Transfer Learning for Entity Recognition of Novel Classes

Juan Diego Rodriguez, Adam Caldwell and **Alexander Liu**
Applied Research Laboratories
The University of Texas at Austin
Austin, Texas 78713
{juan.rodriguez, adam.caldwell, aliu}@arlut.utexas.edu

Abstract

In this reproduction paper, we replicate and extend several past studies on transfer learning for entity recognition. In particular, we are interested in entity recognition problems where the class labels in the source and target domains are different. Our work is the first direct comparison of these previously published approaches in this problem setting. In addition, we perform experiments on seven new source/target corpus pairs, nearly doubling the total number of corpus pairs that have been studied in all past work combined. Our results empirically demonstrate when each of the published approaches tends to do well. In particular, simpler approaches often work best when there is very little labeled target data, while neural transfer approaches tend to do better when there is more labeled target data.

1 Introduction

Transfer learning methods can play an important role in improving the performance of machine learning algorithms in scenarios where in-domain data is scarce, such as entity extraction of protected health information in electronic health records. These methods make use of data in a *source* domain to improve performance in a *target* domain. Many transfer learning techniques have been developed for tasks such as named entity recognition (Daumé III, 2007) and sentiment classification (Blitzer et al., 2007). A broad taxonomy of transfer learning scenarios in machine learning (only some of which are applied to NLP tasks) is given in (Zhang et al., 2017).

In this paper we focus on the entity recognition task[1] in which both the source and target domains have labeled data, but the class label sets are different. As is common in the literature, we frame entity recognition as a sequential classification problem, and restrict ourselves to local (sentence-level) supervised techniques.

Most transfer learning methods have been developed under the assumption that the source and target domains have the same class labels (e.g., (Jiang and Zhai, 2007; Daumé III, 2007)), and many NER corpora only annotate a small number of categories (e.g., 4 labels for the CoNLL 2003 corpus (Sang and Meulder, 2003) and 7 labels for the MUC 6 corpus (Grishman and Sundheim, 1996)). In many situations one would like to recognize categories that were not labeled in a given corpus, but that may be semantically related. This includes, for instance, recognizing companies or sports teams (Ritter et al., 2011), rather than organizations in general, or recognizing actors (Liu et al., 2013b), rather than people in general.

In addition, techniques for transfer learning with different class labels may be useful even when the class label set is the same across domains. As Florian et al. (2004) point out, the same labels may be used in very different ways across corpora due to different annotation guidelines. For example, determiners may or may not be annotated as part of an entity, and annotations may or may not include nominal as well as named entities.

[1] Although some researchers use the terms *entity recognition* and *named entity recognition* (NER) interchangeably, we shall refer to the task of recognizing named and nominal entities as entity recognition.

Proceedings of the 27th International Conference on Computational Linguistics, pages 1974–1985
Santa Fe, New Mexico, USA, August 20-26, 2018.

To the best of our knowledge, a direct comparison of existing transfer learning techniques for entity recognition with different labels does not exist. The results in the literature are hard to compare directly for several reasons: in some cases the datasets used are not publicly available, different evaluation measures are used for evaluation (macro-averaged vs micro-averaged scores, for instance), and in some cases, crucial implementation details are missing.

In this paper we replicate and extend several transfer learning methods for recognizing novel entities, using both standard corpora and datasets that have not been used for this task before. To the best of our knowledge, this is the first direct comparison of these transfer learning methods. In addition, we perform experiments on seven new source/target domain corpus pairs, nearly doubling the total number of corpus pairs that have been studied in all past work combined. We also share our code[2] so that others may verify our results and compare future transfer learning techniques against our benchmarks.

2 Transfer Learning for Novel Classes

2.1 Transfer learning scenarios

It is helpful to distinguish between three different scenarios for the transfer learning problem with different class labels:

1. The source domain entities are more fine-grained than the target domain entities and source domain entities can be mapped to target domain entities.

2. The source domain entities are generally coarser than the target domain entities, yet it may not be possible to map every target entity to a source entity.

3. There is no overlap in the labels in the source and target domains, and entities they correspond to have little in common.

While these three scenarios do not exhaust the possibilities that may be encountered in practice, they are helpful to categorize previous transfer learning experiments in the literature.

Under scenario 1, it is possible to map source domain labels to target domain labels without losing much information, to insure that both domains have the same label set. For example, if transferring from ACE 2005 (Walker et al., 2006) to CoNLL 2003, one could map the GEOPOLITICAL and FACILITY entities to LOCATION. This is often done manually (Daumé III, 2007; Augenstein et al., 2017), though Kim et al. (2015) introduced a method to automate this mapping: labels from both domains are embedded in a common vector space using Canonical Correlation Analysis (CCA), and k-nearest neighbors are used to find the target label closest to each source label. In their experiments the appropriate label mappings were not obvious and the CCA label embedding approach outperformed manual mappings. CCA is a dimensionality reduction technique (Hotelling, 1936) that will be described in Section 2.2.

