To Share or not to Share:
Predicting Sets of Sources for Model Transfer Learning

Lukas Lange[1,2,3]   **Jannik Strötgen**[1]   **Heike Adel**[1]   **Dietrich Klakow**[2]

[1] Bosch Center for Artificial Intelligence, Renningen, Germany
[2] Spoken Language Systems (LSV), Saarland University, Saarbrücken, Germany
[3] Saarbrücken Graduate School of Computer Science, Saarbrücken, Germany
`{Lukas.Lange,Heike.Adel,Jannik.Stroetgen}@de.bosch.com`
`dietrich.klakow@lsv.uni-saarland.de`

Abstract

In low-resource settings, model transfer can help to overcome a lack of labeled data for many tasks and domains. However, predicting useful transfer sources is a challenging problem, as even the most similar sources might lead to unexpected negative transfer results. Thus, ranking methods based on task and text similarity — as suggested in prior work — may not be sufficient to identify promising sources. To tackle this problem, we propose a new approach to automatically determine which and how many sources should be exploited. For this, we study the effects of model transfer on sequence labeling across various domains and tasks and show that our methods based on model similarity and support vector machines are able to predict promising sources, resulting in performance increases of up to 24 F_1 points.

1 Introduction

For many natural language processing applications in non-standard domains, only little labeled data is available. This even holds for high-resource languages like English (Klie et al., 2020). The most popular method to overcome this lack of supervision is transfer learning from high-resource tasks or domains. This includes the usage of resources from similar domains (Ruder and Plank, 2017), domain-specific pretraining on unlabeled text (Gururangan et al., 2020), and the transfer of trained models to a new domain (Bingel and Søgaard, 2017). While having the choice among various transfer sources can be advantageous, it becomes more challenging to identify the most valuable ones as many sources might lead to negative transfer results, i.e., actually reduce performance (Pruksachatkun et al., 2020).

Current methods to select transfer sources are based on text or task similarity measures (Dai et al., 2019; Schröder and Biemann, 2020). The underlying assumption is that similar texts and tasks can support each other. An example for similarity
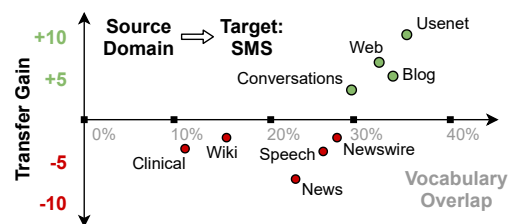

Figure 1: Observed transfer gains by transferring models from a source corpus to SMS texts. The domains are sorted by their vocabulary overlap to the target.

based on vocabulary overlap is shown in Figure 1. However, current methods typically consider text and task similarity in isolation, which limits their application in transfer settings where both the task and the text domain change.

Thus, as a first major contribution, this paper proposes a new model similarity measure that represents text and task similarity jointly. By learning a mapping between two neural models, it captures similarity between domain-specific models across tasks. We perform experiments for different transfer settings, namely zero-shot model transfer, supervised domain adaptation and cross-task transfer across a large set of domains and tasks. Our newly proposed similarity measure successfully predicts the best transfer sources and outperforms existing text and task similarity measures.

As a second major contribution, we introduce a new method to automatically determine which and how many sources should be used in the transfer process, as the transfer can benefit from multiple sources. Our selection method overcomes the limitations of current transfer methods, which solely predict single sources based on rankings. We show the benefits of transfer from sets of sources and demonstrate that support vector machines are able to predict the best sources across domains and tasks. This improves performance with absolute gains of up to 24 F_1 points and effectively prevents negative transfer.

8744

The code for our sequence taggers, prediction methods and the results are publicly available.[1]

2 Related Work

Domain adaptation & transfer learning are typically performed by transferring information and knowledge from a high-resource to a low-resource domain or task (Daumé III, 2007; Ruder, 2019; Hedderich et al., 2021). Recent approaches can be divided into two groups: (i) model transfer (Ruder and Plank, 2017) by reusing trained task-specific weights (Vu et al., 2020; Lange et al., 2021) or by first adapting models on the target domain before training the downstream task (Gururangan et al., 2020; Rietzler et al., 2020) and (ii) multi-task training (Collobert and Weston, 2008) where multiple tasks are trained jointly by learning shared representations (Peng and Dredze, 2017; Meftah et al., 2020). We follow the first approach in this paper.

