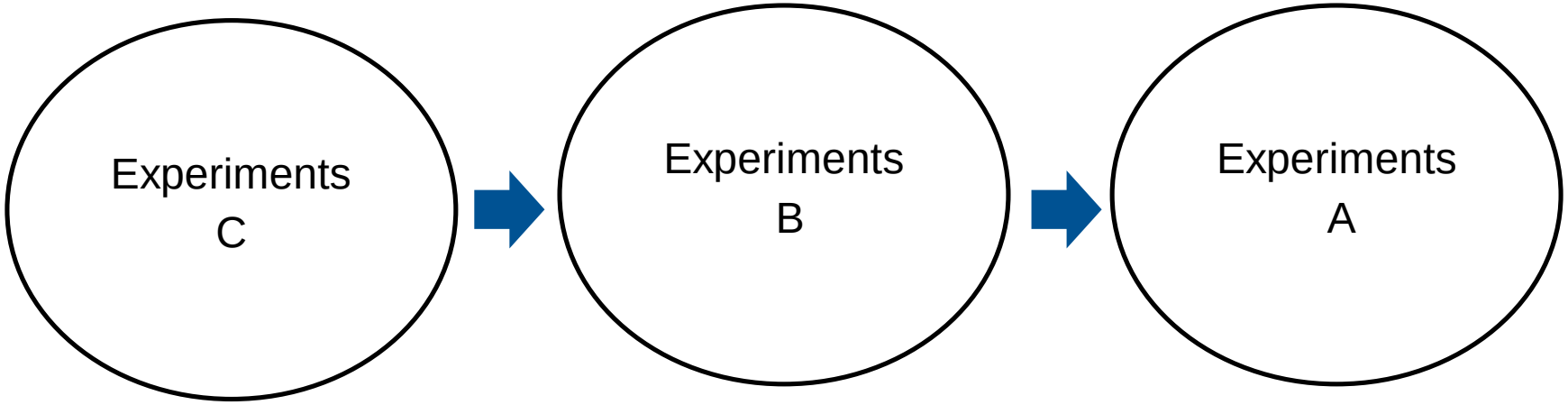
Experiments: Category C

1. Word Embeddings
2. Number of Layers
3. Dropout
4. …other task specific hyperparameters



Single Task (ST)

Experiments: Flow



*The experiments are performed on varying proportion of the training dataset

End of Section 1

- **What** this Thesis is about?
- **Why** are we doing it?
- What is already **existing** to solve the problem?
- **How** do we do it?

Section 2

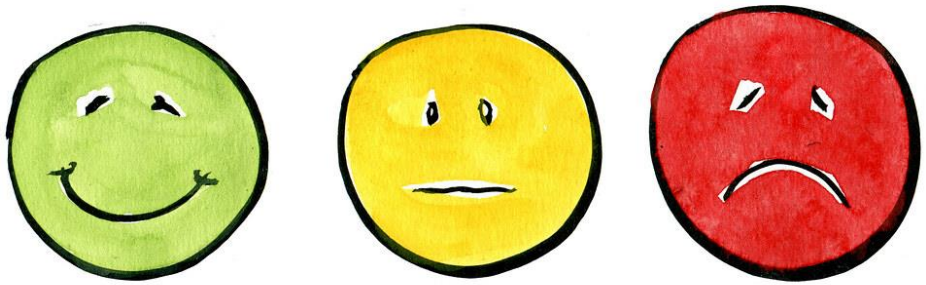
Transfer Learning for NLP

Tasks

My preferred candidate is Cary Moon (PER) , but she wo n't be the next mayor of Seattle (LOC) .

Named Entity Recognition

"Organic foods aren't subjected to the potentially harmful pesticides or drugs that non organic foods generally are ."

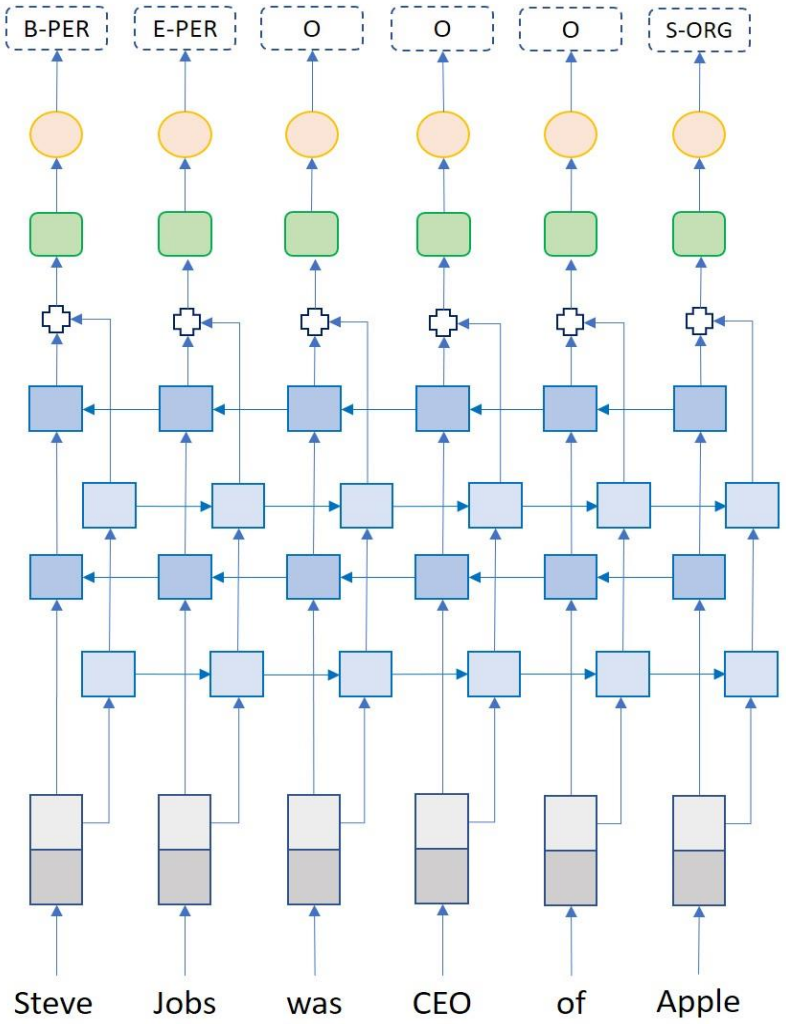


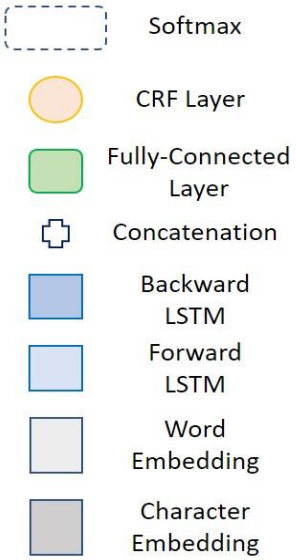
Sentiment Analysis

NER example taken from: https://demo.allennlp.org/named-entity-recognition/NzM1ODMy

