

# An Evaluation of Progressive Neural Networks for Transfer Learning in Natural Language Processing

Gerhard Hagerer<sup>1†</sup>, Abdul Moeed<sup>1†</sup>, Sumit Dugar<sup>1†</sup>, Sarthak Gupta<sup>1†</sup>, Mainak Ghosh<sup>1†</sup>  
Hannah Danner<sup>1</sup> Oliver Mitevski<sup>2</sup>, Andreas Nawroth<sup>2</sup>, Georg Groh<sup>1</sup>

<sup>1</sup> Technical University of Munich, <sup>2</sup> Munich Re

<sup>†</sup> These authors contributed equally.

gerhard.hagerer@tum.de

## Abstract

A major challenge in modern neural networks is the utilization of previous knowledge for new tasks in an effective manner, otherwise known as transfer learning. Fine-tuning, the most widely used method for achieving this, suffers from catastrophic forgetting. The problem is often exacerbated in natural language processing (NLP). In this work, we assess *progressive neural networks* (PNNs) as an alternative to fine-tuning. The evaluation is based on common NLP tasks such as sequence labeling and text classification. By gauging PNNs across a range of architectures, datasets, and tasks, we observe improvements over the baselines throughout all experiments.

**Keywords:** Document Classification, Text categorisation, Named Entity Recognition, Opinion Mining / Sentiment Analysis, Statistical and Machine Learning Methods, Other (Transfer Learning)

## 1. Introduction

Transfer learning is the ability of a model to generalize over previously unseen domains and/or tasks in a competent manner. The intuition is to re-use previously learned knowledge effectively when learning new tasks. The most common approaches to transfer learning include fine-tuning and multi-task learning (MTL). The former, where the weights of the already pre-trained layers are re-trained for a new task, performs well for similar tasks (Min et al., 2017) but fails to transfer over unrelated tasks (Mou et al., 2016). The latter adds terms to the objective function for each new task (Rei, 2017), which means re-training the whole model from scratch each time a new task is added (Chen et al., 2017).

A major problem faced by traditional transfer learning approaches is *catastrophic forgetting* (French, 1999) – a phenomenon where the model loses performance on previously learned tasks when trained on a new task. Catastrophic forgetting is thoroughly documented in artificial neural network literature and a few solutions have been proposed (Kirkpatrick et al., 2016; Awasthi and Sarawagi, 2019). The problem is more prevalent in NLP compared to computer vision; the shallow nature of networks used for NLP has been cited as a possible explanation for this discrepancy (Howard and Ruder, 2018).

Transfer learning approaches other than fine-tuning and MTL have also been explored for neural architectures (Hodas et al., 2017; Riemer et al., 2017). One such example – progressive neural networks (PNNs) (Rusu et al., 2016) – offers a novel solution to catastrophic forgetting. The idea is to train multiple networks – one for each new domain/task – that share information learned from previous tasks with each other through lateral connections. PNNs have gained popularity and have already been used for transfer learning in video summarization (Choi et al., 2018) and emotion recognition (Gideon et al., 2017).

The aim of this work is to evaluate PNNs for transfer learning in the context of various NLP tasks related to sequence labeling and text classification (Aggarwal and Zhai, 2012).

Specifically, three tasks are targeted: named entity recognition (NER), sentiment analysis (SA) and aspect-based sentiment analysis (ABSA). We perform a cross-task, cross-architecture comparison of PNNs with traditional transfer learning methods. The architectures differ significantly between these tasks; NER uses bi-directional LSTMs (BiLSTMs) with no convolutional or attention layers, SA uses convolutional layers and ABSA uses BiLSTMs with attention.

## 2. Background

### 2.1. Sequence labeling

NER falls under sequence labeling – the task of labeling each word in a sentence as a category (part of speech, entity etc.) Classically, solutions to NER have taken the form of Hidden Markov Models (Luo et al., 2015) and Conditional Random Fields (Lafferty et al., 2001; Passos et al., 2014). More recently, end-to-end neural techniques have been deployed that do not require task-specific knowledge (Ma and Hovy, 2016).

### 2.2. Text classification

Sentiment analysis is the task of classifying text according to the sentiment it exhibits. The sentiment labels are usually *positive*, *negative* and *neutral*. While sentiment analysis is well-studied in the NLP literature (Dave et al., 2003; Mäntylä et al., 2016) deep networks have nonetheless proved beneficial lately (Dos Santos and Gatti, 2014; Kim, 2014).

