Transfer Learning for Named-Entity Recognition with Neural Networks

Ji Young Lee[*1], **Franck Dernoncourt**[*1,2], **Peter Szolovits**[1]
[1]MIT, [2]Adobe Research
jjylee@mit.edu, dernonco@adobe.com, psz@mit.edu

Abstract

Recent approaches based on artificial neural networks (ANNs) have shown promising results for named-entity recognition (NER). In order to achieve high performances, ANNs need to be trained on a large labeled dataset. However, labels might be difficult to obtain for the dataset on which the user wants to perform NER: label scarcity is particularly pronounced for patient note de-identification, which is an instance of NER. In this work, we analyze to what extent transfer learning may address this issue. In particular, we demonstrate that transferring an ANN model trained on a large labeled dataset to another dataset with a limited number of labels improves upon the state-of-the-art results on two different datasets for patient note de-identification.

Keywords: named-entity recognition, neural networks, transfer learning

1. Introduction

Electronic health records (EHRs) have been widely adopted in some countries such as the United States and represent gold mines of information for medical research. The majority of EHR data exist in unstructured form such as patient notes (Murdoch and Detsky, 2013). Applying natural language processing on the unstructured data in conjunction with analyzing the other EHR data can lead to a better understanding of health and diseases (Liao et al., 2015), and a more accurate phenotyping of patients to compare tests and treatments (Ananthakrishnan et al., 2013; Pivovarov and Elhadad, 2015; Halpern et al., 2016).

However, before patient notes can be shared with medical investigators, some types of information, referred to as protected health information (PHI), must be removed in order to preserve patient confidentiality. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) (Office for Civil Rights, 2002) defines 18 different types of PHI, ranging from patient names and ID numbers to addresses and phone numbers. The task of removing PHI from a patient note is referred to as *de-identification*. The essence of de-identification is recognizing PHI in patient notes, which is a form of named-entity recognition (NER).

Existing de-identification systems are often rule-based approaches or feature-based machine learning approaches. However, these techniques require additional lead time for developing and fine-tuning the rules or features specific to each new dataset. Meanwhile, recent work using ANNs have yielded state-of-the-art performances without using any manual features (Dernoncourt et al., 2016). Compared to the previous systems, ANNs have a competitive advantage that the model can be fine-tuned on a new dataset without the overhead of manual feature development, as long as some labels for the dataset are available.

However, it may still be inefficient to mass deploy ANN-based de-identification system in practical settings, since creating annotations for patient notes is especially difficult. This is due to the fact that only a restricted set of individuals is authorized to access original patient notes; the annotation task cannot be crowd-sourced, making it slow and expensive to obtain a large annotated corpus. Medical professionals are therefore wary to explore patient notes because of this de-identification barrier, which considerably hampers medical research.

In this paper, we analyze to what extent transfer learning may improve de-identification performances on datasets with a limited number of labels. By training an ANN model on a large dataset (MIMIC) and transferring it to smaller datasets (i2b2 2014 and i2b2 2016), we demonstrate that transfer learning allows to outperform the state-of-the-art results.

2. Related Work

Transfer learning has been studied for a long time. There is no standard definition of transfer learning in the literature (Li, 2012). We follow the definition from (Pan and Yang, 2010): transfer learning aims at performing a task on a target dataset using some knowledge learned from a source dataset. The idea has been applied to many fields such as speech recognition (Wang and Zheng, 2015) and finance (Stamate et al., 2015).

The successes of ANNs for many applications over the last few years have escalated the interest in studying transfer learning for ANNs. In particular, much work has been done for computer vision (Yosinski et al., 2014; Oquab et al., 2014; Zeiler and Fergus, 2014). In these studies, some of the parameters learned on the source dataset are used to initialize the corresponding parameters of the ANNs for the target dataset.

Fewer studies have been performed on transfer learning for ANN-based models in the field of natural language processing. For example, Mou et al. (2016) focused on transfer learning with convolutional neural networks for sentence classification. To the best of our knowledge, no study has analyzed transfer learning for ANN-based models in the context of NER.

[*] These authors contributed equally to this work.