Munich RE

Tasks: Named Entity Recognition

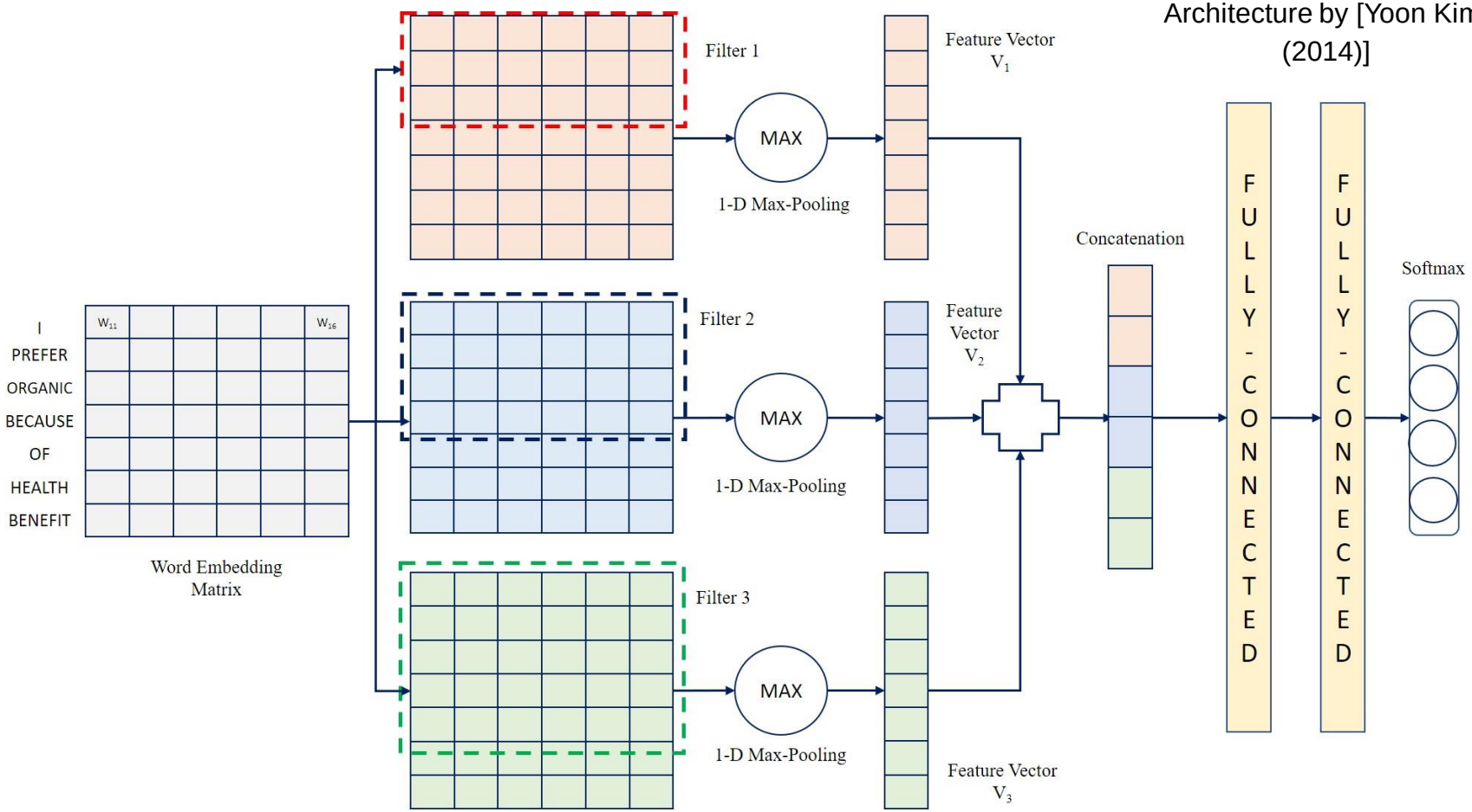




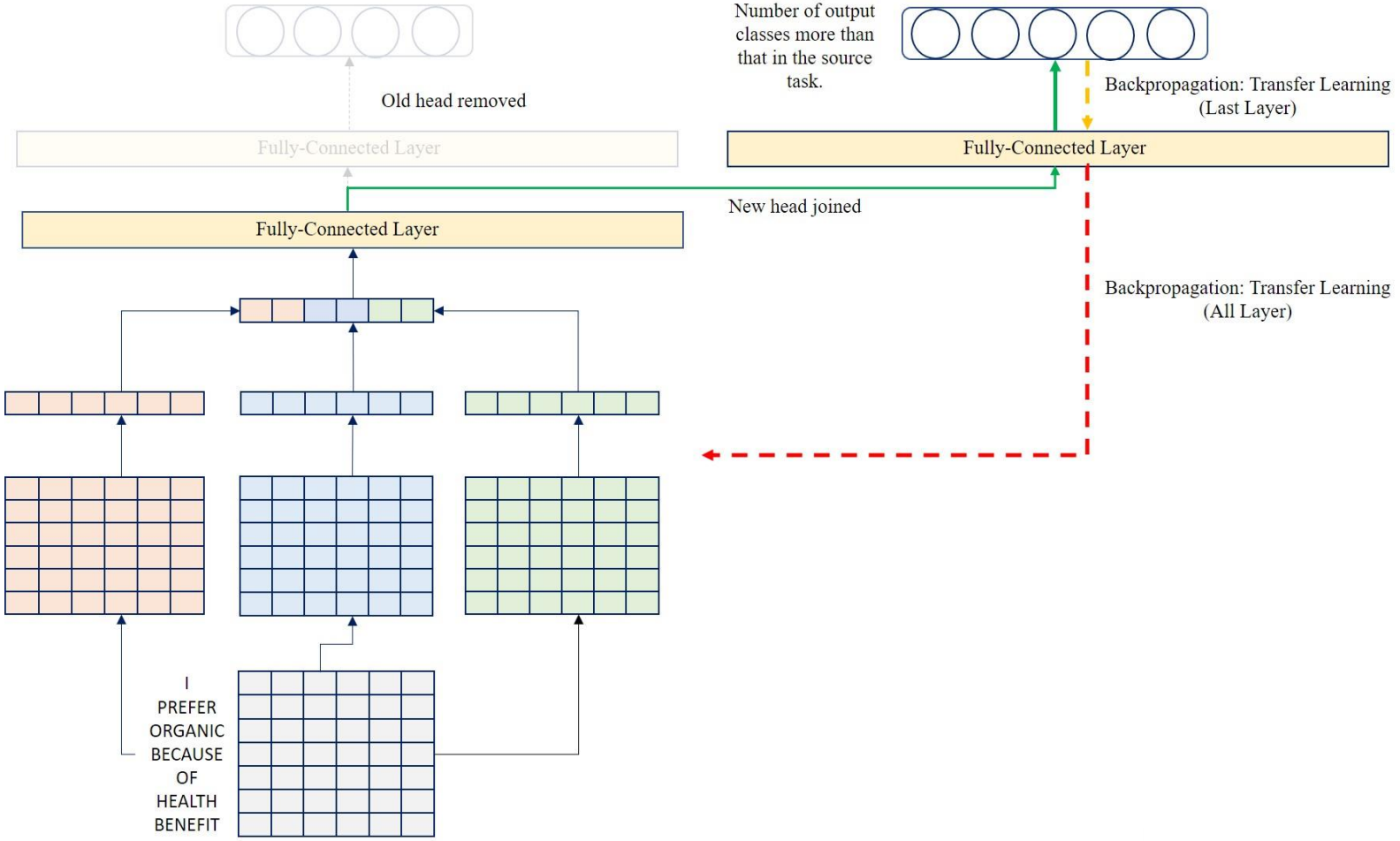
Architecture by [Ma & Hovy, 2016]

Tasks: Sentiment Analysis

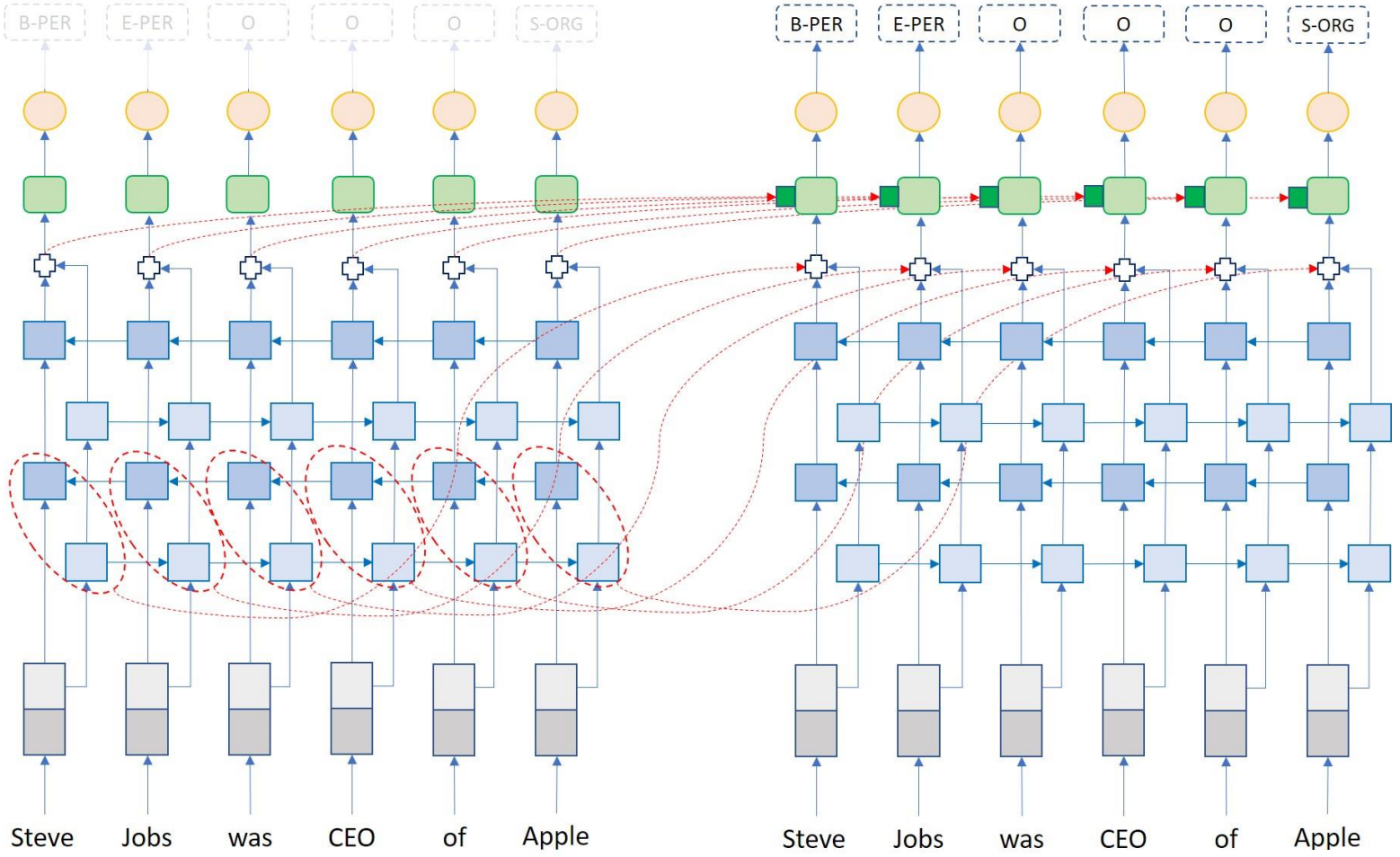
Architecture by [Yoon Kim, (2014)]