Closely related to sentiment analysis is aspect-based sentiment analysis (ABSA), which is more fine-grained. Here, the task is to find relevant aspects (e.g. product) in the text and detect their corresponding sentiments. Traditionally, aspect extraction has been treated as a secondary task for ABSA and the focus has been to classify the sentiment polarities of the aspects (Schouten and Frasincar, 2015; Lakkaraju et al., 2014). Recent work, including that of this paper, departs from

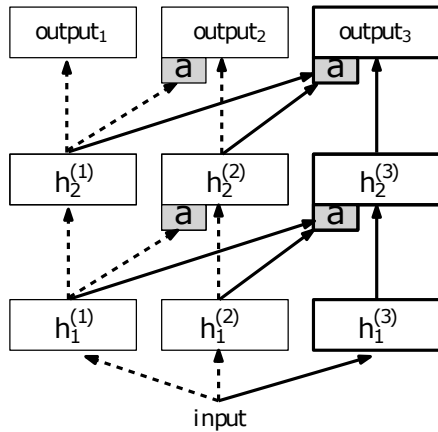


Figure 1: A three column progressive network. The first two columns each are trained on different tasks. The grey box represents the adapter layers. The third column is trained on the target task. Taken from (Rusu et al., 2016).

this formulation and treats aspect extraction as part of the ABSA task (Wojatzki et al., 2017; Schmitt et al., 2018).

### 2.3. Progressive Neural Networks

(Rusu et al., 2016) proposed progressive neural networks (PNNs) as a transfer learning technique for both cross-domain and multi-task purposes – see also (Gupta, 2019). The authors showed the effectiveness of PNNs on reinforcement learning tasks, with the technique demonstrating superior performance to pre-training and fine-tuning. The technique consists of adding lateral connections - coming from networks trained for source tasks - to the network being trained for the target task. Only the parameters of the target network are learned while the source weights are frozen. This ensures the immunity of PNNs to catastrophic forgetting.

The first part of a PNN is a neural network which is trained on the source task containing  $L$  hidden layers. This is called the first column with activations denoted as  $h_i^1$  of layer  $i$ . After the training of the first column is finished, a second so called target column after being initialized randomly is trained on the target task. The activations  $h_i^2$  of the second column are calculated based on the activations from the previous layer of the same column  $h_{i-1}^2$  and from the previous layer of the source column  $h_{i-1}^1$ . Therefore lateral connections between the layers of the source and target column are created. These connections are trained, too, whereas *the weights of the source column are not updated*.

The generalized mathematical formulation for multiple columns is

$$h_i^k = \sigma(U_i^{k:j} \sigma(V_i^{k:j} \alpha_i^{<k} h_{i-1}^{<k}) + W_i^k h_{i-1}^k) \quad (1)$$

where  $\sigma$  is the activation function,  $K$  is the number of columns,  $U_i^{j:k}$  is the weight matrix representing the lateral connections from column  $j$  to  $k$ , and  $W_i^k$  is the weight matrix of the  $i^{th}$  layer in column  $k^{th}$ .

In place of connecting the previous column directly by multiplying its activations  $h_{i-1}^1$  with  $U_i^{k:j}$ , a non-linear downprojection using matrix  $V_i^{k:j}$  is added. This concept,

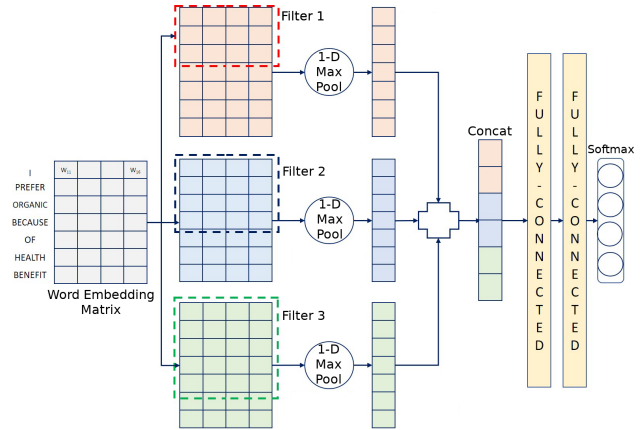


Figure 2: Model architecture for sentiment analysis (Gupta, 2019). The convolutional layer from the source column is passed to the first fully-connected layer (FCL) of the target column. The first FCL is then passed to the second FCL.

termed *adapter*, enhances the lateral connections and reduces the model complexity. The learnable scalar parameter  $\alpha_i$  scales the activations of the source task such that their order of magnitude fits to the target task.