3. Model

The model we use for transfer learning experiments is based on a type of recurrent neural networks called long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), and utilizes both token embeddings and character embeddings. It comprises six major components:

1. **Token embedding layer** maps each token to a token embedding.
2. **Character embedding layer** maps each character to a character embedding.
3. **Character LSTM layer** takes as input character embeddings and outputs a single vector that summarizes the information from the sequence of characters in the corresponding token.
4. **Token LSTM layer** takes as input a sequence of token vectors, which are formed by concatenating the outputs of the token embedding layer and the character LSTM layer, and outputs a sequence of vectors.
5. **Fully connected layer** takes the output of the token LSTM layer as input, and outputs vectors containing the scores of each label for the corresponding tokens.
6. **Sequence optimization layer** takes the sequence of vectors from the output of the fully connected layer and outputs the most likely sequence of predicted labels, by optimizing the sum of unigram label scores as well as bigram label transition scores.

Figure 1 shows how these six components are interconnected to form the model. All layers are learned jointly using stochastic gradient descent. For regularization, dropout is applied before the token LSTM layer, and early stopping is used on the development set with a patience of 10 epochs.

4. Experiments

4.1. Datasets

We use three de-identification datasets for the transfer learning experiments: MIMIC, i2b2 2014, and i2b2 2016. The MIMIC de-identification dataset was introduced in (Dernoncourt et al., 2016), and is a subset of the MIMIC-III dataset (Johnson et al., 2016; Goldberger et al., 2000; Saeed et al., 2011).The i2b2 2014 and 2016 datasets were released as part of the 2014 i2b2/UTHealth shared task Track 1 (Stubbs et al., 2015) and the 2016 i2b2 CEGS N-GRID shared task, respectively. Table 1 presents the datasets' sizes.

	MIMIC	i2b2 2014	i2b2 2016
Vocabulary size	69,525	46,803	61,503
Number of notes	1,635	1,304	1,000
Number of tokens	2,945,228	984,723	2,689,196
Number of PHI instances	60,725	28,867	41,142
Number of PHI tokens	78,633	41,355	54,420

Table 1: Overview of the MIMIC and i2b2 datasets. PHI stands for protected health information.

4.2. Transfer learning

The goal of transfer learning is to leverage the information present in a source dataset to improve the performance of an algorithm on a target dataset. In our setting, we apply transfer learning by training the parameters of the ANN model on the source dataset (MIMIC), and using the same ANN to retrain on the target dataset (i2b2 2014 or 2016) for fine-tuning. We use MIMIC as the source dataset since it is the dataset with the most labels. We perform two sets of experiments to gain insights on how effective transfer learning is and which parameters of the ANN are the most important to transfer.[1]

Experiment 1 Quantifying the impact of transfer learning for various train set sizes of the target dataset. The primary purpose of this experiment is to assess to what extent transfer learning improves the performances on the target dataset. We experiment with different train set sizes to understand how many labels are needed for the target dataset to achieve reasonable performances with and without transfer learning.

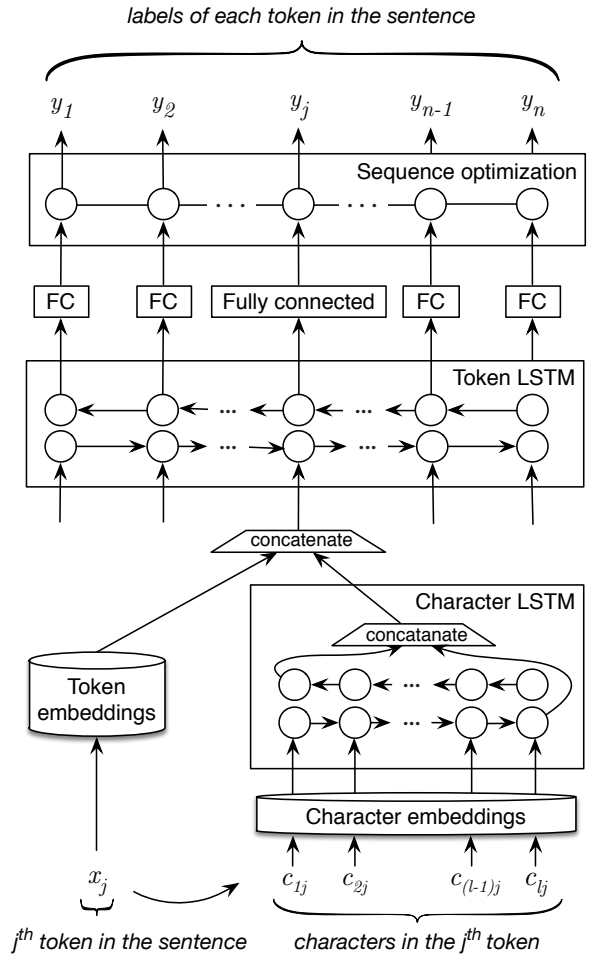

Figure 1: ANN model for NER. For transfer learning experiments, we train the parameters of the model on a source dataset, and transfer all or some of the parameters to initialize the model for training on a target dataset.