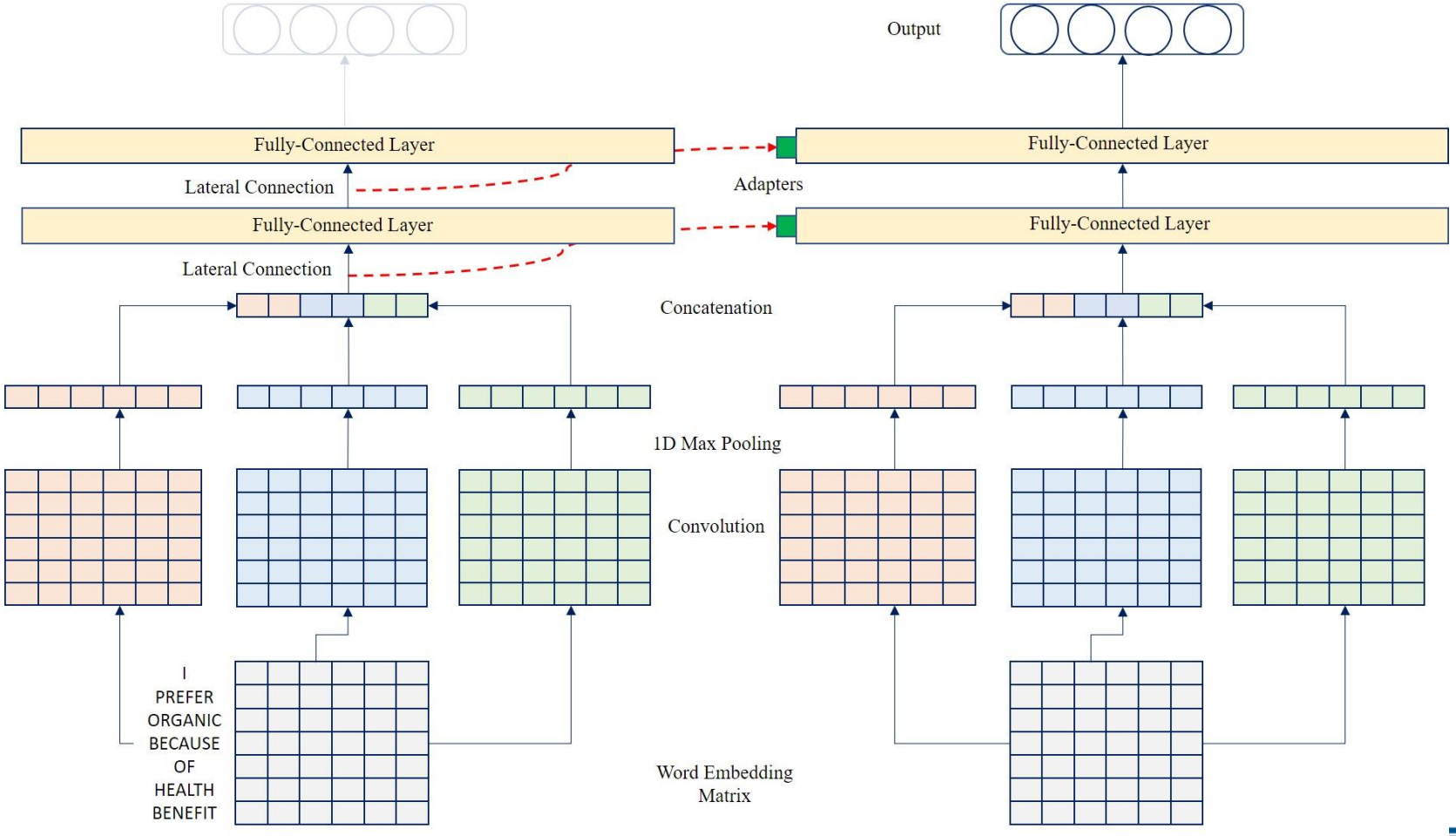
Experiments: TL-L, TL-A for Sentiment Analysis