## 3. Experiments

### 3.1. Transfer Learning

For each task mentioned in the paragraphs below, we evaluate transfer learning using progressive neural networks as they have been introduced above. As a baseline to show improvements using that technique, an appropriate neural network model is trained on that task. For each task there are at least two or more domains or sub-datasets given. Transfer learning is evaluated by training on a source domain and fine-tuning on the respective target domain of the same task. Therefore we apply normal fine-tuning of all layers (FT) and progressive neural networks with one (1PNN) and two (2PNN) source columns.

The latter is only done for the task of named entity recognition (NER), since, as shown in the results section, the increase in performance is small while the increase of the model complexity is big. For NER with 1PNN the best performing source column is chosen to be connected laterally to the target network.

For NER we further investigate the effect of catastrophic forgetting, i.e., the degradation of the prediction performance of a model which is firstly trained on the source task, then fine-tuned to a target task, and then again evaluated on the source task. Due to the fine-tuning on the target domain, a decrease of prediction accuracies is expected to happen due to the modification of the network weights. This is done to shed light upon what happens to the source networks during training on NLP tasks.

### 3.2. Named Entity Recognition

**Modeling** For all evaluations on named entity recognition (NER), micro F1 score is used as the metric. GloVe 100-D (Pennington et al., 2014) is used for word embeddings. We make a slight modification to (Ma and Hovy, 2016)’s architecture, having two LSTM layers instead of

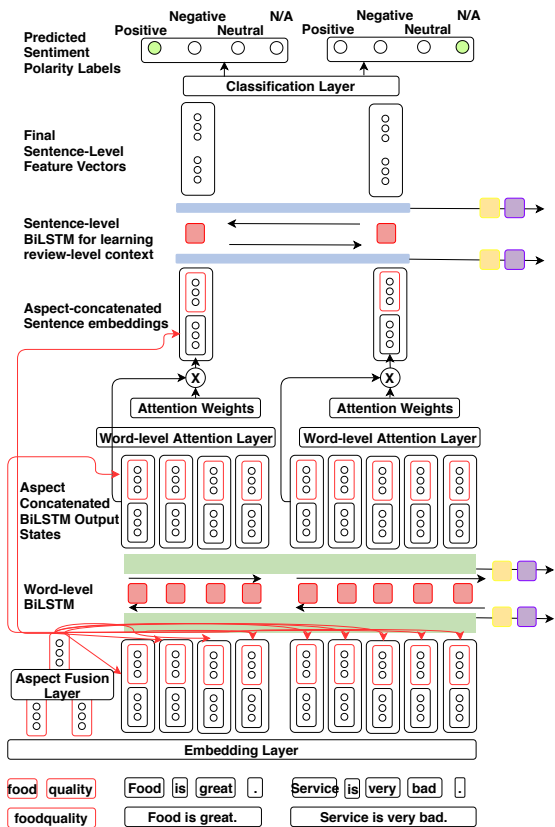


Figure 3: Model architecture for ABSA (Dugar, 2019). Green and blue rectangles denote lateral connections of BiLSTM layers on word and sentence representations. Red squares represent BiLSTM cells, yellow represent learnable scalar, purple represent the adapter layers.

the original’s one (Gupta, 2019). The architecture is illustrated in illustration 4.

Regarding the data, the experiments for named entity recognitions are executed on three different publicly available biomedical datasets as they are provided by (Crichton et al., 2017).

**BC5CDR Dataset** The BioCreative V Chemical-Disease Relation dataset (BC5CDR) is released along with the CDR task of the BioCreativ V challenge in 2015 (Li et al., 2016). The overall goal of the challenge is to find relations between chemicals and their associated diseases. Thus, the entity classes *chemical* and *disease* have been annotated manually which is the ground truth for NER in our experiments.

**NCBI Dataset** The NCBI Disease Corpus from 2014 aims at evaluating the task of disease name recognition. It comes with manual annotations for all mentioned diseases and according classifications based on 793 PubMed abstracts (Doğan et al., 2014). In our experiments, only the target entity *disease* is classified for each word.

**JNLPBA Dataset** This dataset was published for the JNLPBA challenge of bio-entity recognition in 2004. The data is based on the GENIA v3 named entity corpus of MEDLINE abstracts (Kim et al., 2004). The target classes in this dataset are *DNA*, *RNA*, *cell line*, *cell type*, and *protein*.