[1] Code: https://github.com/Franck-Dernoncourt/NeuroNER

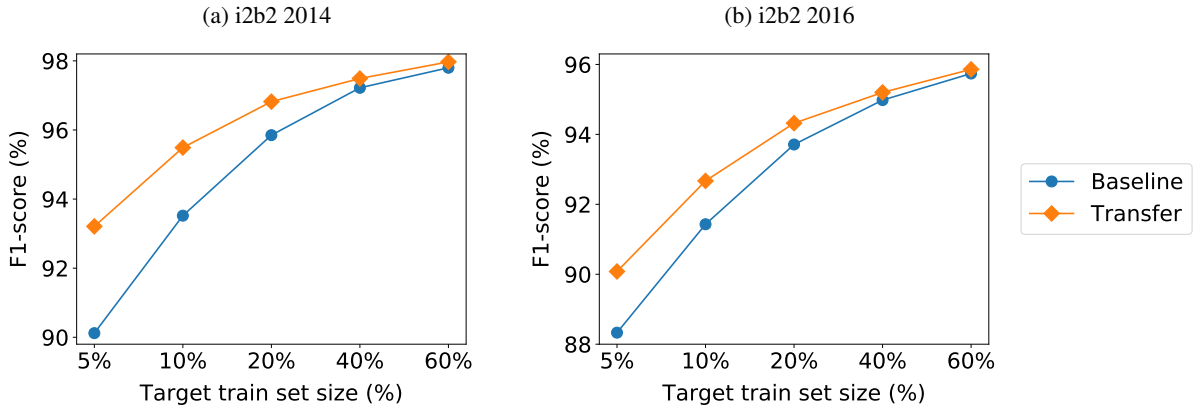

Figure 2: Impact of transfer learning on the F1-scores. Baseline corresponds to training the ANN model only with the target dataset, and transfer learning corresponds to training on the source dataset followed by training on the target dataset. The target train set size is the percentage of train set in the whole dataset, and 60% corresponds to the full official train set.

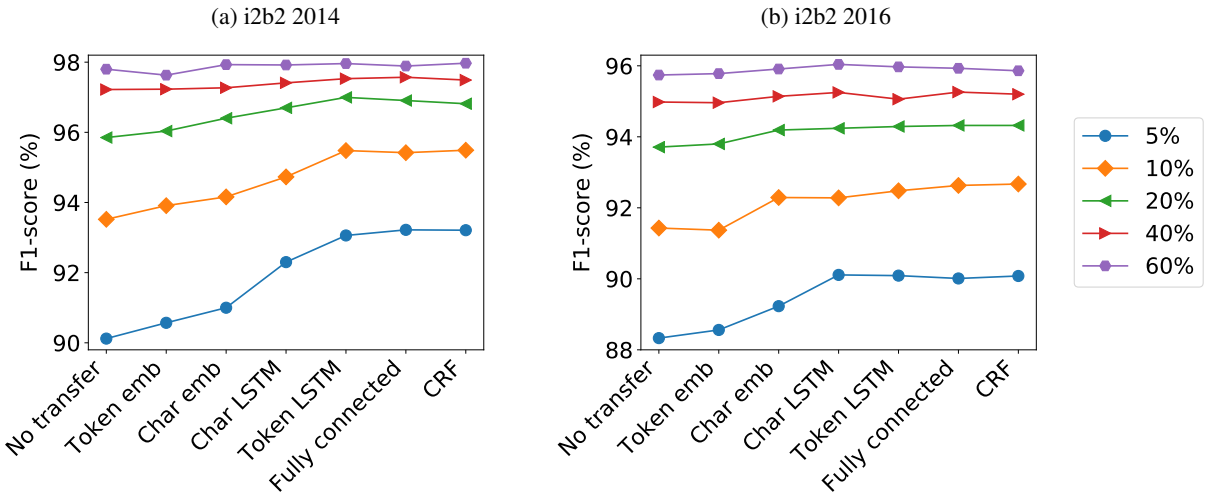

Figure 3: Impact of transferring the parameters up to each layer of the ANN model using various train set sizes on the target dataset: 5%, 10%, 20%, 40%, and 60% (official train set). "CRF" refers to the sequence optimization layer in Figure 1.