For transfer learning, the **selection of sources** is utterly important. Text and task similarity measures (Ruder and Plank, 2017; Bingel and Søgaard, 2017) are used to select the best sources for cross-task transfer (Jiang et al., 2020), multi-task transfer (Schröder and Biemann, 2020), cross-lingual transfer (Chen et al., 2019) and language modeling (Dai et al., 2019). Alternatively, neural embeddings for corpora can be compared (Vu et al., 2020). In prior work, the set of domains is usually limited and the focus is on the single-best source. In contrast, we exploit sources from a larger set of domains and also explore the prediction of sets of sources, as using multiple sources is likely to be beneficial, as also shown by Parvez and Chang (2021) contemporaneously to this work.

3 Similarity Measures and Predictors

In this section, we describe the sequencer tagger model and similarity measures along with metrics for the evaluation. Finally, we introduce our new prediction method for sets of transfer sources.

3.1 Terminology

We consider two dimensions of datasets: the task T, which defines the label set, and the input text coming from a specific domain D. We thus define a dataset as a tuple $\langle T, D \rangle$, and specify in our experiments which of the two dimensions are changed.

3.2 Similarity Measures

We apply the following measures to rank sources according to their similarity with the target data.

Baselines. We use the most promising domain similarity measures reported by Dai et al. (2020): *Vocabulary* and *Annotation overlap*, *Language model perplexity* (Baldwin et al., 2013), *Dataset size* (Bingel and Søgaard, 2017) and *Term distribution* (Ruder and Plank, 2017). We also compare to domain similarity via *Text embedding*s and task similarity using *Task embedding*s (Vu et al., 2020).

Model similarity. As a new strong method, we propose *Model similarity* that is able to combine domain and task similarity. For this, feature vectors f for a target dataset t are extracted from the last layer of two models m_s, m_t which have been trained on the source and target datasets, respectively. The features are then aligned by a linear transformation W, a learned mapping, between the feature spaces using the Procrustes method (Schönemann, 1966) to minimize their pointwise differences:

$$\arg \min_W |W(f(m_s, t)) - f(m_t, t)| \quad (1)$$

The resulting transformation W is the optimal mapping between the features $f(m_s, t)$ to $f(m_t, t)$. If both feature spaces are the same, W would be the identity matrix I, i.e., no change is required for the transformation. Larger changes indicate dissimilarity, thus the distance between the two models is the difference of the mapping W and the identity matrix I: $\text{diff}(m_s, m_t) := |W - I|$.

Similar mappings have been used for the alignment of different embedding spaces (Mikolov et al., 2013; Artetxe et al., 2018) as they inherently carry information on the relatedness between models.

3.3 Prediction Methods for Sets of Sources

While these similarity measures can be applied to create rankings and select similar datasets, they still have a major shortcoming in practice: None of them provides explicit insights when positive or negative transfer can be expected.

Typically, the most similar source is selected for training based on a given similarity measure. We call this method *Top*-1. This might introduce only a low risk of selecting a negative transfer source, but it also cannot benefit from further positive transfer sources. Thus, we also test its extension to an arbitrary selection of the n best sources denoted by *Top*-n. However, it is unclear how to choose n,

and increasing n comes with the risk of including sources that lead to negative transfer results.

As a solution, we propose two methods that predict whether positive transfer is likely for a given distance between datasets: The first method models the prediction as a 3-class classification task, and the second one as a regression task predicting the transfer gain. For classification, we split the transfer gain g into the three classes positive ($g \geq \theta$), neutral ($|g| < \theta$) and negative ($g \leq -\theta$) based on a predefined threshold θ. (In our experiments, we set $\theta = 0.5$.) We introduce the neutral class for classification to cope with small transfer gains $|g| < \theta$ that do not provide additional information, but would increase the training time.

To solve these tasks, we propose to use support-vector machines (SVM) for classification (-C) and regression (-R) and compare to k-nearest-neighbour classifiers (k-NN) as well as logistic and linear regression in our experiments.[2] For each method, the input to the model is a similarity value between source and target. The training label is either the observed transfer gain (for regression) or the corresponding class (for classification) for the source-target pair. Given a new similarity value, a trained model can then be used to predict which kind of transfer can be expected.[3] The predictions for a target and a set of sources can then be used to select the subset of sources with expected positive transfer.

4 Experimental Setup

In this section, we introduce the tasks, datasets and transfer settings used in our experiments.