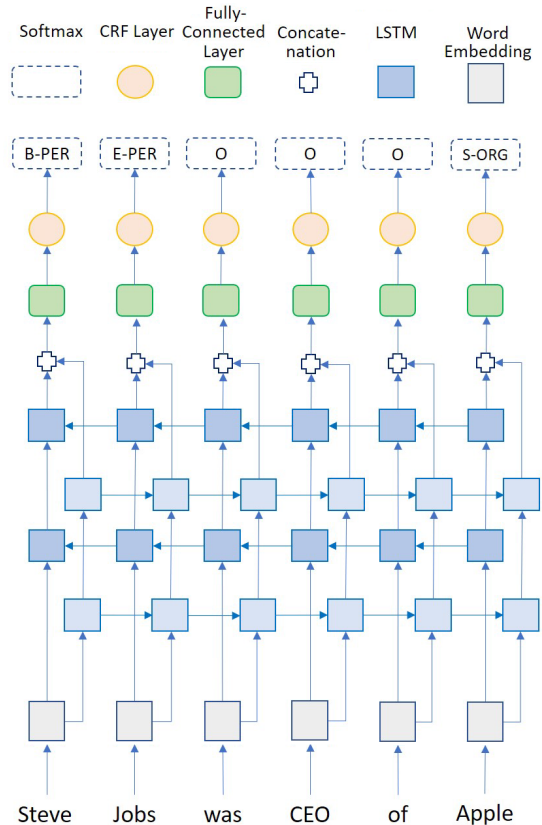


Figure 4: Model architecture for the NER experiments (Gupta, 2019). For the PNN, the first BiLSTM layer of the source column is passed to the second RNN layer of the target column, and the second RNN layer to the fully-connected layer analogously.

### 3.3. Sentiment Analysis

**Modeling** For all the evaluations of sentiment analysis, accuracy score is used as the evaluation metric. Similar to NER, the model is based on pre-trained GloVe 100-D embeddings. The model architecture is inspired by (Kim, 2014) which uses three one dimensional convolutional kernels with varying sizes in the first layer to capture local features as can be observed in figure 2.

**Amazon Dataset** As data the Amazon product review dataset as provided by (Blitzer et al., 2007) is used in the experiments. For transfer learning, the categories ‘kitchen houseware’ and ‘personal healthcare’ are considered. The annotated sentiment target classes are positive and negative.

### 3.4. Aspect-Based Sentiment Analysis

**Modeling** We also consider aspect-based sentiment analysis (ABSA) as a task for our experiments (Dugar, 2019). The utilized architecture is a hierarchical neural network as shown in figure 3 which is inspired by (Yang et al., 2016). It also uses GloVe word vectors BiLSTMs and attention (Wang et al., 2016), with a joint end-to-end formulation of ABSA similar to (Schmitt et al., 2018).

**SemEval Dataset** The dataset for ABSA is taken from the SemEval 2016 challenge task 5 subtask 1 (Pontiki et al., 2016). The subtask is defined as aspect extraction and sentiment polarity classification with regard to that aspect.

Category	%	Summary Label
sentiment	39%	neutral/ambiguous
	32%	positive
	29%	negative
entity	83%	organic
	11%	conventional
	5%	genetic engineering
attribute	33%	general
	28%	healthiness
	12%	trustworthiness
	11%	quality
	10%	environment
	6%	price

Table 1: Annotation distribution on the organic of all sentences to which at least one opinion triplet (entity+attribute+sentiment) was assigned, i.e., 53% of all 10,000 sentences. 668 of the annotated sentences contain two or more opinion triplets.

Task-Dataset	Train	Val	Test
NER-JNLPBA	16691	1853	3856
NER-BC5DR	5423	922	939
NER-NCBI	4559	4580	4796
SA-Amazon	2880	320	800
ABSA-SemEval-R	5654	106	106
ABSA-SemEval-L	55136	6892	6892
ABSA-Organic	8824	712	908

Table 2: Data splits of the named entity recognition (NER), sentiment analysis (SA) and aspect-based sentiment analysis (ABSA) tasks. SemEval-R and SemEval-L refer to the restaurant and laptop datasets respectively.

For our experiments, we solve the subtask as a whole by jointly classifying the aspect and its related sentiment. For the evaluation of transfer learning, we utilize both given domains, i.e., *laptops* and *restaurants*.

