Multi-task Learning of Pairwise Sequence Classification Tasks Over Disparate Label Spaces

Augenstein, Isabelle; Ruder, Sebastian; Søgaard, Anders

Published in: Proceedings, 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

DOI: 10.18653/v1/N18-1172

Publication date: 2018

Document version
Early version, also known as pre-print

Citation for published version (APA):
Multi-task Learning of Pairwise Sequence Classification Tasks Over Disparate Label Spaces

Isabelle Augenstein1, Sebastian Ruder23, Anders Søgaard1
1Department of Computer Science, University of Copenhagen, Denmark
2Insight Research Centre, National University of Ireland, Galway
3Aylien Ltd., Dublin, Ireland
{augenstein|soegaard}@di.ku.dk, sebastian@ruder.io

Abstract

We combine multi-task learning and semi-supervised learning by inducing a joint embedding space between disparate label spaces and learning transfer functions between label embeddings, enabling us to jointly leverage unlabelled data and auxiliary, annotated datasets. We evaluate our approach on a variety of sequence classification tasks with disparate label spaces. We outperform strong single and multi-task baselines and achieve a new state-of-the-art for topic-based sentiment analysis.

1 Introduction

Multi-task learning (MTL) and semi-supervised learning are both successful paradigms for learning in scenarios with limited labelled data and have in recent years been applied to almost all areas of NLP. Applications of MTL in NLP, for example, include partial parsing (Søgaard and Goldberg, 2016), text normalisation (Bollman et al., 2017), neural machine translation (Luong et al., 2016), and keyphrase boundary classification (Augenstein and Søgaard, 2017).

Contemporary work in MTL for NLP typically focuses on learning representations that are useful across tasks, often through hard parameter sharing of hidden layers of neural networks (Collobert et al., 2011; Søgaard and Goldberg, 2016). If tasks share optimal hypothesis classes at the level of these representations, MTL leads to improvements (Baxter, 2000). However, while sharing hidden layers of neural networks is an effective regulariser (Søgaard and Goldberg, 2016), we potentially lose synergies between the classification functions trained to associate these representations with class labels. This paper sets out to build an architecture in which such synergies are exploited.

Contributions Our contributions are: a) We model the relationships between labels by inducing a joint label space for multi-task learning. b) We propose a Label Transfer Network that learns to transfer labels between tasks and propose to use semi-supervised learning to leverage them for training. c) We evaluate MTL approaches on a variety of classification tasks and shed new light on settings where multi-task learning works. d) We perform an extensive ablation study of our model.

The first two authors contributed equally.
e) We report state-of-the-art performance on topic-based sentiment analysis.

2 Related work

Learning task similarities Existing approaches for learning similarities between tasks enforce a clustering of tasks (Evgeniou et al., 2005; Jacob et al., 2009), induce a shared prior (Yu et al., 2005; Xue et al., 2007; Daume III, 2009), or learn a grouping (Kang et al., 2011; Kumar and Daume III, 2012). These approaches focus on homogeneous tasks and employ linear or Bayesian models. They can thus not be directly applied to our setting with tasks using disparate label sets.

Multi-task learning with neural networks Recent work in multi-task learning goes beyond hard parameter sharing (Caruana, 1993) and considers different sharing structures, e.g. only sharing at lower layers (Søgaard and Goldberg, 2016) and induces private and shared subspaces (Liu et al., 2017; Ruder et al., 2017). These approaches, however, are not able to take into account relationships between labels that may aid in learning. Another related direction is to train on disparate annotations of the same task (Chen et al., 2016; Peng et al., 2017). In contrast, the different nature of our tasks requires a modelling of their label spaces.

Semi-supervised learning There exists a wide range of semi-supervised learning algorithms, e.g., self-training, co-training, tri-training, EM, and combinations thereof, several of which have also been used in NLP. Our approach is probably most closely related to an algorithm called co-forest (Li and Zhou, 2007). In co-forest, like here, each learner is improved with unlabeled instances labeled by the ensemble consisting of all the other learners. Note also that several researchers have proposed using auxiliary tasks that are unsupervised (Plank et al., 2016; Rei, 2017), which also leads to a form of semi-supervised models.