One expects that transfer learning would be hardest in scenario 3, given that the source and target entities have little in common. For example, Qu et al. (2016) show that transfer from CoNLL 2003 (news) to CADEC (a biomedical corpus) is very difficult.

In this paper we focus on scenario 2, which is more challenging than scenario 1 and also of great practical importance. Many corpora have been published with coarse-grained entities, and it would be advantageous to harness this data to recognize other kinds of entities in other domains.

Figure 1 shows the combinations of datasets that have been used in previous transfer learning experiments relevant to the problem of disparate label sets, together with the additional experiments in this paper. We were unable to access TAC-KBP 2015 (Ji et al., 2015), i2b2 2016 (Stubbs et al., 2017) or MIMIC (Dernoncourt et al., 2017b). The TAC-KBP 2015 corpus consists mainly of news text, so we used MUC 6 and NIST IE-ER (NIST, 1999) as substitutes to see if we could obtain results consistent with those of (Obeidat et al., 2016). MIMIC is a manually corrected subset of the MIMIC-III dataset (Johnson et al., 2016). It was unclear how to replicate the manual corrections, so we leave experiments with

[2] https://github.com/ciads-ut/transfer-learning-ner

MIMIC to future work. We decided not to use ACE 2005; the annotated entities in this corpus are nested, and it was unclear which entities should be selected to fit the sequential classification framework.[3]

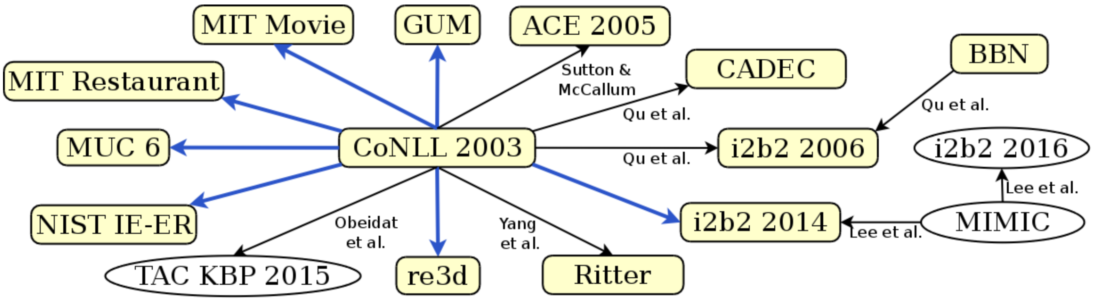

Figure 1: Source and target domains used in transfer learning experiments

Of the datasets shown in Figure 1, only CoNLL 2003, MIMIC and BBN (Weischedel and Brunstein, 2005) have been used as the source domain. Most past experiments use CoNLL 2003 as the source corpus. We replicate most of these past experiments, and extend experiments using CoNLL 2003 as a source corpus to seven additional target corpora that have not previously been used in transfer learning with novel classes for entity recognition. In future work, we plan to extend our study to additional source/target corpus pairs, particularly those that do not use CoNLL 2003 as the source corpus. Details on the datasets that we used for experiments are provided in Section 3.1.

2.2 Transfer learning frameworks

Here we briefly describe the transfer learning frameworks which are able to handle the problem of novel entity types. We use the term *framework*, rather than *algorithm*, because each of these can be implemented with a variety of machine learning models. It is important to emphasize that most existing transfer learning methods assume the label sets are the same across domains (Day and Khoshgoftaar, 2017; Weiss et al., 2016; Zhang et al., 2017). In addition, most of the methods which can handle disparate labels are designed for image and text classification rather than sequence classification problems (Day and Khoshgoftaar, 2017); some examples include (Shi et al., 2009; Shi et al., 2010; Quadrianto et al., 2010; Tommasi et al., 2010; Jie et al., 2011; Qi et al., 2011; Xiang et al., 2011; Patricia and Caputo, 2014; Feuz and Cook, 2015; Bhatt et al., 2016; Moon and Carbonell, 2016). To the best of our knowledge, the only two transfer learning approaches that have been proposed for entity recognition with different class labels are:

- **PRED:** train one classifier on the source domain, and use its predictions on the target domain as features for a second classifier.

- **Pre-training:** train a neural network on the source domain, use the trained weights to initialize a new neural network, and fine-tune the weights on the target domain.

In this paper we shall compare two variants of PRED against the Pre-training approach.

We note that transfer learning is closely related to work in *multi-task learning*. Multi-task learning consists in simultaneously training a classifier (e.g., a neural network with shared weights) to solve more than one task. In multi-task learning, however, the goal is to improve performance across all domains, rather than to optimize performance on the target domain. (Yang et al., 2017) is the only work we are aware of that uses multi-task learning on NER tasks with different label sets. For the sake of restricting the scope of this paper, in our experiments we restrict ourselves to approaches that train on the source and target data separately. A comparison to multi-task learning techniques that train simultaneously on source and target data is planned in future work.

[3] From a practical perspective, using the largest entity span is not always desirable.