**Organic Dataset** One important goal of transfer learning is to improve performance on custom datasets of the respective target domain of interest. In that regard it is not always possible to provide many expert annotations of high reliability on noisy real world data. In that regard aspect-based sentiment analysis can be considered as an interesting use case due to reasons such as high number of classes, multi-labeling classification, and small number of annotated samples.

Therefore, we collected 10,000 social media comments from the well-known question-and-answer website Quora which contain opinions about *organic food* and related consumer issues. After being thoroughly instructed, each of 10 labelers annotated relevance, entity, attribute, and sentiment for 1000 sentences. Relevance is merely a binary flag to indicate if the sentence contains a relevant opinion. The other classes and their respective distributions are enlisted in table 1.

Transfer learning is evaluated by first training on the laptop and restaurant dataset jointly and then fine-tuning and evaluating on the organic dataset.

Tasks-Dataset	Baseline	FT	1PNN	2PNN
NER-JNLPBA	73.8	73.1	74.0	<b>74.3</b>
NER-BC5DR	81.7	83.7	<b>84.8</b>	84.0
NER-NCBI	84.9	85.0	85.6	<b>85.7</b>
SA-Kitchen	79.0	79.0	82.5	-
SA-Healthcare	80.5	81.1	<b>82.9</b>	-
ABSA-SemEval-R	39.4	32.4	<b>47.1</b>	-
ABSA-SemEval-L	24.6	22.9	<b>27.6</b>	-
ABSA-Organic	12.9	6.2	<b>17.0</b>	-

Table 3: Summary of the results across different tasks and datasets. NER and ABSA results are reported using micro F1 scores, while SA results use model accuracy. SemEval-R and SemEval-L correspond to the restaurant and laptop datasets of SemEval respectively. PNN outperforms both the baseline and fine-tuned models across all tasks and datasets. Bold entries indicate the best performing architecture for that row.

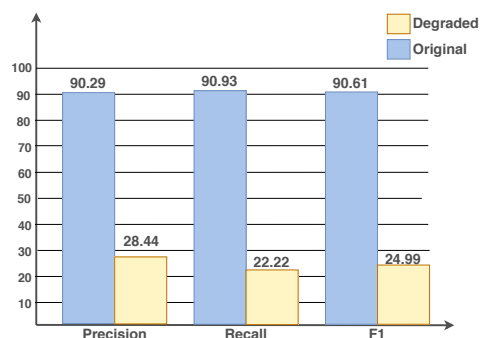


Figure 5: Graph showing catastrophic forgetting on the CoNLL03 domain after fine-tuning on JNLPBA.

## 4. Results

As a general overview, table 3 provides a summary of the results. It can clearly be seen that PNNs exceed the baseline and the standard fine-tuning approach for transfer learning throughout all tasks, domains, and architectures. This is denoted as bold number in table table 3.

### 4.1. Named Entity Recognition

For JNLPBA as the target domain, 2PNN fares marginally better than 1PNN. We train the source columns on the NCBI and BC5CDR datasets. Using NCBI as the target domain, 1PNN and 2PNN are comparable. Finally, 1PNN outperforms all other transfer techniques for BC5CDR as target domain. Varying the source datasets between JNLPBA and NCBI does not change performance in any significant manner (Gupta, 2019).

### 4.2. Sentiment Analysis

Similar to the results of NER, PNNs outperform fine-tuning the model. Fine-tuning yields results not too dissimilar to the baseline (Gupta, 2019).

### 4.3. Aspect-based Sentiment Analysis

Results for ABSA are no different; PNN surpasses fine-tuning notably in terms of performance. The performance gain of PNN is varied, however, and depends on the domain. For the 'Restaurant' dataset as target, PNN achieves

a micro F1 score of 47.1% vs. 32.4% achieved by fine-tuning. The second experiment, however, does not show as remarkable a difference, with micro F1 scores of 27.6% vs 22.9% for PNN and fine-tuning respectively.

In our experiments, PNNs with two source columns are not evaluated for ABSA; we hypothesize, however, that similar to NER 2PNN performs at least as well as 1PNN in most cases (Dugar, 2019).

#### 4.4. Catastrophic forgetting

As demonstrated in figure 5, we confirm the occurrence of catastrophic forgetting for NER. Initially, the model is trained on the CoNLL03 (Sang and De Meulder, 2003) dataset. After being subsequently fine-tuned on JNLPBA, the model's performance is crippled on the original domain. A performance degradation of approximately 70% can be observed (Gupta, 2019).