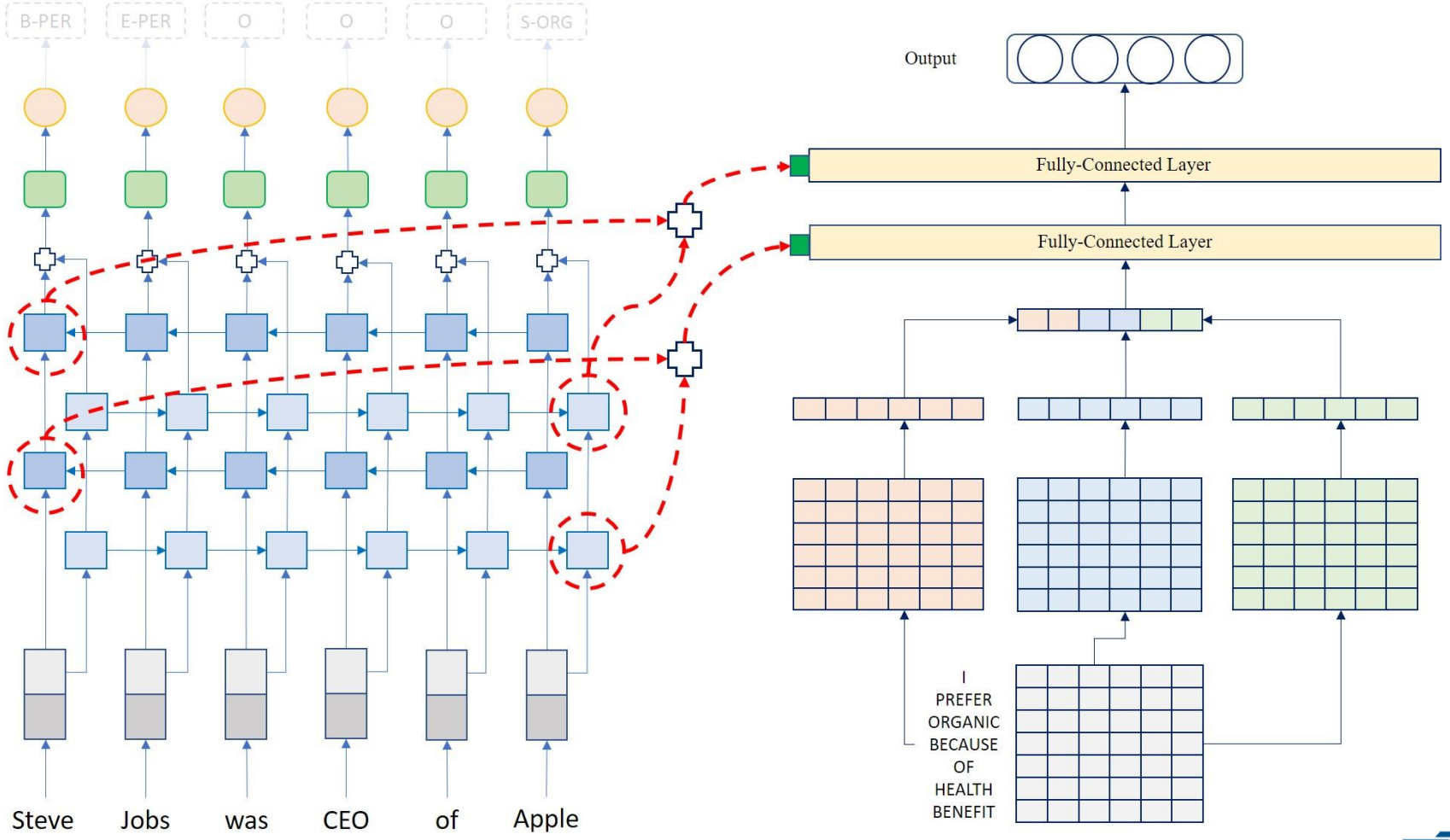
Experiments: PNN for NER



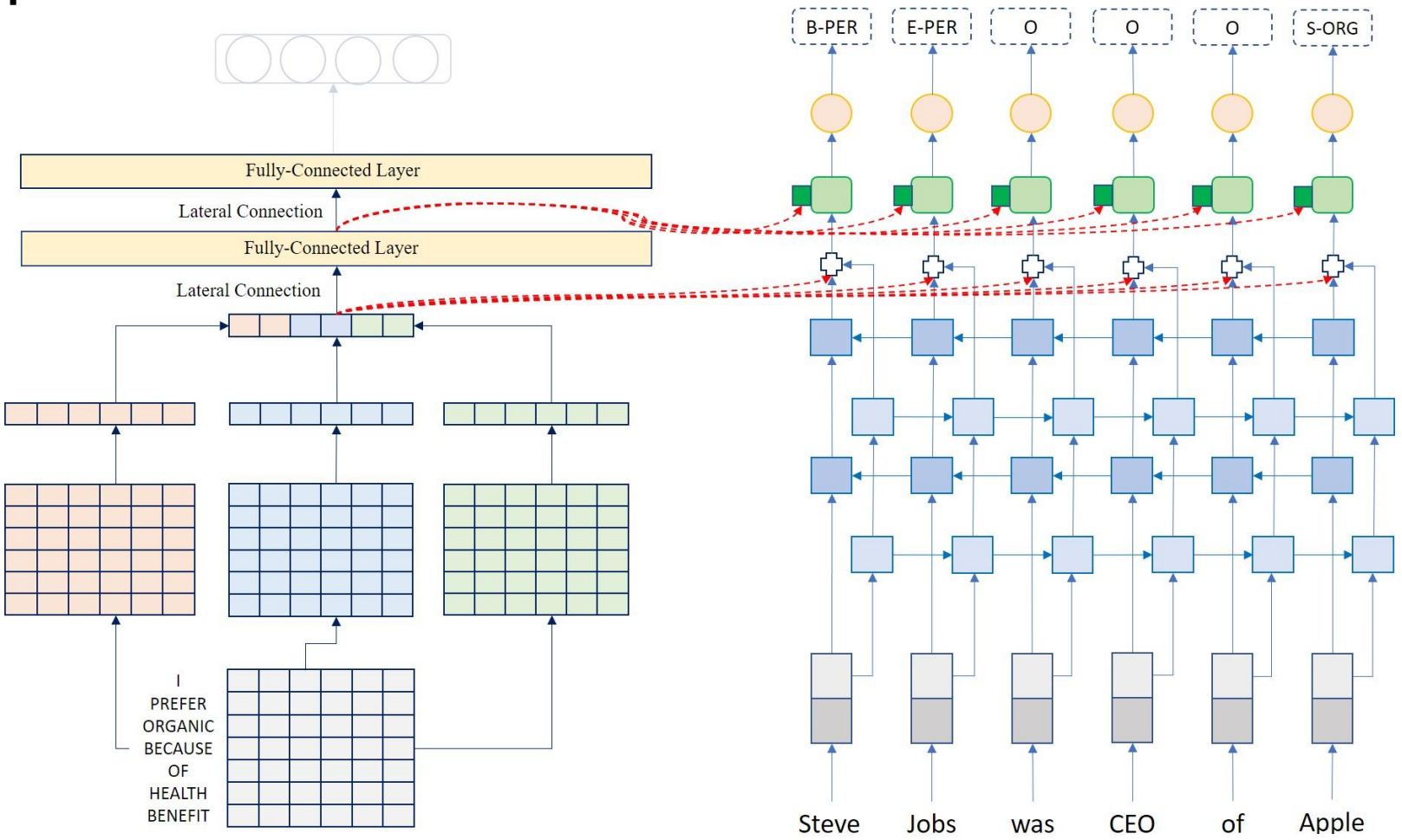
Experiments: PNN for Sentiment Analysis



Experiments: PNN from NER to SA



Experiments: PNN from SA to NER

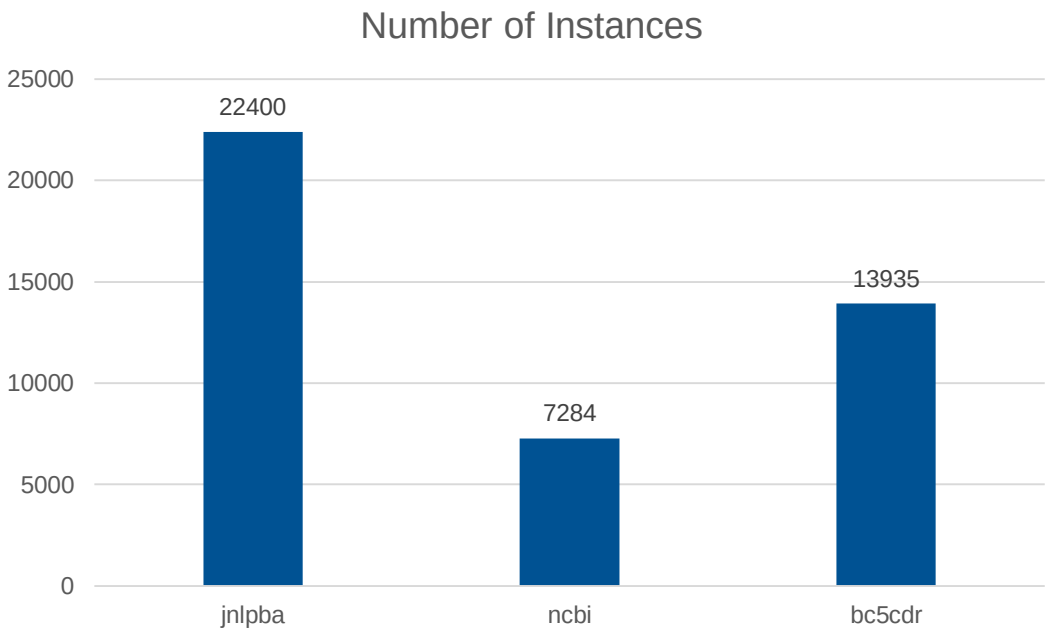


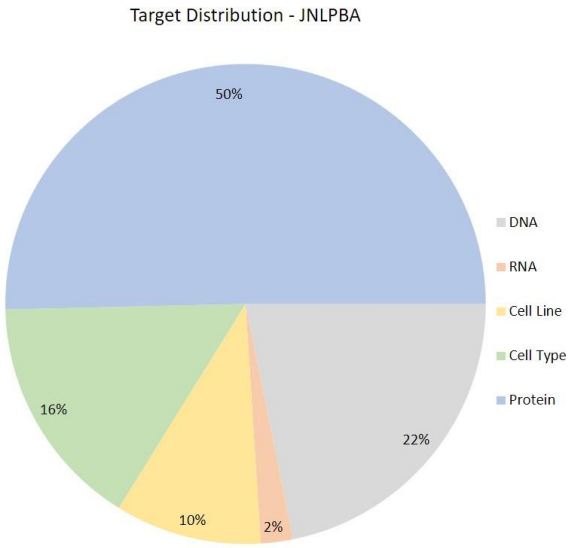
Datasets: Named Entity Recognition

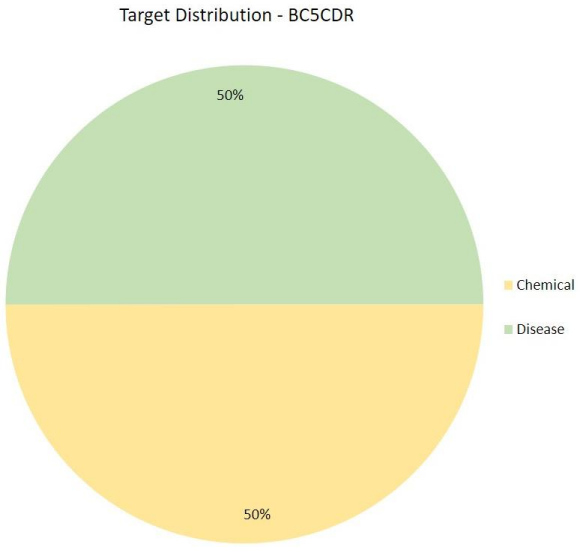
Corpora	Labels	Example
JNLPBA	Cell Type, Cell Line, DNA, RNA, Protein	“[T cells]$_{\text{CELLTYPE}}$ lack active [NF-kappa]$_{\text{PROTEIN}}$ B but express [Sp1]$_{\text{PROTEIN}}$ as expected .”
NCBI	Disease	“The risk for [ovarian cancer]$_{\text{DISEASE}}$ was 2 .”
BC5CDR	Disease, Chemical	“Worsening of [levodopa]$_{\text{CHEMICAL}}$ - induced [dyskinesias]$_{\text{DISEASE}}$ by motor and mental tasks .”