Label transformations The idea of manually mapping between label sets or learning such a mapping to facilitate transfer is not new. Zhang et al. (2012) use distributional information to map from a language-specific tagset to a tagset used for other languages, in order to facilitate cross-lingual transfer. More related to this work, Kim et al. (2015) use canonical correlation analysis to transfer between tasks with disparate label spaces. There has also been work on label transformations in the context of multi-label classification problems (Yeh et al., 2017).

3 Multi-task learning with disparate label spaces

3.1 Problem definition

In our multi-task learning scenario, we have access to labelled datasets for \( T \) tasks \( T_1, \ldots, T_T \) at training time with a target task \( T_T \) that we particularly care about. The training dataset for task \( T_i \) consists of \( N_k \) examples \( X_{T_i} = \{x_{i1}^{T_i}, \ldots, x_{N_k}^{T_i}\} \) and their labels \( Y_{T_i} = \{y_{i1}^{T_i}, \ldots, y_{N_k}^{T_i}\} \). Our base model is a deep neural network that performs classic hard parameter sharing (Caruana, 1993): It shares its parameters across tasks and has task-specific softmax output layers, which output a probability distribution \( p^{T_i} \) for task \( T_i \) according to the following equation:

\[
p^{T_i} = \text{softmax}(W^{T_i}h + b^{T_i})
\]

where \( \text{softmax}(x) = e^x / \sum_{i=1}^{||x||} e^x \), \( W^{T_i} \in \mathbb{R}^{L_i \times h}, b^{T_i} \in \mathbb{R}^{L_i} \) is the weight matrix and bias term of the output layer of task \( T_i \) respectively, \( h \in \mathbb{R}^h \) is the jointly learned hidden representation, \( L_i \) is the number of labels for task \( T_i \), and \( h \) is the dimensionality of \( h \).

The MTL model is then trained to minimise the sum of the individual task losses:

\[
\mathcal{L} = \lambda_1 \mathcal{L}_1 + \ldots + \lambda_T \mathcal{L}_T
\]

where \( \mathcal{L}_i \) is the negative log-likelihood objective \( \mathcal{L}_i = \mathcal{H}(p^{T_i}, y^{T_i}) = -\frac{1}{N} \sum_n \sum_j \log p_{j|y^{T_i}_n} \) and \( \lambda_i \) is a parameter that determines the weight of task \( T_i \). In practice, we apply the same weight to all tasks. We show the full set-up in Figure 14.

3.2 Label Embedding Layer

In order to learn the relationships between labels, we propose a Label Embedding Layer (LEL) that embeds the labels of all tasks in a joint space. Instead of training separate softmax output layers as above, we introduce a label compatibility function \( c(\cdot, \cdot) \) that measures how similar a label with embedding \( I \) is to the hidden representation \( h \):

\[
c(1, h) = 1 \cdot h
\]

where \( \cdot \) is the dot product. This is similar to the Universal Schema Latent Feature Model introduced by Riedel et al. (2013). In contrast to
other models that use the dot product in the objective function, we do not have to rely on negative sampling and a hinge loss (Collobert and Weston 2008) as negative instances (labels) are known. For efficiency purposes, we use matrix multiplication instead of a single dot product and softmax instead of sigmoid activations:

$$p = \text{softmax}(Lh)$$  \hspace{1cm} (4)

where $L \in \mathbb{R}^{(\sum_{i} L_i) \times l}$ is the label embedding matrix for all tasks and $l$ is the dimensionality of the label embeddings. In practice, we set $l$ to the hidden dimensionality $h$. We use padding if $l < h$. We apply a task-specific mask to $L$ in order to obtain a task-specific probability distribution $p_T$. The LEL is shared across all tasks, which allows us to learn the relationships between the labels in the joint embedding space. We show MTL with the LEL in Figure [1].