4.1 Tasks and Evaluation Metrics

We perform experiments on 33 datasets for three tasks: Named entity recognition (NER), part-of-speech tagging (POS), and temporal expression extraction (TIME).

For TIME tagging and for POS tagging, we use the English corpora described by Strötgen and Gertz (2016) and the four publicly available universal dependencies corpora with the UPOS tag (Nivre et al., 2016), respectively. Following Lange et al. (2020), we convert the TIMEX corpora into the BIO format for sequence tagging. For NER with different label sets, we collected several datasets

from a wide range of domains, including clinical (I2B2, Stubbs and Uzuner, 2015), social media (WNUT, Strauss et al., 2016a) and materials science corpora (SOFC, Friedrich et al., 2020). The GUM (Zeldes, 2017) and ACE'05 (Walker et al., 2006) can be split easily into multiple domains. Thus, we perform experiments for all subcorpora. The GUM corpus has multi-layer annotations and includes named entity annotations as well. We use this to study the effects of NER transfer when the label set is shared. All datasets are listed in the appendix with information on their domain and size with respect to the label set and number of sentences in the training, development, and test splits.

The metric for all experiments is micro F_1. We use the difference in F_1 to measure transfer effects and also report *transfer gain* (Vu et al., 2020), i.e., the relative improvement of a transferred model compared to the single-task performance.

In Section 5.2, we rank sources according to their similarity to the target. These rankings are evaluated with two metrics, following Vu et al. (2020): (1) the average rank of the best performing model in the predicted ranking denoted by ρ and (2) the normalized discounted cumulative gain (NDCG, Järvelin and Kekäläinen, 2002). The latter is a ranking measure commonly used in information retrieval, which evaluates the complete ranking while ρ only considers the top.

4.2 Sequence Tagger Model

For sequence tagging, we follow Devlin et al. (2019) and use BERT-base-cased as the feature extractor and a linear mapping to the label space followed by a softmax as the classifier.

Models are trained using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $2e - 5$. The training is performed for a maximum of 100 epochs. We apply early stopping after 5 epochs without change of the F_1-score on the development set. We use the same hyperparameters across all settings.[4]

4.3 Transfer Settings

Zero-shot model transfer. We apply a model trained on a source dataset to a target with the same task but a different domain: $\langle T_i, D_i \rangle \rightarrow \langle T_i, D_j \rangle$.

Task	Min.		Avg.		Max.	
Zero-shot Model Transfer						
NER	-57.3	(-37.9)	-17.7	(-10.1)	18.1	(8.0)
POS	-8.7	(-8.4)	-2.8	(-2.7)	1.6	(1.5)
TIME	-100.0	(-83.2)	-42.7	(-29.6)	38.6	(13.7)
Supervised Domain Adaptation						
NER	-5.2	(-2.7)	3.8	(1.9)	14.5	(6.3)
POS	-0.3	(-0.3)	0.4	(0.4)	1.8	(1.7)
TIME	-15.3	(-10.1)	3.4	(2.0)	32.7	(15.1)
Cross-Task Transfer						
NER→NER	-9.1	(-4.1)	-0.2	(-0.2)	6.8	(3.1)
POS→NER	-5.9	(-4.8)	-0.5	(-0.3)	2.6	(1.2)
TIME→NER	-7.2	(-3.3)	-0.9	(-0.5)	0.9	(0.6)

Table 1: Statistics on transfer gains (F_1 differences) for the three transfer settings. The average is aggregated over all domains and the 5 random seeds resulting in 210 task-specific experiments for NER and up to 780 for TIME.

Supervised domain adaptation. A model trained on a source domain is adapted to a target domain by finetuning its weights on target training data: $\langle T_i, D_i \rangle \rightarrow \langle T_i, D_j \rangle$.

Cross-task transfer. For applying a model to a different task, we replace the classification layer with a randomly initialized layer and adapt it to the new target task: $\langle T_i, D_i \rangle \rightarrow \langle T_j, D_j \rangle$.[5]

5 Results

This section presents the results of the different transfer settings and analyzes how similarity measures can be used to predict transfer sources.

5.1 Analysis of Transfer Performance

Table 1 shows the observed performance gains compared to the single-task performance. For zero-shot model transfer, we observe severe performance drops. In addition to domain-specific challenges, this setting is impaired by differences in the underlying annotation schemes.[6]

Supervised domain adaptation, i.e., adapting a model to the target domain, improves performance across all settings independent of the source domain. Table 1 shows that the average transfer gains are positive for all tasks and that the maximum transfer gain is 32.7 points for TIME.