### 5. Conclusion

Transfer learning ensures a learning algorithm's ability to generalize over new domains and tasks in a competent manner. In this paper, we evaluate progressive neural networks as a transfer learning approach with reference to natural language processing tasks. We observe that progressive networks consistently outperform the conventional transfer technique of fine-tuning the network on named entity recognition, sentiment analysis, and aspect-based sentiment analysis. We further observe that PNNs with two source networks and according lateral connections produce marginally better results than with a single source network.

### 6. Acknowledgements

We are thanking Munich Re AG for their support of the experiments regarding named entity recognition and sentiment analysis.

### 7. Bibliographical References

- Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer.
- Awasthi, A. and Sarawagi, S. (2019). Continual learning with neural networks: A review. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 362–365. ACM.
- Blitzer, J., Dredze, M., and Pereira, F. (2007). Domain adaptation for sentiment classification. In *45th Annv. Meeting of the Assoc. Computational Linguistics (ACL'07)*.
- Chen, Z., Badrinarayanan, V., Lee, C.-Y., and Rabinovich, A. (2017). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. *arXiv preprint arXiv:1711.02257*.
- Choi, J., Oh, T.-H., and Kweon, I. S. (2018). Contextually customized video summaries via natural language. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1718–1726. IEEE.
- Crichton, G., Pyysalo, S., Chiu, B., and Korhonen, A. (2017). A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368.
- Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.
- Doğan, R. I., Leaman, R., and Lu, Z. (2014). Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.
- Dos Santos, C. and Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78.
- Dugar, S. (2019). Aspect-based sentiment analysis using deep neural networks and transfer learning, 3. Master's thesis under supervision of Gerhard Hagerer, M.Sc, and PD Dr. Georg Groh.
- French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.
- Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., and Provost, E. M. (2017). Progressive neural networks for transfer learning in emotion recognition. *CoRR*, abs/1706.03256.
- Gupta, S. (2019). Neural transfer learning for natural language processing, 4. Master's thesis under supervision of Gerhard Hagerer, M.Sc, and PD Dr. Georg Groh.
- Hodas, N. O., Shaffer, K., Yankov, A., Corley, C. D., Anderson, A., and Cheney, W. (2017). Beyond fine tuning: Adding capacity to leverage few labels. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States).
- Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004). Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Citeseer.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2016). Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.
- Lafferty, J. D., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.
- Lakkaraju, H., Socher, R., and Manning, C. (2014). Aspect specific sentiment analysis using hierarchical deep learning. In *NIPS Workshop on deep learning and representation learning*.

- Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wieggers, T. C., and Lu, Z. (2016). Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.
- Luo, G., Huang, X., Lin, C.-Y., and Nie, Z. (2015). Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 879–888, Lisbon, Portugal, September. Association for Computational Linguistics.
- Ma, X. and Hovy, E. H. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354.
- Mäntylä, M. V., Graziotin, D., and Kuuttila, M. (2016). The evolution of sentiment analysis - A review of research topics, venues, and top cited papers. *CoRR*, abs/1612.01556.
- Min, S., Seo, M., and Hajishirzi, H. (2017). Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171*.
- Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., and Jin, Z. (2016). How transferable are neural networks in nlp applications? *arXiv preprint arXiv:1603.06111*.
- Passos, A., Kumar, V., and McCallum, A. (2014). Lexicon infused phrase embeddings for named entity resolution. *CoRR*, abs/1404.5367.
- Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Pontiki, M., Galanis, D., Papageorgiou, H., Androutopoulos, I., Manandhar, S., Mohammad, A.-S., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.
- Rei, M. (2017). Semi-supervised multitask learning for sequence labeling. *arXiv preprint arXiv:1704.07156*.
- Riemer, M., Khabiri, E., and Goodwin, R. (2017). Representation stability as a regularizer for improved text analytics transfer learning. *arXiv preprint arXiv:1704.03617*.
- Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. (2016). Progressive neural networks. *arXiv preprint arXiv:1606.04671*.
- Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.
- Schmitt, M., Steinheber, S., Schreiber, K., and Roth, B. (2018). Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. *CoRR*, abs/1808.09238.
- Schouten, K. and Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.
- Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, November. Association for Computational Linguistics.
- Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., and Biemann, C. (2017). Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. *Proceedings of the GermEval*, pages 1–12.
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June. Association for Computational Linguistics.