### 3.3 Label Transfer Network

The LEL allows us to learn the relationships between labels. In order to make use of these relationships, we would like to leverage the predictions of our auxiliary tasks to estimate a label for the target task. To this end, we introduce the Label Transfer Network (LTN). This network takes the auxiliary task outputs as input. In particular, we define the output label embedding $o_i$ of task $T_i$ as the sum of the task’s label embeddings $l_j$ weighted with their probability $p_{j,i}$:

$$o_i = \sum_{j=1}^{L_i} p_{j,i} l_j$$  \hspace{1cm} (5)

The label embeddings $L$ encode general relationship between labels, while the model’s probability distribution $p_T$ over its predictions encodes fine-grained information useful for learning (Hinton et al. 2015). The LTN is trained on labelled target task data. For each example, the corresponding label output embeddings of the auxiliary tasks are fed into a multi-layer perceptron (MLP), which is trained with a negative log-likelihood objective $L_{\text{LTN}}$ to produce a pseudo-label $z_T$ for the target task $T_T$:

$$L_{\text{LTN}} = \text{MLP}([o_1, \ldots, o_{T-1}])$$  \hspace{1cm} (6)

where $[\cdot, \cdot]$ designates concatenation. The mapping of the tasks in the LTN yields another signal that can be useful for optimisation and act as a regulariser. The LTN can also be seen as a mixture-of-experts layer (Jacobs et al. 1991) where the experts are the auxiliary task models. As the label embeddings are learned jointly with the main model, the LTN is more sensitive to the relationships between labels than a separately learned mixture-of-experts model that only relies on the...
3.4 Semi-supervised MTL

The downside of the LTN is that it requires additional parameters and relies on the predictions of the auxiliary models, which impacts the runtime during testing. Instead of using the LTN for prediction directly, we can use it to provide pseudo-labels for unlabelled or auxiliary task data by utilising auxiliary predictions for semi-supervised learning.

We train the target task model on the pseudo-labelled data to minimise the squared error between the model predictions $p^T_i$ and the pseudo labels $z^T_i$ produced by the LTN:

$$L_{\text{pseudo}} = MSE(p^T_i, z^T_i) = ||p^T_i - z^T_i||^2$$

(7)

We add this loss term to the MTL loss in Equation 2. As the LTN is learned together with the MTL model, pseudo-labels produced early during training will likely not be helpful as they are based on unreliable auxiliary predictions. For this reason, we first train the base MTL model until convergence and then augment it with the LTN. We show the full semi-supervised learning procedure in Figure 1c.

3.5 Data-specific features

When there is a domain shift between the datasets of different tasks as is common for instance when learning NER models with different label sets, the output label embeddings might not contain sufficient information to bridge the domain gap.

To mitigate this discrepancy, we augment the LTN’s input with features that have been found useful for transfer learning (Ruder and Plank 2017). In particular, we use the number of word types, type-token ratio, entropy, Simpson’s index, and Rényi entropy as diversity features. We calculate each feature for each example. The features are then concatenated with the input of the LTN.

3.6 Other multi-task improvements

Hard parameter sharing can be overly restrictive and provide a regularisation that is too heavy when jointly learning many tasks. For this reason, we propose several additional improvements that seek to alleviate this burden: We use skip-connections, which have been shown to be useful for multi-task learning in recent work (Ruder et al., 2017). Furthermore, we add a task-specific layer before the output layer, which is useful for learning task-specific transformations of the shared representations (Søgaard and Goldberg, 2016; Ruder et al., 2017).

4 Experiments

For our experiments, we evaluate on a wide range of text classification tasks. In particular, we choose pairwise classification tasks—i.e. those that condition the reading of one sequence on another sequence—as we are interested in understanding if knowledge can be transferred even for these more complex interactions. To the best of our knowledge, this is the first work on transfer learning between such pairwise sequence classification tasks. We implement all our models in Tensorflow (Abadi et al., 2016) and release the code at https://github.com/coastalcph/mtl-disparate.

4.1 Tasks and datasets

We use the following tasks and datasets for our experiments, show task statistics in Table 1 and summarise examples in Table 2:

<table>
<thead>
<tr>
<th>Task</th>
<th>Domain</th>
<th>N</th>
<th>L</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic-2</td>
<td>Twitter</td>
<td>4,346</td>
<td>2</td>
<td>$p^{PN}$</td>
</tr>
<tr>
<td>Topic-5</td>
<td>Twitter</td>
<td>6,000</td>
<td>5</td>
<td>$MAE^M$</td>
</tr>
<tr>
<td>Target</td>
<td>Twitter</td>
<td>6,248</td>
<td>3</td>
<td>$F^M$</td>
</tr>
<tr>
<td>stance</td>
<td>Twitter</td>
<td>2,914</td>
<td>3</td>
<td>$F^A$</td>
</tr>
<tr>
<td>ABSA-L</td>
<td>Reviews</td>
<td>2,909</td>
<td>3</td>
<td>Acc</td>
</tr>
<tr>
<td>ABSA-R</td>
<td>Reviews</td>
<td>2,507</td>
<td>3</td>
<td>Acc</td>
</tr>
<tr>
<td>FNC-1</td>
<td>News</td>
<td>39,741</td>
<td>4</td>
<td>Acc</td>
</tr>
<tr>
<td>MultiNLI</td>
<td>Diverse</td>
<td>392,702</td>
<td>3</td>
<td>Acc</td>
</tr>
</tbody>
</table>

Table 1: Training set statistics and evaluation metrics of every task. N: # of examples. L: # of labels.

For more information regarding the feature calculation, refer to Ruder and Plank (2017).

Topic-based sentiment analysis Topic-based sentiment analysis aims to estimate the sentiment of a tweet known to be about a given topic. We use the data from SemEval-2016 Task 4 Subtask B and C (Nakov et al., 2016) for predicting on a two-point scale of positive and negative (Topic-2) and five-point scale ranging from highly negative
<table>
<thead>
<tr>
<th>Topic-based sentiment analysis: Tweet: No power at home, sat in the dark listening to AC/DC in the hope it’ll make the electricity come back again Topic: AC/DC Label: positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target-dependent sentiment analysis: Text: how do you like settlers of catan for the wii? Target: wii Label: neutral</td>
</tr>
<tr>
<td>Aspect-based sentiment analysis: Text: For the price, you cannot eat this well in Manhattan Aspects: restaurant prices, food quality Label: positive</td>
</tr>
<tr>
<td>Stance detection: Tweet: Be prepared - if we continue the policies of the liberal left, we will be #Greece Target: Donald Trump Label: favor</td>
</tr>
<tr>
<td>Fake news detection: Document: Dino Ferrari hooked the whopper wels catfish, (...), which could be the biggest in the world. Headline: Fisherman lands 19 STONE catfish which could be the biggest in the world to be hooked Label: agree</td>
</tr>
<tr>
<td>Natural language inference: Premise: Fun for only children Hypothesis: Fun for adults and children Label: contradiction</td>
</tr>
</tbody>
</table>

Table 2: Example instances from the datasets described in Section 4.1

to highly positive (Topic-5) respectively. An example from this dataset would be to classify the tweet “No power at home, sat in the dark listening to AC/DC in the hope it’ll make the electricity come back again” known to be about the topic “AC/DC”, which is labelled as a positive sentiment. The evaluation metrics for Topic-2 and Topic-5 are macro-averaged recall ($\rho_{PN}$) and macro-averaged mean absolute error ($MAE^M$) respectively, which are both averaged across topics.

**Aspect-based sentiment analysis** Aspect-based sentiment analysis is the task of identifying whether an aspect, i.e. a particular property of an item is associated with a positive, negative, or neutral sentiment (Ruder et al., 2016). We use the data of SemEval-2016 Task 5 Subtask 1 Slot 3 (Pontiki et al., 2016) for the laptops (ABSA-L) and restaurants (ABSA-R) domains. An example is the sentence “For the price, you cannot eat this well in Manhattan”, labelled as positive towards both the aspects “restaurant prices” and “food quality”. The evaluation metric for both domains is accuracy ($Acc$).

**Stance detection** Stance detection (Stance) requires a model, given a text and a target entity, which might not appear in the text, to predict whether the author of the text is in favour or against the target or whether neither inference is likely (Augenstein et al., 2016). We use the data of SemEval-2016 Task 6 Subtask B (Mohammad et al., 2016). An example from this dataset would be to predict the stance of the tweet “Be prepared - if we continue the policies of the liberal left, we will be #Greece” towards the topic “Donald Trump”, labelled as “favor”. The evaluation metric is the macro-averaged $F_1$ score of the “favour” and “against” classes ($F_1^{FA}$).