PRED and PRED-CCA

It is common in NLP to use the outputs of one classifier as inputs to another one, but usually this is done for different tasks (e.g., using part-of-speech tags as features for NER), rather than for the same task with different class labels. Sutton and McCallum (2005) refer to this as "cascaded transfer"; we shall denote it PRED, following (Daumé III, 2007). A few researchers have used this approach for the entity recognition task with different class labels. Florian et al. (2004) use the output of entity taggers trained on corpora with different label types as features for a second classifier to tag entities in the ACE 2003 evaluation. They report an increase of 2% in F1 score over a baseline using lexical features with a Maximum Entropy classifier. Sutton and McCallum (2005) use PRED to transfer from CoNLL 2003 to ACE 2005, and show an improvement in F1 scores for several classes (e.g., an increase of 5.9% for person names and an increase of 1.2% for person nominals). Daumé III (2007) uses PRED to transfer from ACE 2005 to CoNLL 2003 (though source entities are mapped to target entities before the transfer[4]), and obtain a 0.82% increase in accuracy.

PRED-CCA[5] was introduced by Obeidat et al. (2016) as a variant of PRED. In PRED-CCA label embeddings rather than the predicted entities themselves are used as features for the second classifier. Label embeddings are obtained through CCA as in (Kim et al., 2015). CCA is a dimensionality reduction technique for paired data. In this case, every word is paired with its label. CCA is used to obtain label embeddings by finding a sequence of projections of the words and labels such that the correlations between the projections are maximized.

Obeidat et al. (2016) shows an improvement of 9% in micro-averaged F1 score of PRED-CCA over CCA when transferring from CoNLL 2003 to TAC-KBP 2015.

Pre-training

Fine tuning pre-trained neural networks has been used successfully in computer vision tasks, but seems comparatively less common in NLP. Mou et al. (2016) use pre-training to improve the performance of sentence and sentence-pair classification; two of their experiments transfer between corpora with different label sets. Kim et al. (2015) also experiment with pre-training for slot tagging, but only after the source domain labels have been mapped to target domain labels.

The only works we are aware of that apply pre-training to the entity recognition task are (Lee et al., 2017) and (Qu et al., 2016). Lee et al. (2017) use pre-trained neural networks to improve the performance of de-identification of protected health information in medical records, though they use two datasets with the same label set.[6] Qu et al. (2016) use pre-training in experiments with novel entities, using two conditional random fields (CRFs) rather than two neural networks. Unfortunately, we were not able to obtain the code used in Qu et al.'s experiments or to replicate their approach using new code.

Lee et al. (2017) use a bidirectional LSTM CRF (BiLSTM-CRF) neural network for their transfer learning experiments. Their network architecture is described in (Dernoncourt et al., 2017b), and is similar to the BiLSTM-CRF in (Lample et al., 2016). They experiment with transferring successive layers of the network: character embeddings, word embeddings, the character BiLSTM layer, the word BiLSTM layer, the fully connected layer, and the CRF layer. They experiment with transferring from MIMIC to i2b2 2014 and from MIMIC to i2b2 2016, and show that transferring weights of a pre-trained network boosts performance, with better results as more layers are transferred. While transferring the last (most task-specific) layers does not help as much as transferring the lower layers, it also does not hurt performance.

The results in (Lee et al., 2017) were produced with the NeuroNER program[7] (Dernoncourt et al., 2017a). Unfortunately, NeuroNER requires both source and target corpora to have the same label set, so we were not able to use it for our experiments. To run experiments similar to those in (Lee et al., 2017),

[4] (Daumé III, 2007) does not explicitly describe this preprocessing step, but we infer it based on the other methods that PRED is compared against.

[5] Obeidat et al. (2016) refer to it as "AugmntTr"; we call it PRED-CCA in order to emphasize its relation to PRED.

[6] The label sets were slightly different, but in the few cases where entities were different they were mapped to a common type.

[7] Available at `https://github.com/Franck-Dernoncourt/NeuroNER/`

but with different label sets, we implemented a variant of their neural network architecture. Since our goal is to replicate the transfer learning approach, not the exact network architecture, we made one simplification: we replaced the character-level embeddings with casing embeddings following (Reimers and Gurevych, 2017). This has two advantages: it makes the neural network experiments more comparable to the CRF experiments with PRED and PRED-CCA (which use simple token features), and it reduces the number of layers that can be transferred. In our experiments the only layers that can be transferred are the word embedding layer and the word BiLSTM layer, since the last two layers depend on the target domain's label set. Details of our implementation and our choice of hyperparameters are listed in Section 3.2.

3 Experimental Setup

Our experiments compare the performance of the transfer learning methods in different settings: we vary both the target corpus and the size of the target training set. We run every experiment 5 times, using different subsets of the target data to train with, and average results.

3.1 Datasets

We used CoNLL 2003 with the standard train/test split as the source corpus for our experiments . CoNLL 2003 is annotated with Person, Location, Organization and Miscellaneous entities. Table 1 shows the corpora that we used as the target, together with their entity types. The entities which overlap with the CoNLL 2003 entities are italicized; we refer to the others as "novel entities" [8].