Munich RE

Datasets: NER



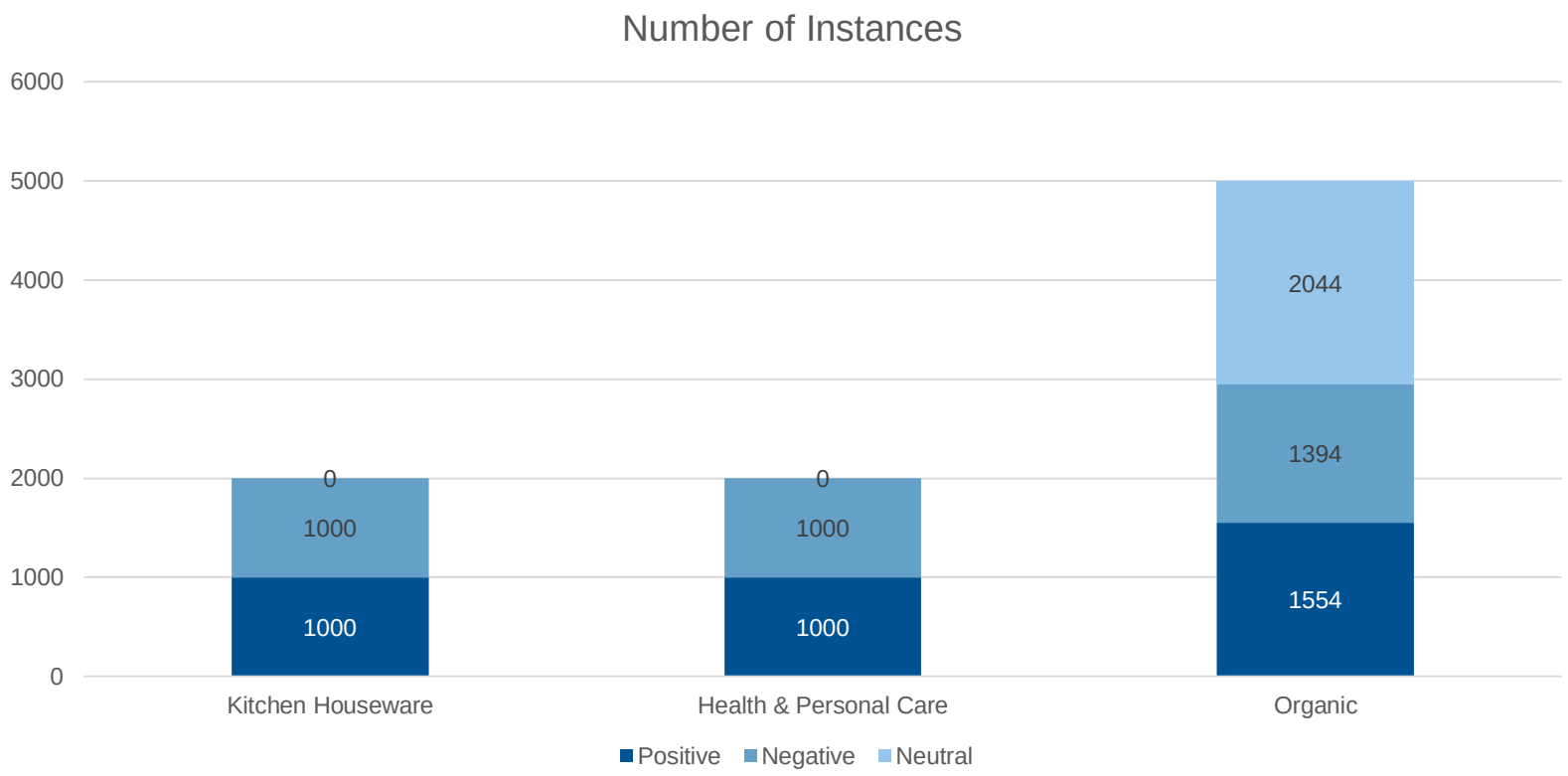




Datasets: Sentiment Analysis

Corpora	Labels	Example
Kitchen Housewares	Positive, **Negative**	*"this steamer performs very poorly . The wrinkles in my shirts just laughed when they saw it coming . now it sits in my closet collecting wasting space ."*
Personal Healthcare	**Positive**, Negative	*"this is one of the nicest electric shavers that I've owned and the first wet-type shaver ."*
Organic	**Positive**, Negative, Neutral	*"Organic foods aren't subjected to the potentially harmful pesticides or drugs that non organic foods generally are ."*

Datasets: Sentiment Analysis



End of Section 2

- Named Entity Recognition
- Sentiment Analysis
- How to apply Transfer Learning for Cross-Task and Same-Task Transfer
- Datasets for evaluation

Section 3

Evaluation, Results, Discussion and Conclusion

Evaluation

- **Precision:** Percentage correct out of all the positives found
- **Recall:** Percentage of all the positives found
- **F1 score:** Harmonic mean of Precision and Recall
- **Accuracy:** Percentage of correct found

Results: Category C

Pamameter	Value 1	Value 2
LSTM layers	*Two*	*One*
	90.7	90.34
Tagging	*IOBES*	*IOB2*
	90.8	90.7
CRF	*Yes*	*No*
	91.21	90.8
Casing	*Yes*	*No*
	91.1	**91.21**

Experiments for Named Entity Recognition

Pamameter	Value 1	Value 2
Conv. layers	*Two*	*One*
	79.25	**80.5**
Word Embedding	*Word2Vec (300D)*	*GloVe (100D)*
	82.9	80.5

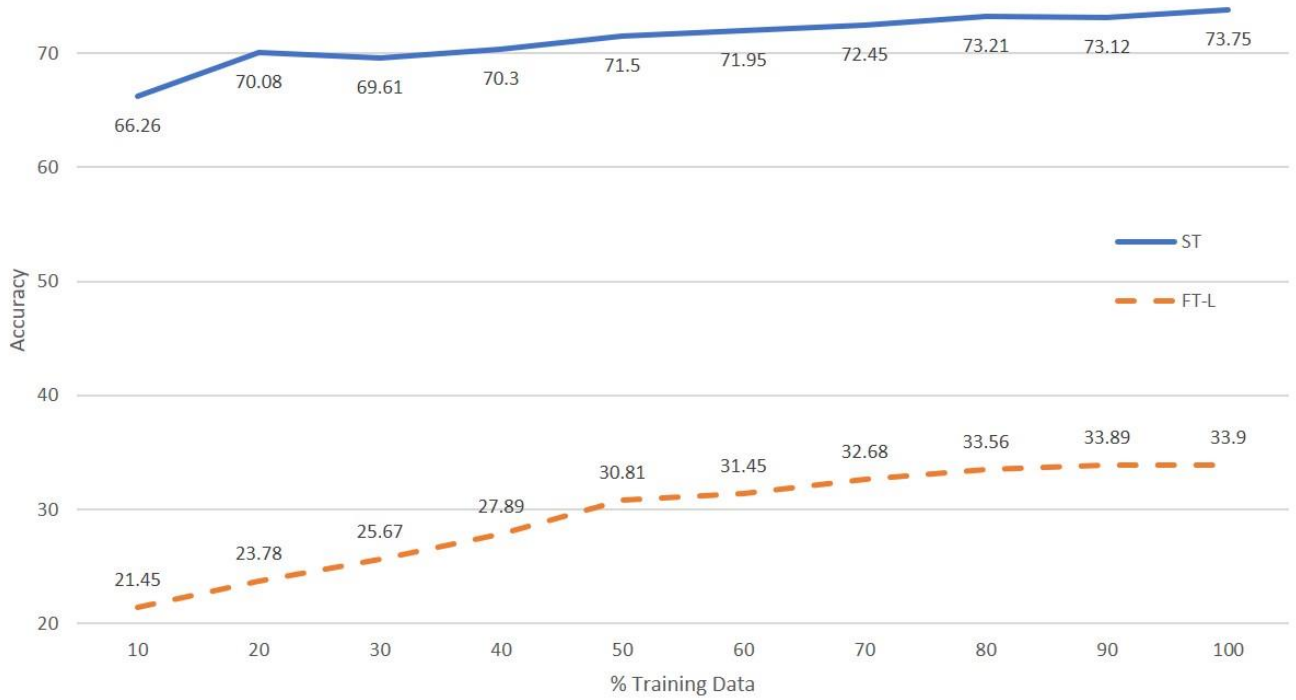
Experiments for Sentiment Analysis

Results: Category B

Pamameter	Value 1	Value 2
Lateral Connections	*Two*	*One*
	85.1	84.3
Adapter	*Non-Linear*	*Linear*
	85.3	85.1
Source Column Weights	*Frozen*	*Trainable*
	85.3	**86.4**
Target Column Weights	*Best*	*Random*
	85.33	85.31

Experiments for Progressive Neural Networks. The experiments were done in 1PNN-B setting where JNLPBA dataset is used as the source column and NCBI dataset as the target column

Results: Category B



The graph shows the performance of the NER architecture trained on NCBI dataset transferring information to JNLPBA dataset using the FT-L setting

Results: Category A

Target	ST	TL-A	1PNN-R	1PNN-B	2PNN-B
JNLPBA	73.75	73.14 (N)	73.87	74.1 (N)	**74.3 (N, B)**
NCBI	84.9	85.1 (J)	84.4	85.46 (B)	**85.7 (B, J)**
BC5CDR	81.7	83.74 (N)	84.1	**84.8 (B)**	84.1 (B, J)

Scores for Transfer experiments on NER datasets. **ST:** Single Task, **TL-A:** Transfer Learning All-Layer, **1PNN-R:** Progressive Neural Network with 1 source column initialized randomly, **1PNN-B:** Progressive Neural Network with 1 source column, initialized optimally, **2PNN-B:** 2 source columns, **N:** NCBI dataset, **J:** JNLPBA dataset, **B:** BC5CDR dataset. **All scores are shown for 100% training dataset.**

Munich RE

Results: Category A

Target	ST	TL-A	1PNN-B	2PNN-B
Kitchen Housewares	79	79 (H)	82.56 (K)	**83.18 (H, O)**
Personal Healthcare	84.9	81.3 (K)	**82.9 (K)**	81.8 (K, O)
Organic	45.83	48.61 (K)	53.76 (K)	**54.6 (K, H)**

Scores for Transfer experiments on Sentiment Analysis datasets. **ST:** Single Task, **TL-A:** Transfer Learning All-Layer, **1PNN-B:** Progressive Neural Network with 1 source column, initialized optimally, **2PNN-B:** 2 source columns.
H: Personal Healthcare, **K:** Kitchen Housewares, **O:** Organic dataset. **All scores are shown for 100% training dataset**

Munich RE

Results: Category A

Source (NER)/Target (SA)	Kitchen Housewares **(79)**	Personal Healthcare **(84.9)**	Organic **(45.83)**
JNLPBA	74.2	71.4	32.5
NCBI	64.19	-	-
BC5CDR	-	-	41.3

Scores for transfer from NER to Sentiment Analysis using PNN. **100% training dataset**