**Fake news detection** The goal of fake news detection in the context of the Fake News Challenge is to estimate whether the body of a news article agrees, disagrees, discusses, or is unrelated towards a headline. We use the data from the first stage of the Fake News Challenge (FNC-1). An example for this dataset is the document “Dino Ferrari hooked the whopper wels catfish, (...), which could be the biggest in the world.” with the headline “Fisherman lands 19 STONE catfish which could be the biggest in the world to be hooked” labelled as “agree”. The evaluation metric is accuracy ($Acc$).

**Natural language inference** Natural language inference is the task of predicting whether one sentence entails, contradicts, or is neutral towards another one. We use the Multi-Genre NLI corpus (MultiNLI) from the RepEval 2017 shared task (Nangia et al., 2017). An example for an instance would be the sentence pair “Fun for only children”, “Fun for adults and children”, which are...
in a “contradiction” relationship. The evaluation metric is accuracy (Acc).

4.2 Base model

Our base model is the Bidirectional Encoding model (Augenstein et al., 2016), a state-of-the-art model for stance detection that conditions a bidirectional LSTM (BiLSTM) encoding of a text on the BiLSTM encoding of the target. Unlike Augenstein et al. (2016), we do not pre-train word embeddings on a larger set of unlabelled in-domain text for each task as we are mainly interested in exploring the benefit of multi-task learning for generalisation.

4.3 Training settings

We use BiLSTMs with one hidden layer of 100 dimensions, 100-dimensional randomly initialised word embeddings, a label embedding size of 100. We train our models with RMSProp, a learning rate of 0.001, a batch size of 128, and early stopping on the validation set of the main task with a patience of 3.

5 Results

Our main results are shown in Table 3 with a comparison against the state of the art. We present the results of our multi-task learning network with label embeddings (MTL + LEL), multi-task learning with label transfer (MTL + LEL + LTN), and the semi-supervised extension of this model. On 7/8 tasks, at least one of our architectures is better than single-task learning; and in 4/8, all our architectures are much better than single-task learning.

The state-of-the-art systems we compare against are often highly specialised, task-dependent architectures. Our architectures, in contrast, have not been optimised to compare favourably against the state of the art, as our main objective is to develop a novel approach to multi-task learning leveraging synergies between label sets and knowledge of marginal distributions from unlabeled data. For example, we do not use pre-trained word embeddings (Augenstein et al., 2016; Palogiannidi et al., 2016; Vo and Zhang, 2015), class weighting to deal with label imbalance (Balikas and Amini, 2016), or domain-specific sentiment lexicons (Brun et al., 2016; Kumar et al., 2016). Nevertheless, our approach outperforms the state-of-the-art on two-way topic-based sentiment analysis (Topic-2).

The poor performance compared to the state-of-the-art on FNC and MultiNLI is expected; as we alternate among the tasks during training, our model only sees a comparatively small number of examples of both corpora, which are one and two orders of magnitude larger than the other datasets. For this reason, we do not achieve good performance on these tasks as main tasks, but they are still useful as auxiliary tasks as seen in Table 4.

6 Analysis

6.1 Label Embeddings

Our results above show that, indeed, modelling the similarity between tasks using label embeddings sometimes leads to much better performance. Figure 2 shows why. In Figure 2 we visualise the label embeddings of an MTL+LEL model trained on all tasks, using PCA. As we can see, similar labels are clustered together across tasks, e.g.
there are two positive clusters (middle-right and top-right), two negative clusters (middle-left and bottom-left), and two neutral clusters (middle-top and middle-bottom).

Our visualisation also provides us with a picture of what auxiliary tasks are beneficial, and to what extent we can expect synergies from multi-task learning. For instance, the notion of positive sentiment appears to be very similar across the topic-based and aspect-based tasks, while the conceptions of negative and neutral sentiment differ. In addition, we can see that the model has failed to learn a relationship between MultiNLI labels and those of other tasks, possibly accounting for its poor performance on the inference task. We did not evaluate the correlation between label embeddings and task performance, but Bjerva (2017) recently suggested that mutual information of target and auxiliary task label sets is a good predictor of gains from multi-task learning.