Corpus	Domain	Annotated entities
Ritter	Twitter	*Person*, *geo-loc*, facility, company, sportsteam, band, product, tv-show, movie, other
i2b2 2006	Medical	Date, doctor, hospital, ID, *location*, patient, phone
i2b2 2014	Medical	Age, username, ID, patient, doctor, profession, hospital, street, state, zip, city, country, *organization*, date, medical record, phone
CADEC	Medical	Adverse reaction, disease, drug, finding, symptom
MUC 6	News	*Person*, *location*, *organization*, date, percent, money
NIST IE-ER	News	*Person*, *location*, *organization*, date, duration, percent, money, cardinal
GUM	Wikinews wikihow wikivoyage	Abstract, animal, event, object, *organization*, *person*, *place*, plant, quantity, substance, time
MIT Movie	Spoken queries	Actor, character, director, genre, plot, year, soundtrack, opinion, award, origin, quote, relationship
MIT Restaurant	Spoken queries	Amenity, cuisine, dish, hours, *location*, price, rating, restaurant name
re3d	Defense and security	Document reference, *location*, military platform, money, nationality, *organization*, *person*, quantity, temporal, weapon

Table 1: Corpora used as the target domain in our experiments

Entities which occured less than 20 times in a corpus were removed[9]. We used the standard train/test splits for the i2b2 2006 (Uzuner et al., 2007), i2b2 2014 (Stubbs and Uzuner, 2015), MIT Movie (Liu et al., 2013b) and MIT Restaurant (Liu et al., 2013a) corpora. For the Ritter corpus (Ritter et al., 2011)

[8] We consider subtypes of CoNLL 2003 entities such as "doctor" and "hospital" to be novel. Although "geo-loc" and "place" are used in slightly different ways, we identify them with "location", so do not consider them to be novel.

[9] In particular, "time" was removed from both MUC 6 and NIST IE-ER; "age" was removed from i2b2 2006; "healthplan", "URL", "fax", "email", "device", "location-other" and "bioID" were removed from i2b2 2014; "commsIdentifier", "frequency" and "vehicle" were removed from re3d.

we used the same train/test split as in (Yang et al., 2017)[10]. We used stratified random sampling at the sentence level to create suitable test sets for CADEC (Karimi et al., 2015), GUM (Zeldes, 2017), NIST IE-ER (NIST, 1999), re3d (DSTL, 2017) and MUC 6 (Grishman and Sundheim, 1996)[11]. Unlike i2b2 2006, the i2b2 2014 corpus is not tokenized. To stay consistent with (Lee et al., 2017), we used the scripts included with NeuroNER with the *spacy* tokenizer to convert i2b2 2014 to the CoNLL 2003 format.

CoNLL 2003, Ritter, and i2b2 2014 also have default train/test/dev splits. In order to have more control over the sizes of the training and development sets (in particular, to have the same train/dev ratio over increasing sizes of the training set), we decided not to use the standard train/dev splits.

The available target training data is randomly shuffled, and for every shuffle we use increasing subsets of the target corpus (with 20, 50, 100, 250, 500, 1000, 1500 and 2000 sentences) as target training data. In order not to give the neural network methods an unfair advantage over PRED and PRED-CCA, we do not use a hold-out development set for them, and instead use 20% of the available training data for validation.

3.2 Implementation Details

We used the IOB2 tagging scheme for all our experiments. For every experiment we compute the micro and macro-averaged chunk-based precision, recall and F1 scores, as well as micro and macro-averaged scores over novel classes, as was done in (Qu et al., 2016). Due to space limitations we only include the micro-averaged scores in this paper, but our full results (including scores for specific labels) will be available online.

PRED and PRED-CCA

Obeidat et al.'s experiments comparing PRED and PRED-CCA were performed using *Stanford NER*, which is an implementation of a second-order linear CRF. Since we are replicating the PRED and PRED-CCA approaches rather than Obeidat's exact setup, we perform our experiments with *CRFSuite*. Although Stanford NER has an extensive set of built-in features, it is harder to modify them than it is for CRFSuite. We use the following features when evaluating PRED and PRED-CCA: (1) the current token x_i and tokens in a window of size 2, (2) initial capitalization of tokens in a window of size 2, (3) word shape information for x_i (uppercase, alphanumeric, or all digits), and (4) token prefixes of lengths 3 and 4, and token suffixes of lengths 1 through 4.

These features were used in (Zhang and Johnson, 2003)[12], and were also used as baseline features in (Turian et al., 2010). Augenstein et al. (2017) also evaluate CRFSuite on NER across a variety of datasets with features similar to these. Due to limitations of CRFSuite, we did not include second-order label transitions or features of the form $\mathbb{I}_{tag_i \wedge x_i}$. We also opted against using gazeteer, word embedding, part-of-speech or chunking features, since the goal of this paper is to compare transfer learning approaches and not achieve state-of-the-art results.