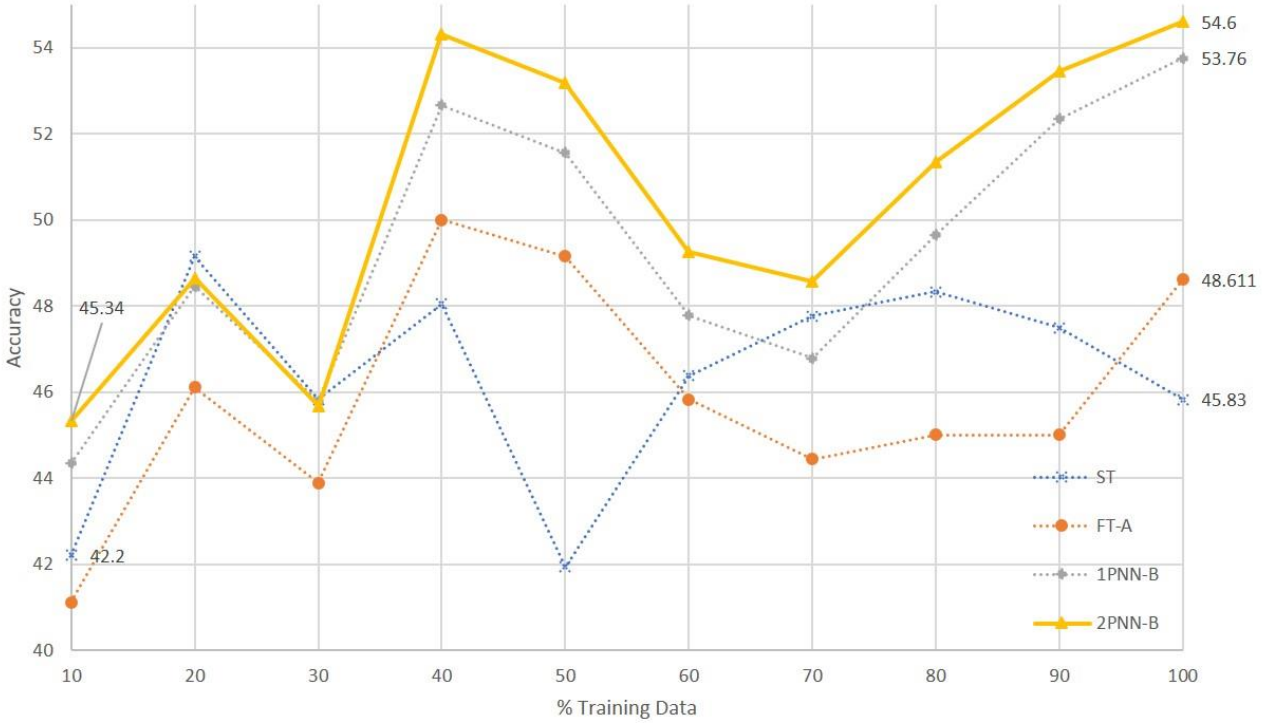
Source (SA)/Target (NER)	JNLPBA **(73.75)**	NCBI **(84.9)**	BC5CDR **(81.7)**
Kitchen Housewares	63.17	71.6	57.2
Personal Healthcare	-	-	-
Organic	52.15	43.7	49.4

Scores for transfer from Sentiment Analysis to NER using PNN. **100% training dataset**

Munich RE

Research Questions

Q1. Can Transfer Learning help in improving the performance when the training data is less?

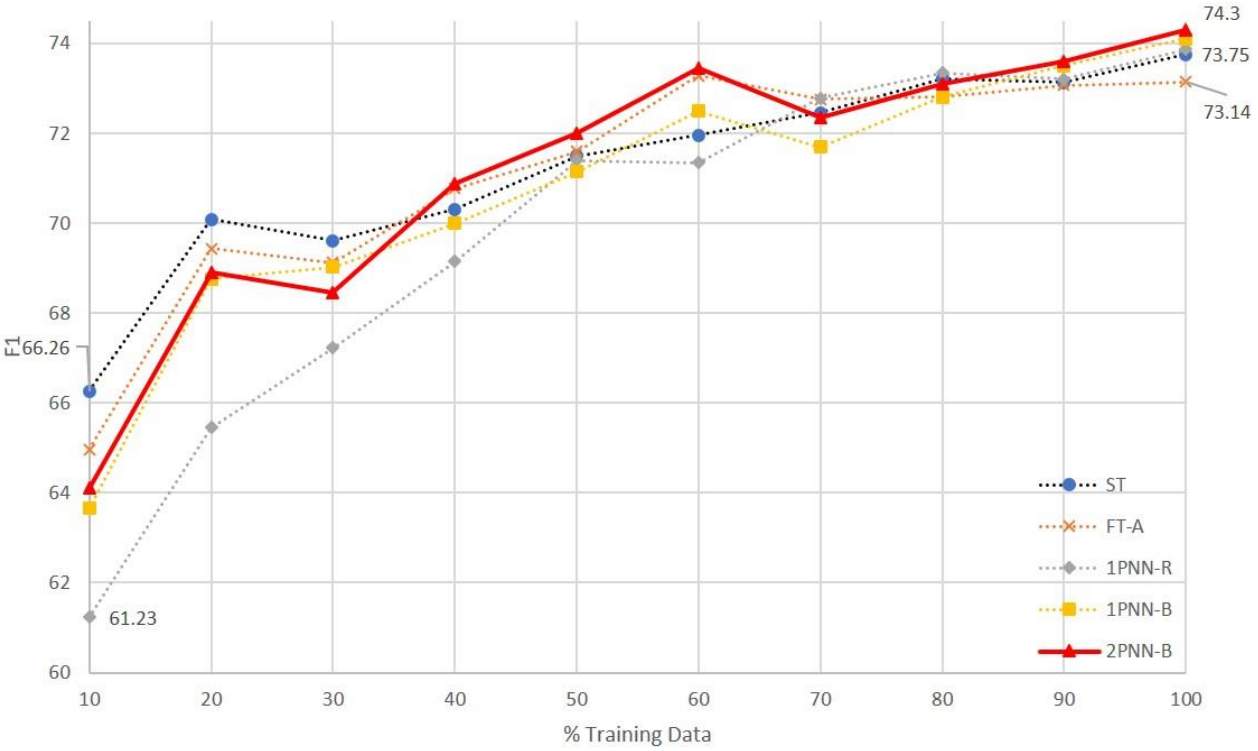


Graph shows comparison of various Transfer Learning techniques for varying proportion of the **Organic** training data (x-axis). The performance is evaluated in terms of the accuracy score (y-axis). **Transfer Learning helps in improving the performance when the training data is less by as much as 10%.** The best performing technique is shown in the **solid** line.

The target is the Organic dataset from the Sentiment Analysis task.

Research Questions

Q2. Can Transfer Learning still improve the performance when the training data is sufficiently large?



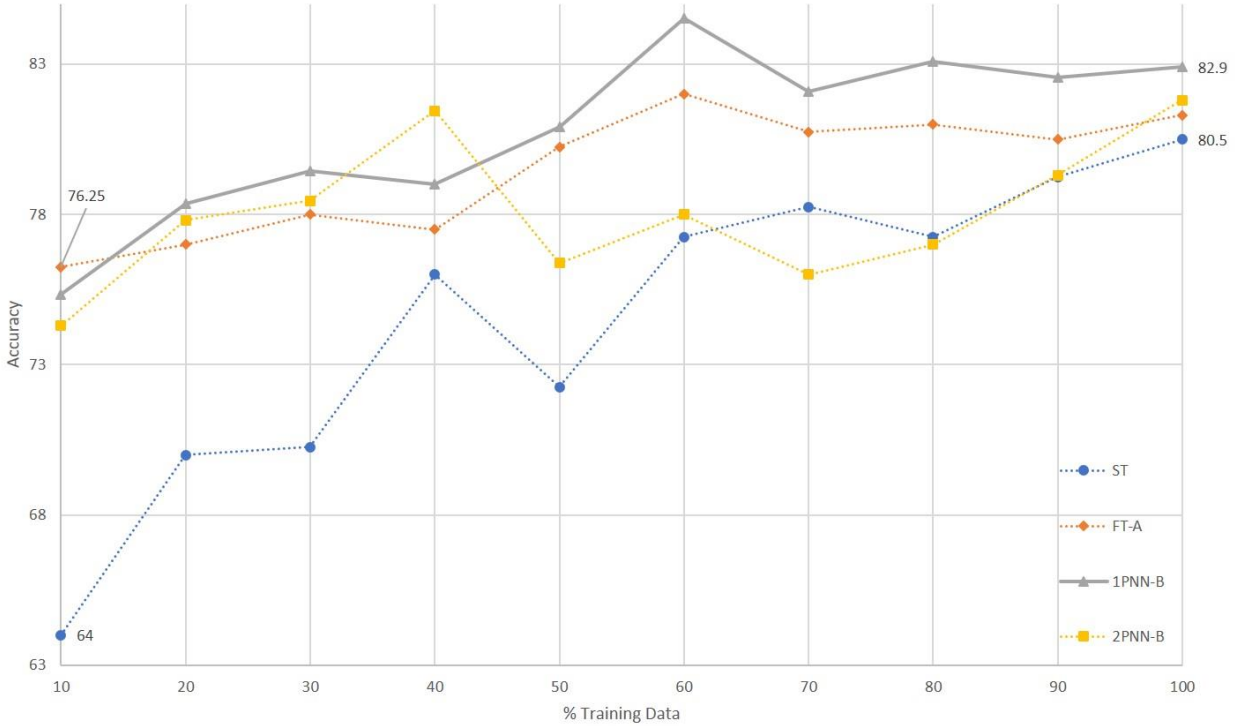
Graph shows comparison of various Transfer Learning techniques for varying proportion of the **JNLPBA** training data (x-axis). The performance is evaluated in terms of the accuracy score (y-axis). **The improvement in performance is not visible when the training data is sufficiently large.** The best performing technique is shown in the **solid** line.