Experiment 2 Analyzing the importance of each parameter of the ANN in the transfer learning. Instead of transferring all the parameters, we experiment with transferring different combinations of parameters. The goal is to understand which components of the ANN are the most important to transfer. The lowest layers of the ANN tend to represent task-independent features, whereas the topmost layers are more task-specific. As a result, we try transferring the parameters starting from the bottommost layer up to the topmost layer, adding one layer at a time.

5. Results

Experiment 1 Figure 2 compares the F1-scores of the ANN trained only on the target dataset against the ANN trained on the source dataset followed by the target dataset. Transfer learning improves the F1-scores over training only with the target dataset, though the improvement diminishes as the number of training samples used for the target dataset increases. This implies that the representations learned from the source dataset are efficiently transferred and exploited for the target dataset.

Therefore, when transfer learning is adopted, fewer annotations are needed to achieve the same level of performance as when the source dataset is unused. For example, on the i2b2 2014 dataset, performing transfer learning and using 16% of the i2b2 train set leads to similar performance as not using transfer learning and using 34% of the i2b2 train set. Transfer learning thus allows to cut by half the number of labels needed on the target dataset in this case.

For both the i2b2 2014 and 2016 datasets, the performance gains from transfer learning are greater when the train set size of the target dataset is small. The largest improvement can be observed for i2b2 2014 when using 5% of the dataset as the train set (consisting of around 2k PHI tokens out of 50k tokens), where transfer learning increases the F1-score by around 3.1 percent point, from 90.12 to 93.21. Even when all of the train set is used, the F1-score improves when using transfer learning, albeit by just 0.17 percent point, from 97.80 to 97.97.

Experiment 2 Figure 3 shows the importance of each layer of the ANN in transfer learning. We observe that transferring a few lower layers is almost as efficient as transferring all layers. For i2b2 2014, transferring up to the token LSTM shows great improvements for each layer,

but there is less improvement for each added layer beyond that. For i2b2 2016, larger improvements can be observed up to the character LSTM and less so beyond that layer.

The parameters in the lower layers therefore seems to contain most information that are relevant to the de-identification task in general, which supports the common hypothesis that higher layers of ANN architectures contain the parameters that are more specific to the task as well as the dataset used for training.

Despite the observation that transferring a few lower layers may be sufficient for efficient transfer learning, it is interesting to see that adding the topmost layers to the transfer learning does not hurt the performance. When retraining the model on the target dataset, the ANN is able to adapt to the target dataset quite well despite some the higher layers being initialized to parameters that are likely to be more specific to the source dataset.

6. Conclusion

In this work, we have studied transfer learning with ANNs for NER, specifically patient note de-identification, by transferring ANN parameters trained on a large labeled dataset to another dataset with limited human annotations. We demonstrated that transfer learning improves the performance over the state-of-the-art results on two datasets. Transfer learning may be especially beneficial for a target dataset with small number of labels.

7. Bibliographical References

Ananthakrishnan, A. N., Cai, T., Savova, G., Cheng, S.-C., Chen, P., Perez, R. G., Gainer, V. S., Murphy, S. N., Szolovits, P., Xia, Z., et al. (2013). Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflammatory bowel diseases*, 19(7):1411.

Dernoncourt, F., Lee, J. Y., Uzuner, O., and Szolovits, P. (2016). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, page ocw156.

Halpern, Y., Horng, S., Choi, Y., and Sontag, D. (2016). Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, page ocw011.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Li, Q. (2012). Literature survey: domain adaptation algorithms for natural language processing. *Department of Computer Science The Graduate Center, The City University of New York*, pages 8–10.

Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., Karlson, E. W., Ananthakrishnan, A. N., Gainer, V. S., Shaw, S. Y., Xia, Z., Szolovits, P., et al. (2015). Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *bmj*, 350:h1885.

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*.

Murdoch, T. B. and Detsky, A. S. (2013). The inevitable application of big data to health care. *Jama*, 309(13):1351–1352.

Office for Civil Rights, H. (2002). Standards for privacy of individually identifiable health information. final rule. *Federal Register*, 67(157):53181.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Pivovarov, R. and Elhadad, N. (2015). Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Stamate, C., Magoulas, G. D., and Thomas, M. S. (2015). Transfer learning approach for financial applications. *arXiv preprint arXiv:1509.02807*.

Wang, D. and Zheng, T. F. (2015). Transfer learning for speech and language processing. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*, pages 1225–1237. IEEE.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

8. Language Resource References

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Johnson, A. E. W., Pollard, T. J., Shen, L., wei Lehman, L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, (in press).

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. (2011). Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. *Critical care medicine*, 39(5):952.

Stubbs, A., Kotfila, C., and Uzuner, Ö. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.