While Kim et al. (2015) and Obeidat et al. (2016) use the predicted entities without the IOB prefix as features, we experimented with using the IOB-prefixed entities for both PRED and PRED-CCA. Preliminary experiments suggested the differences were minimal, so we ran all our experiments with entity labels without the IOB prefix as features. In addition, Obeidat et al. (2016) do not specify whether they included the "O" entity label when computing the CCA label embeddings. If "O" is not included in the calculation, one must find a suitable label embedding for "O" separately. We experimented with excluding "O" from the computation and mapping "O" to the zero vector. The difference was minimal, so we ran our full experiments with the "O" label included in the CCA computation.

We use the same embedding dimension as (Obeidat et al., 2016), k=5, for all our experiments.

[10] Available at `http://kimi.ml.cmu.edu/transfer/data.tar.gz`

[11] We used 1000 sentences for the test sets of CADEC, GUM, and MUC6; for NIST IE-ER and re3d we used 690 and 200 sentences, respectively.

[12] Specifically the combination B+D+E+F, listed in their Table 1.

Pre-training

We used the same network architecture and hyperparameters as in (Lee et al., 2017) [13], with the following exceptions: (1) we replaced the character embedding layer with a casing layer as in (Reimers and Gurevych, 2017), (2) we used the IOB2 tagging scheme rather than the BIOES tagging scheme, and (3) we used RMSprop rather than stochastic gradient descent (SGD) for fine-tuning on the target dataset. Preliminary experiments showed that SGD with a learning rate of 0.005 and 100 epochs (the defaults in (Lee et al., 2017)) failed to converge for many of our datasets. We experimented with learning rates of 0.05 and 0.5, as well as with optimizers with adaptive learning rates (Adam, Adagrad, and RMSprop). Since RMSprop tended to have the highest F1 scores, and in order not to have to tune the learning rate of SGD for each dataset, we ran our final fine-tuning experiments with RMSprop, with a learning rate of 0.001.

We ran experiments with three settings: no transfer (training the network from scratch on the target corpus), transferring the word embedding layer, and transferring both the word embedding layer and the BiLSTM layer. We observed that the difference between not transferring any layers and transferring one layer was often small; this was also noted in (Lee et al., 2017). We therefore only include the results with transferring both layers and when discussing pretraining below we are referring to this case.

4 Results

We are particularly interested in comparing the behavior of the transfer learning methods as the size of the target dataset varies. This can help practitioners decide which approach is best suited to their situation. Plots of micro-averaged precision, recall and F1 scores with error bars for a few target domain datasets are shown in Figures 2, 3 and 4 [14]. We denote the no-transfer (training on the target only) baselines for the CRF and the BiLSTM-CRF as CRF-TGT and BiLSTM-TGT, respectively.

4.1 Comparing PRED/PRED-CCA and neural network approaches

The PRED/PRED-CCA approaches have greater precision than the neural network approaches in most cases (the only exceptions being GUM, with 1500 and 2000 sentences), with the greatest difference occurring when the number of target training sentences is small. For recall, the situation is not so clear. For MIT Restaurant, MIT Movie, CADEC, i2b2 2014 and i2b2 2006 (Figure 2), the neural network approaches have better recall than PRED and PRED-CCA across all dataset sizes; this results in the neural network approaches having a greater F1 score for these datasets in most cases (the exceptions being MIT Restaurant with under 500 sentences, and CADEC with 500 sentences).

On the other hand, for GUM, re3d, NIST IE-ER, MUC 6 (Figure 4) and Ritter (Figure 3), we observe that initially PRED and PRED-CCA have higher recall than the neural network approaches, but the neural networks' recall surpasses them once the target domain dataset is large enough. This occurs at 500 sentences for Ritter, NIST IE-ER and MUC 6, and at 250 sentences for re3d and GUM; it leads to a similar pattern in F1 scores: for Ritter, GUM, MUC 6 and NIST IE-ER, PRED and PRED-CCA have higher F1 scores than either of the neural network approaches for smaller target training sets, but the situation is reversed once the target domain dataset is large enough.

4.2 Comparing PRED and PRED-CCA

For most corpora, the PRED/PRED-CCA approaches improved the F1 score over the "target-only" (CRF-TGT) baseline. MIT Restaurant and MIT Movie were the only corpora where the F1 scores for CRF-TGT, PRED, and PRED-CCA were about the same. For CADEC the F1 scores for the 20 and 50 sentence settings decreased.

PRED-CCA outperformed PRED in most cases, but not always by very much. MUC 6, NIST IE-ER and Ritter show the greatest increase in F1 score, due to an increase in recall. The improvements in F1

[13] The hyperparameters can be found in the `parameters.ini` file in NeuroNER; they include using dropout of 0.5 after the word embedding layer, using the 100-dimensional Glove 6B word embeddings, and using 100 LSTM hidden units. Training on the source corpus is done using SGD with 100 epochs and patience of 10, with a learning rate of 0.005 and a gradient clipping value of 5.0.