6.2 Auxiliary Tasks

For each task, we show the auxiliary tasks that achieved the best performance on the development data in Table 4. In contrast to most existing work, we did not restrict ourselves to performing multi-task learning with only one auxiliary task (Søgaard and Goldberg, 2016; Bingel and Søgaard, 2017). Indeed we find that most often a combination of auxiliary tasks achieves the best performance. In-domain tasks are less used than we assumed; only Target is consistently used by all Twitter main tasks. In addition, tasks with a higher number of labels, e.g. Topic-5 are used more often. Such tasks provide a more fine-grained reward signal, which may help in learning representations that generalise better. Finally, tasks with large amounts of training data such as FNC-1 and MultiNLI are also used more often. Even if not directly related, the larger amount of training data that can be indirectly leveraged via multi-task learning may help the model focus on relevant parts of the representation space (Caruana, 1993). These observations shed additional light on when multi-task learning may be useful that go beyond existing studies (Bingel and Søgaard, 2017).

6.3 Ablation analysis

We now perform a detailed ablation analysis of our model, the results of which are shown in Table 5. We ablate whether to use the LEL (+ LEL), whether to use the LTN (+ LTN), whether to use the LEL output or the main model output for prediction (main model output is indicated by main model), and whether to use the LTN as a regulariser or for semi-supervised learning (semi-supervised learning is indicated by semi). We further test whether to use diversity features (– diversity feats) and whether to use main model predictions for the LTN (+ main model feats).

Overall, the addition of the Label Embedding Layer improves the performance over regular MTL in almost all cases.

6.4 Label transfer network

To understand the performance of the LTN, we analyse learning curves of the relabelling function vs. the main model. Examples for all tasks without semi-supervised learning are shown in Figure 3. One can observe that the relabelling...
Table 5: Ablation results with task-specific evaluation metrics on test set with early stopping on dev set. LTN means the output of the relabelling function is shown, which does not use the task predictions, only predictions from other tasks. LTN + main preds feats means main model predictions are used as features for the relabelling function. LTN, main model means that the main model predictions of the model that trains a relabelling function are used. Note that for MultiNLI, we down-sample the training data. *: lower is better. Bold: best. Underlined: second-best.

Table 6: Error analysis of LTN with and without semi-supervised learning for all tasks. Metric shown: percentage of correct predictions only made by either the relabelling function or the main model, respectively, relative to the number of all correct predictions.

To further analyse the performance of the LTN, we look into to what degree predictions of the main model and the relabelling model for individual instances are complementary to one another. Or, said differently, we measure the percentage of correct predictions made only by the relabelling model or made only by the main model, relative to the number of correct predictions overall. Results of this for each task are shown in Table 6 for the LTN with and without semi-supervised learning. One can observe that, even though the relabelling function overall contributes to the score to

Figure 3: Learning curves with LTN for selected tasks, dev performances shown. The main model is pre-trained for 10 epochs, after which the relabelling function is trained.
a lesser degree than the main model, a substantial number of correct predictions are made by the relabelling function that are missed by the main model. This is most prominently pronounced for ABSA-R, where the proportion is 14.6.

7 Conclusion

We have presented a multi-task learning architecture that (i) leverages potential synergies between classifier functions relating shared representations with disparate label spaces and (ii) enables learning from mixtures of labeled and unlabeled data. We have presented experiments with combinations of eight pairwise sequence classification tasks. Our results show that leveraging synergies between label spaces sometimes leads to big improvements, and we have presented a new state of the art for topic-based sentiment analysis. Our analysis further showed that (a) the learned label embeddings were indicative of gains from multi-task learning, (b) auxiliary tasks were often beneficial across domains, and (c) label embeddings almost always led to better performance. We also investigated the dynamics of the label transfer network we use for exploiting the synergies between disparate label spaces.

Acknowledgments

Sebastian Ruder is supported by the Irish Research Council Grant Number EBPPG/2014/30 and Science Foundation Ireland Grant Number SFI/12/RC/2289. Anders Søgaard is supported by the ERC Starting Grant Number 313695. Isabelle Augenstein is supported by Eurostars grant number E10138. We further gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References


Ayush Kumar, Sarah Kohail, Amit Kumar, Asif Ekbal, and Chris Biemann. 2016. IIT-TUDA at SemEval-2016 Task 5: Beyond Sentiment Lexicon: Combining Domain Dependency and Distributional Semantics Features for Aspect Based Sentiment Analysis. *Proceedings of SemEval*.