The target is the JNLPBA dataset from the NER task.

Munich RE

Research Questions

Q3. Can Transfer Learning be applied under domain shift?

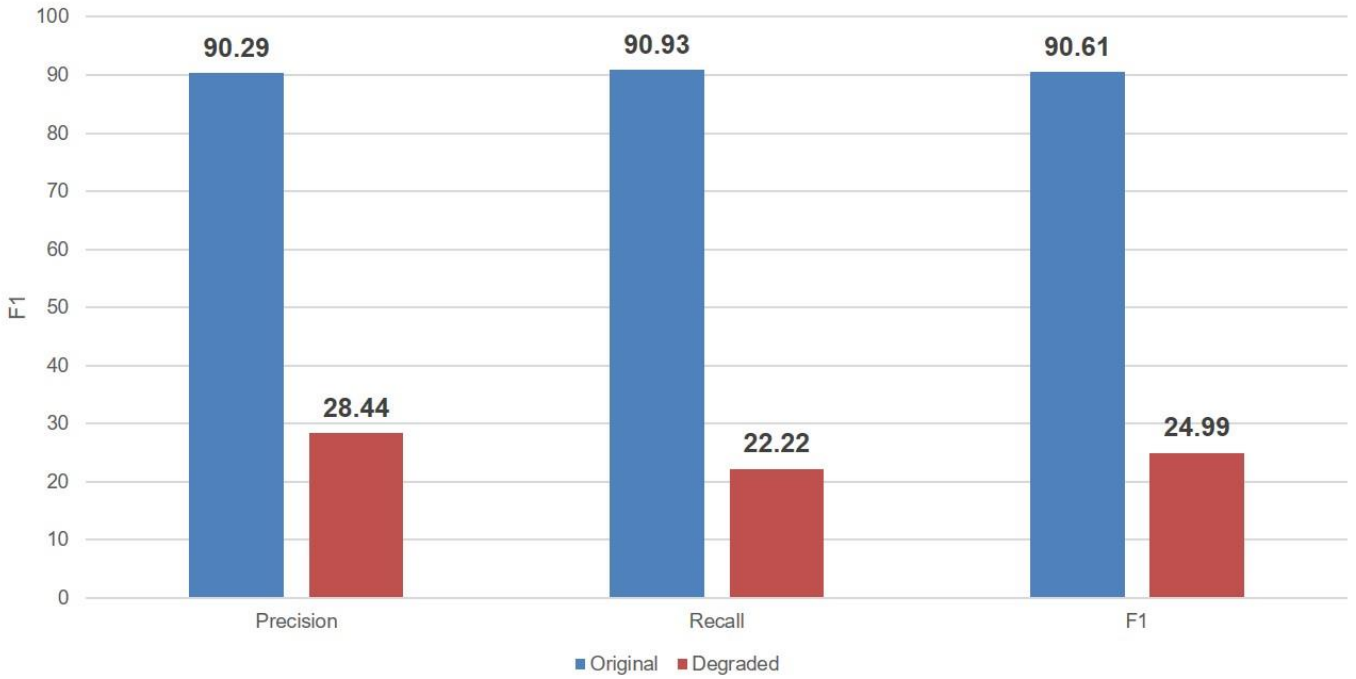


Graph shows comparison of various Transfer Learning techniques for varying proportion of the **Health** training data (x-axis). The performance is evaluated in terms of the accuracy score (y-axis). **The Transfer Learning helps in improving the performance under the Domain Shift.** The best performing technique is shown in the **solid** line.

The target is the Personal Healthcare dataset from the Sentiment Analysis task.

Research Questions

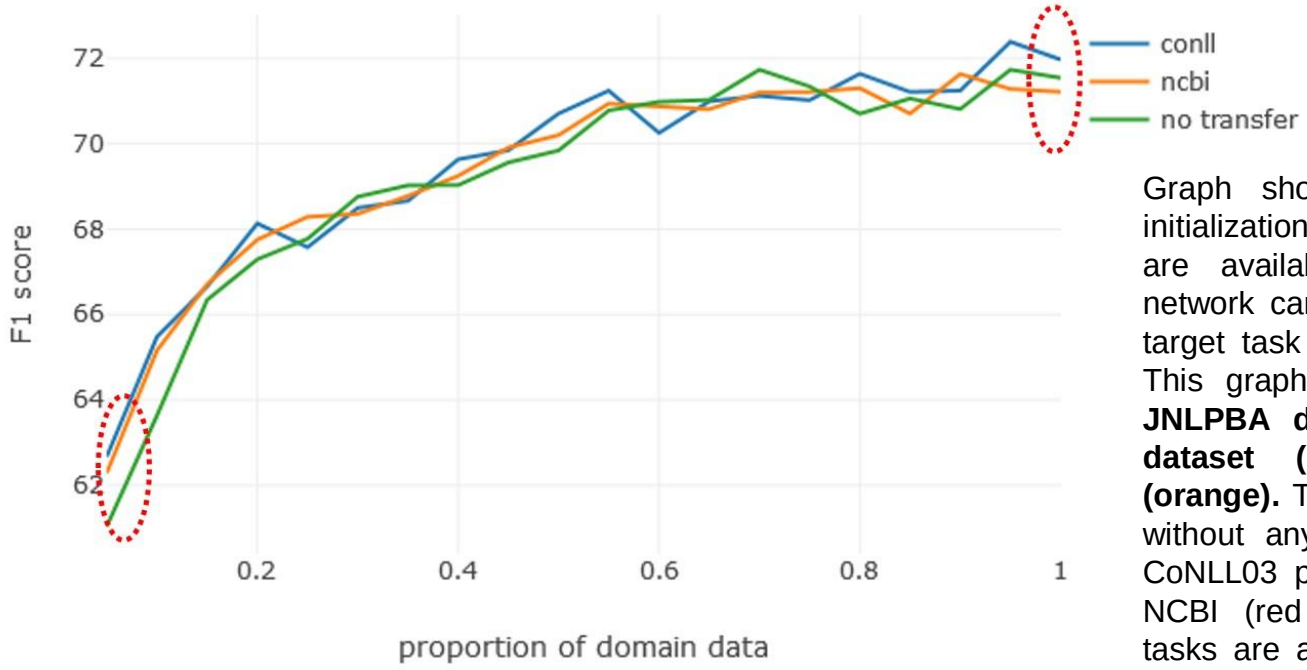
Q4. Is there any Catastrophic Forgetting in the source network when using PTFT?



Graph shows the Catastrophic Forgetting on CoNLL03 data after NER model trained on CoNLL03 data is fine-tuned on the JNLPBA dataset. **The performance degradation is approximately 70%.**

Research Questions

Q5. How to chose the source task in TL-A and TL-L settings?



Graph shows the problem of source initialization when multiple source networks are available. The choice of source network can vary the performance of the target task as shown by the red circles. This graph shows the **transfer to the JNLPBA dataset using the CoNLL03 dataset (blue) and NCBI dataset (orange).** The green line shows the score without any transfer. Here we see that CoNLL03 provides a better transfer than NCBI (red circles). If multiples source tasks are available, the task of choosing one becomes cumbersome.