[14] We only include three due to space limitations; the plots for the other transfer learning experiments will be available online.

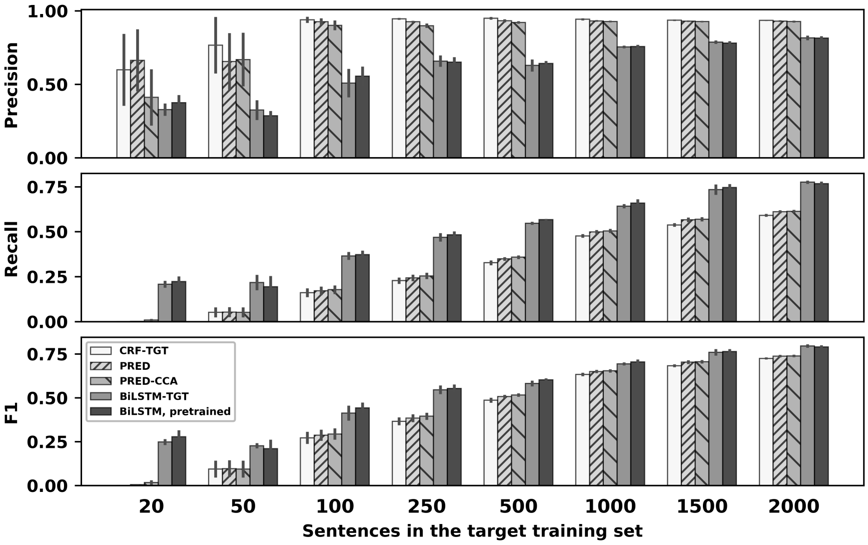

Figure 2: Micro-averaged scores for the i2b2 2006 corpus

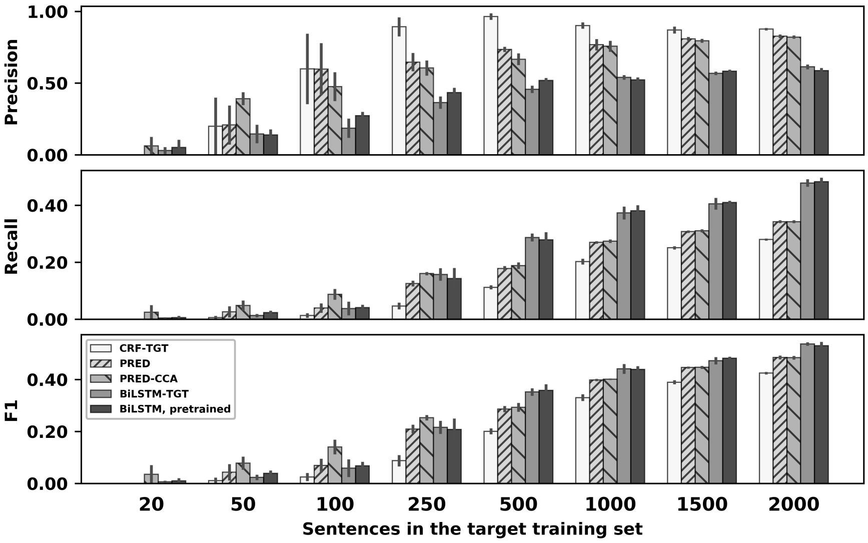

Figure 3: Micro-averaged scores for the Ritter corpus

score are highest for smaller target training datasets, and rapidly decrease with increasing dataset size. Figure 5 shows the change in F1 score between PRED and PRED-CCA for the MUC 6 and Ritter corpora. NIST IE-ER displayed a pattern similar to MUC 6, so it was omitted. re3d showed an improvement of 2% under the 20 sentence setting; the improvement for the other datasets was under 1%.

We note that in most cases, using PRED-CCA rather than PRED results in a small drop in precision; this is consistent with the results in (Obeidat et al., 2016), who observe a drop of 1% between PRED and PRED-CCA for the TAC-KBP 2015 corpus. The large increase in F1 score for MUC 6 and NIST IE-ER are consistent with the 9% increase for TAC-KBP 2015 reported in (Obeidat et al., 2016). Although the label sets are different in each case, MUC 6, NIST IE-ER and TAC-KBP 2015 are all news corpora, as is CoNLL 2003. We hypothesize that the reason PRED-CCA does so much better than PRED in these cases is that there is more of an overlap in the source and target domains' vocabularies. This would lead to more semantically similar source and target labels being mapped closer together.[15]

[15] We did not lowercase the tokens before computing the label embeddings. We expect to obtain higher F1 scores in this case, particularly for the MIT Movie and MIT Restaurant corpora, but leave this for future work.

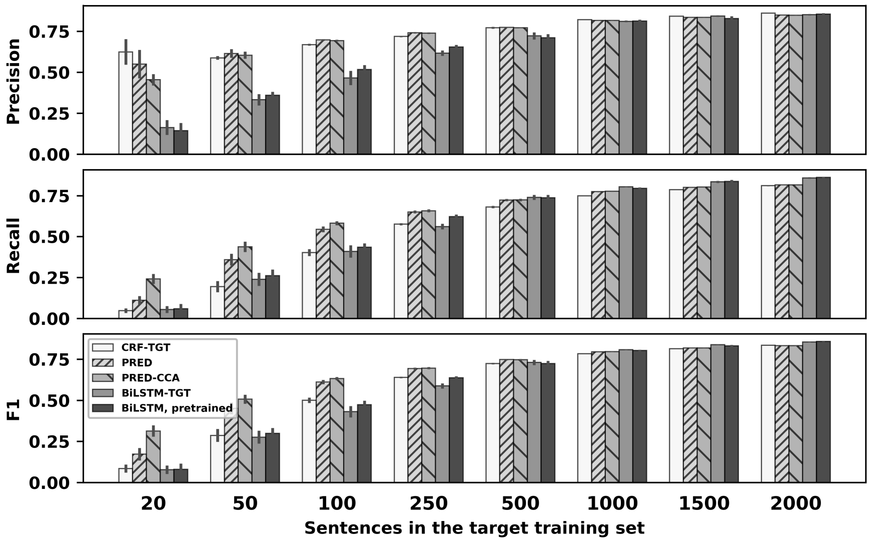

Figure 4: Micro-averaged scores for the MUC 6 corpus

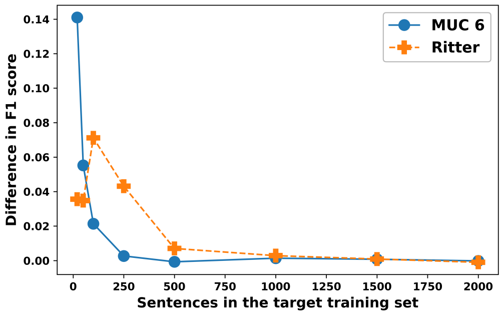

Figure 5: Difference in F1 score between PRED and PRED-CCA

4.3 Comparing the BiLSTM-CRFs with and without pre-training

The results for the pre-trained neural network were less clear. On i2b2 2006, i2b2 2014, and MIT Movie, using pre-training generally led to an improvement in F1 scores over the BiLSTM-TGT baseline. When these datasets had over 500 sentences the improvement was minimal (under 1%); for less than 500 sentences the improvement varied, but was often between 2% and 3%. In these scenarios pre-training also performed better than PRED and PRED-CCA as well. The improvement in F1 score of i2b2 2014 with 1500 sentences was 0.7%. For comparison, Lee et al. (2017) report an increase of 3% when transferring from MIMIC to i2b2 2014, when using 5% of their target set (roughly 1600 sentences) and reusing all layers up to the token layer.

For MUC 6, NIST IE-ER, re3d, Ritter, MIT Restaurant, and GUM, pre-training often achieved a higher F1 score than training from scratch; this mostly occurred when there were 500 target training sentences or less. The increase in F1 score was highest for MUC 6 (2%-5%) and for re3d (2-3%). However, in nearly all of these cases pre-training was outperformed by PRED-CCA.

5 Summary and Conclusion

In this reproduction paper, we have compared three existing methods that can be applied to the setting of transfer learning with novel entities in the target domain. These methods have not been compared against each other before in the literature.

In summary, our results show that in some situations (particularly when there is less labeled target data), PRED and PRED-CCA outperform pre-training neural networks for the entity recognition task.[16]

[16] This may not generalize to other neural network architectures, hyperparameter choices, or domain pairs.

With sufficient labeled data, the neural transfer approaches tend to do well, but of course, these correspond to cases with less data scarcity.

Thus, we encourage researchers to compare neural network approaches with simpler baselines such as PRED and PRED-CCA, rather than only comparing against the "no tranfer" option, and to report not only F1 scores, but also precision and recall, which may be useful when deciding which transfer learning approach to use.

A number of extensions to our study are possible in future work. For example, we hope to run similar comparisons with multi-task and few-shot learning approaches. In particular, we believe methods for few-shot learning in image classification could potentially be adapted to sequential classification problems such as entity recognition. A more thorough study of hyperparameter tuning (versus replicating previously published hyperparameters) would also be interesting future work, particularly for the neural network approaches, which are known to be sensitive to choices in hyperparameters.

References

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Himanshu Sharad Bhatt, Manjira Sinha, and Shourya Roy. 2016. Cross-domain text classification with multiple domains and disparate label sets. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1641–1650.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Oscar Day and Taghi M. Khoshgoftaar. 2017. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4(1):29.

Defence Science and Technology Laboratory. 2017. Relationship and Entity Extraction Evaluation Dataset. `https://github.com/dstl/re3d`. Accessed: January 2018.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017a. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 97–102, Copenhagen, Denmark, September. Association for Computational Linguistics.

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017b. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Kyle D. Feuz and Diane J. Cook. 2015. Transfer learning across feature-rich heterogeneous feature spaces via feature-space remapping (FSR). *ACM Transactions on Intelligent Systems and Technology*, 6(1):3:1–3:27.