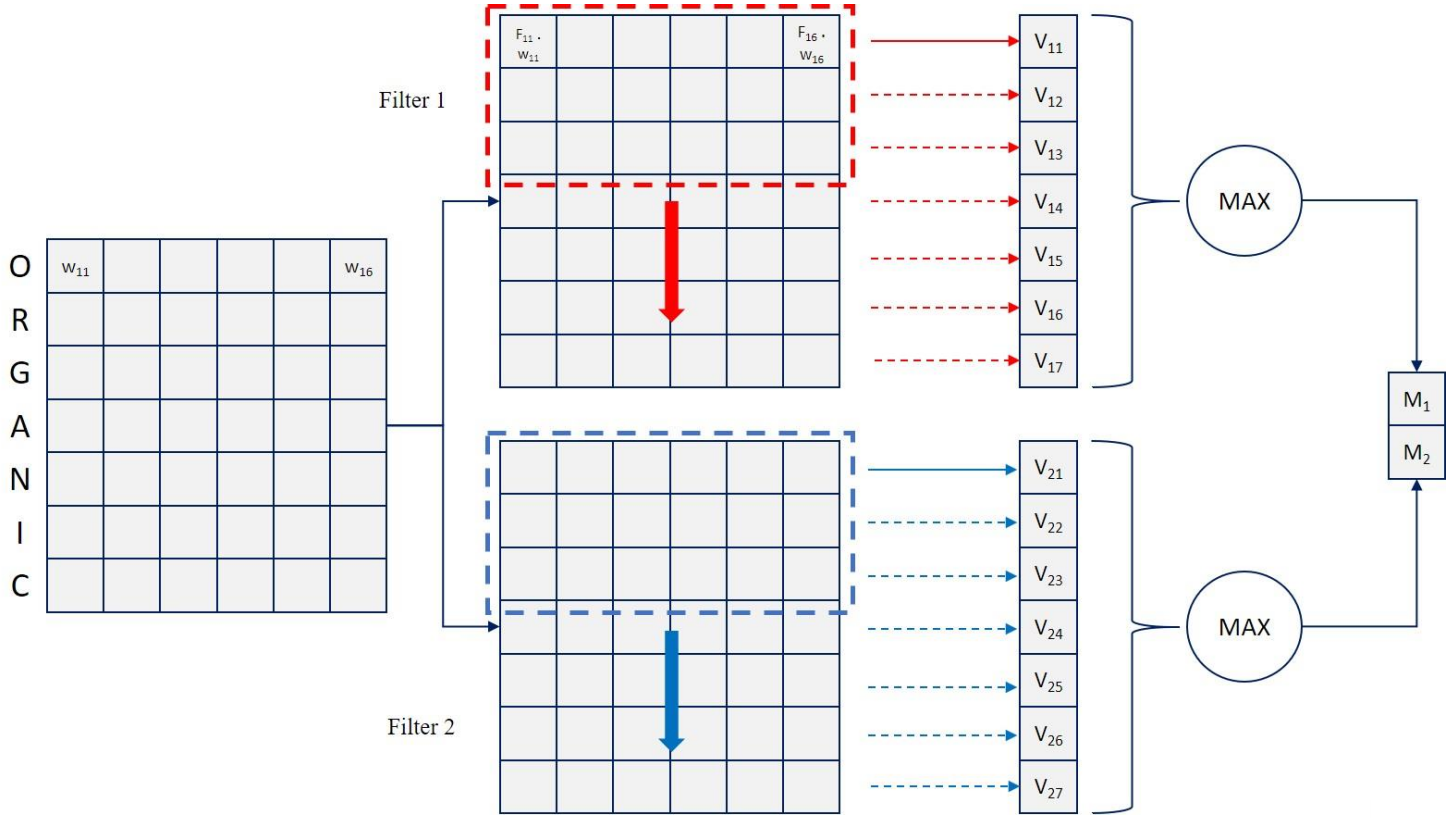
Munich RE

Research Questions

Q6. Can Progressive Neural Networks also be used for cross-task transfer?

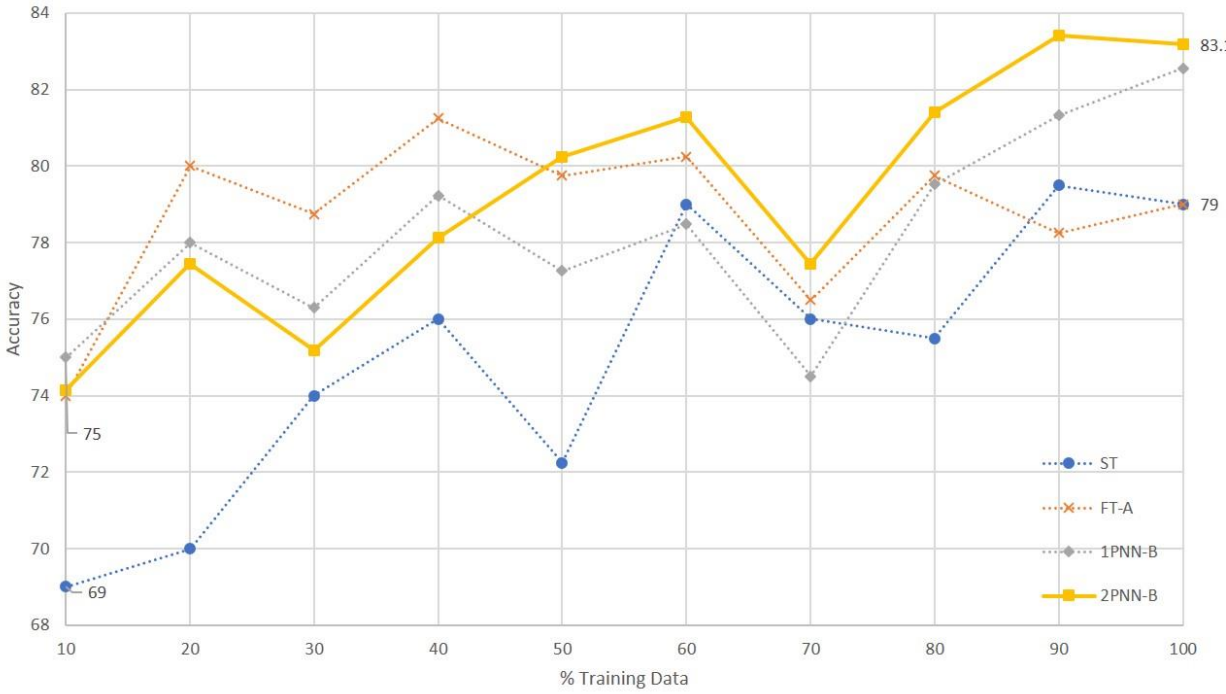
Research Questions

Q7. Does Character information help in improving the performance of NER?



Research Questions

Q8. Under what circumstances PNN should be used instead of TL-A



Graph shows comparison of various Transfer Learning techniques for varying proportion of the **Kitchen Housewares** training data (x-axis). The performance is evaluated in terms of the accuracy score (y-axis). **PNN is effective when the training data is sufficiently larger compared to the TL-A setting.** The best performing technique is shown in the **solid** line.

The target is the Kitchen Housewares dataset from the Sentiment Analysis task.

Munich RE

Discussion and Conclusion

- TL-A setup provides a benefit primarily in **slightly related domains.**
  - network is initialized with the weights from the source dataset that are near optimal for the target dataset which is true only when the source and the target domains are highly related.

- PNN can **avoid Negative Transfer**, whereas PTFT cannot
  - Adapters might block the information from source

- 2PNN-B often **performs better** than the 1PNN-B setting.
  - More network capacity
  - More information to adapt

Munich RE

Discussion and Conclusion

- 1PNN-R regularly **performs worse** than the 1PNN-B and 2PNN-B settings.
  - Only provides extra capacity which might not be needed by less complex tasks
  - No information to adapt
- The benefits of Transfer Learning are **more visible in the Sentiment Analysis** tasks than NER tasks
  - Sentiment Analysis dataset are smaller
- **PNN performs worse than TL-A at smaller splits** of training data (<30%)
  - Requires learning target network from scratch
  - Requires learning lateral weights in addition
- Cross-Task transfer between NER and Sentiment Analysis is **not YET successful**
  - Try using Language model as one of the tasks.

Munich RE

Contributions

1. **Outperform the best model** on the NCBI-Disease dataset using Transfer Learning reported by (Giorgi and Bader, 2018) which is one of the most established Transfer Learning studies for NER in the biomedical domain.
2. **The first-ever transfer from NER to Sentiment Analysis (vice-versa)** using Progressive Neural Network. This transfer cannot be realized using the PTFT technique.
3. **The first-ever domain adaptation on the Multi-Domain Sentiment Analysis** dataset using the Progressive Neural Network.

Munich RE

Limitations

1. Cross-Task transfer from NER to Sentiment Analysis (and vice-versa) does not improve results.
2. All combinations of Source and Target tasks / domains are not experimented with.
3. The source columns contains the same task.
4. MDSA dataset is not explored completely.
5. The approach is not compared to the Multi-Task Learning

Future Work

1. Use Language Model in the PNN architecture. Similar to [Peters et. al., 2018]
2. Expert Gates [Aljundi et. al., 2017]: Automatically find the task to execute at the inference time.
3. Elastic Weight Consolidation [James & Pascanu et. al., 2016]: Create a single network for learning the tasks continually by consolidating the information from new tasks in the existing network.
4. Domain Adversarial Learning [Ganin et. al., 2014]: Attach a Domain Classifier with a Gradient Reversal layer.

Munich RE

References

[1] Jia, Robin & Liang, Percy. (2017). Adversarial Examples for Evaluating Reading Comprehension Systems. 2021-2031. 10.18653/v1/D17-1215.
[2] Belinkov, Yonatan & Bisk, Yonatan. (2017). Synthetic and Natural Noise Both Break Neural Machine Translation.
[3] S. Ruder. (2019). Neural Transfer Learning for Natural Language Processing, National University of Ireland, Galway
[4] Rusu, Andrei & C. Rabinowitz, Neil & Desjardins, Guillaume & Soyer, Hubert & Kirkpatrick, James & Kavukcuoglu, Koray & Pascanu, Razvan & Hadsell, Raia. (2016). Progressive Neural Networks.
[5] Ma, Xuezhe & Hovy, Eduard. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. 1064-1074. 10.18653/v1/P16-1101.
[6] Kim, Yoon. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 10.3115/v1/D14-1181.
[7] Peters, Matthew & Neumann, Mark & Iyyer, Mohit & Gardner, Matt & Clark, Christopher & Lee, Kenton & Zettlemoyer, Luke. (2018). Deep Contextualized Word Representations. 2227-2237. 10.18653/v1/N18-1202.
[8] Aljundi, Rahaf & Chakravarty, Punarjay & Tuytelaars, Tinne. (2017). Expert Gate: Lifelong Learning with a Network of Experts. 7120-7129. 10.1109/CVPR.2017.753.
[9] James & Pascanu, Kirkpatrick, Razvan & Rabinowitz, Neil & Veness, Joel & Desjardins, Guillaume & Rusu, Andrei & Milan, Kieran & Quan, John & Ramalho, Tiago & Grabska-Barwinska, Agnieszka & Hassabis, Demis & Clopath, Claudia & Kumaran, Dharshan & Hadsell, Raia. (2016). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences. 114. 10.1073/pnas.1611835114.
[10] Ganin, Yaroslav & Lempitsky, Victor. (2014). Unsupervised Domain Adaptation by Backpropagation.
[11] Giorgi, John & Bader, Gary. (2018). Transfer learning for biomedical named entity recognition with neural networks. 10.1101/262790.

Munich RE