Radu Florian, Hany Hassan, Abraham Ittycheriah, Hongyan Jing, Nanda Kambhatla, Xiaoqiang Luo, Nicolas Nicolov, and Salim Roukos. 2004. A statistical model for multilingual entity detection and tracking. In Susan Dumais, Daniel Marcu, and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 1–8, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP 2015 tri-lingual entity discovery and linking. In *Proceedings of the Eighth Text Analysis Conference (TAC 2015)*.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic, June. Association for Computational Linguistics.

Luo Jie, Tatiana Tommasi, and Barbara Caputo. 2011. Multiclass transfer learning from unconstrained priors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1863–1870. IEEE.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81. Available at `https://data.csiro.au` Accessed: November 2017.

Young-Bum Kim, Karl Stratos, Ruhi Sarikaya, and Minwoo Jeong. 2015. New transfer learning techniques for disparate label sets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 473–482, Beijing, China, July. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2017. Transfer Learning for Named-Entity Recognition with Neural Networks. *arXiv preprint arXiv:1705.06273*.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013a. Asgard: A portable architecture for multilingual dialogue systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8386–8390. IEEE. Available at `https://groups.csail.mit.edu/sls/downloads/restaurant/` Accessed: January 2018.

Jingjing Liu, Panupong Pasupat, Yining Wang, Scott Cyphers, and Jim Glass. 2013b. Query understanding enhanced by hierarchical parsing structures. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 72–77. IEEE. Available at `https://groups.csail.mit.edu/sls/downloads/movie/` We used the trivia10k13 portion. Accessed: January 2018.

Seungwhan Moon and Jaime Carbonell. 2016. Proactive transfer learning for heterogeneous feature and label spaces. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 706–721. Springer.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in NLP applications? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 479–489, Austin, Texas, November. Association for Computational Linguistics.

NIST. 1999. Information Extraction - Entity Recognition Evaluation. `http://www.nist.gov/speech/tests/ieer/er_99/er_99.htm`. We used the newswire development test data included in the NLTK package.

Rasha Obeidat, Xiaoli Z. Fern, and Prasad Tadepalli. 2016. Label embedding approach for transfer learning. In *Proceedings of the Joint International Conference on Biological Ontology and BioCreative, Corvallis, Oregon, United States, August 1-4, 2016.*

Novi Patricia and Barbara Caputo. 2014. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proceedings of the Computer Vision and Pattern Recognition*.

Guo-Jun Qi, Charu Aggarwal, Yong Rui, Qi Tian, Shiyu Chang, and Thomas Huang. 2011. Towards cross-category knowledge propagation for learning visual concepts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 897–904. IEEE.

Lizhen Qu, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, and Timothy Baldwin. 2016. Named entity recognition for novel types by transfer learning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 899–905, Austin, Texas, November. Association for Computational Linguistics.

Novi Quadrianto, James Petterson, Tibério S. Caetano, Alex J. Smola, and S.V.N. Vishwanathan. 2010. Multitask learning without label correspondences. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1957–1965. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2017. Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks. *arXiv preprint arXiv:1707.06799*.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.

Xiaoxiao Shi, Wei Fan, Qiang Yang, and Jiangtao Ren. 2009. Relaxed transfer of different classes via spectral partition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 366–381. Springer.

Xiaoxiao Shi, Qi Liu, Wei Fan, Philip S. Yu, and Ruixin Zhu. 2010. Transfer learning on heterogenous feature spaces via spectral transformation. In *2010 IEEE International Conference on Data Mining*, pages 1049–1054.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of biomedical informatics*, 58:S20–S29. Available at `https://www.i2b2.org/NLP/DataSets/` Accessed: February 2018.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of Biomedical Informatics*, 75:S4 – S18. A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.

Charles Sutton and Andrew McCallum. 2005. Composition of conditional random fields for transfer learning. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 748–754, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. 2010. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3081–3088. IEEE.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July. Association for Computational Linguistics.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563. Available at `https://www.i2b2.org/NLP/DataSets/` Accessed: February 2018.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. *Linguistic Data Consortium, Philadelphia*.

Ralph Weischedel and Ada Brunstein. 2005. BBN pronoun coreference and entity type corpus. *Linguistic Data Consortium, Philadelphia*.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data*, 3(1):9.

Evan Wei Xiang, Sinno Jialin Pan, Weike Pan, Jian Su, and Qiang Yang. 2011. Source-selection-free transfer learning. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2355–2360. AAAI Press.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. *arXiv preprint arXiv:1703.06345*.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612. Available at `https://github.com/amir-zeldes/gum/tree/master/coref/tsv/` Accessed: November 2017.

Tong Zhang and David Johnson. 2003. A robust risk minimization based named entity recognition system. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 204–207.

Jing Zhang, Wanqing Li, and Philip Ogunbona. 2017. Transfer Learning for Cross-Dataset Recognition: A Survey. *arXiv preprint arXiv:1705.04396*.