---

[5] We restrict the cross-task transfer to NER targets with different label sets, as the combination of all tasks quickly becomes computationally infeasible given the large number of different settings.

[6] For example, the TIMEX2 (Ferro et al., 2005) and TIMEX3 (Pustejovsky et al., 2005) guidelines disagree about including preceding words in the annotated mentions as "in".

Distance	Model Transfer		Domain Adapt.		Cross -Task		Avg.	
	ρ	N	ρ	N	ρ	N	ρ	N
Vocabulary	**2.4**	**92.1**	**2.8**	88.9	6.4	84.9	3.9	88.7
Annotation	**2.4**	91.7	3.1	**89.3**	6.1	85.3	3.9	**89.1**
Dataset size	3.6	86.4	3.8	85.9	7.2	82.3	4.9	84.9
Term Dist	2.8	90.5	4.2	87.5	6.7	85.2	4.5	87.7
LM Perp.	3.9	85.6	3.4	88.2	5.9	84.4	4.4	86.1
Text Emb.	4.0	88.1	4.6	85.0	7.1	84.6	5.2	85.9
Task Emb.	4.1	88.5	4.7	84.8	6.6	84.5	5.1	85.6
Model Sim.	2.8	90.8	3.3	88.7	**5.1**	**85.4**	**3.7**	88.3

Table 2: Ranking results for different similarity measures in the three transfer settings. Corpus-based measures are listed first and model-based ones below. The values displayed are the average rank of the best model (ρ) and the NDCG-score (N).

The gains for cross-task transfer are smaller than for supervised domain adaptation. While we still observe some performance increases, the average transfer gains are negative for all tasks. This shows that it is likely that the adaptation of models from other tasks will decrease performance. These results demonstrate the need for reliable similarity measures and methods to predict the expected transfer gains given the source task and domain. We will explore them in Section 5.2 and Section 5.3.

5.2 Similarity-based Ranking

To evaluate the prospects of different sources for model transfer, we compute the pairwise distances between all datasets using the similarity measures presented in Section 3.2 and rank them accordingly.

Table 2 shows that the text-based methods vocabulary and annotation overlap are most suited for in-task transfer, i.e., model transfer and domain adaptation, while our model similarity is most useful for cross-task transfer. This shows that task similarity alone is not the most decisive factor for predicting promising transfer sources and domain similarity is equally or even more important, in particular, when more distant domains are considered. Our model similarity is able to capture both properties. It is the best similarity measure on average across all transfer settings according to the predicted rank of the top-performing source (ρ) and the best neural method according to NDCG. A more fine-grained analysis is given in Table 7 in the appendix.

In general, we find that selecting only the top source(s) based on a ranking from a distance measure, as done in current research, gives no informa-

Target	Method	Sources	F_1 (increase)
TIME - ACE-UN	Single-Task Performance	*No source corpora for pretraining*	60.5 F_1
	Top-1	WWARS	+ 10.9
	Top-2	WWARS, ACE-BN	+ 11.2
	Top-3	WWARS, ACE-BN, TIMEBANK	+ 15.8
	All	WWARS, ACE-ALL, TIMEBANK, AQUAINT, I2B2-TIME, ANCIENT, TIME4SCI, TIME4SMS	+ 10.2
	SVM (Classifier)	WWARS, ACE-ALL, TIMEBANK, AQUAINT, TIME4SCI, TIME4SMS	**+ 24.0**
	Logistic Regression	WWARS, ACE-ALL, TIMEBANK, AQUAINT, ANCIENT, TIME4SCI, TIME4SMS	+ 18.2
	k-Nearest-Neighbor	WWARS, ACE-ALL, TIMEBANK, AQUAINT, I2B2-TIME, TIME4SCI, TIME4SMS	+ 17.1
	SVM (Regression)	WWARS, ACE-ALL, TIMEBANK, AQUAINT, TIME4SCI, TIME4SMS	**+ 24.0**
	Linear Regression	WWARS, ACE-ALL, TIMEBANK, AQUAINT, ANCIENT, TIME4SCI, TIME4SMS	+ 18.2

Table 3: Predicted transfer sources for TIME domain adaptation with target ACE-UN (usenet).

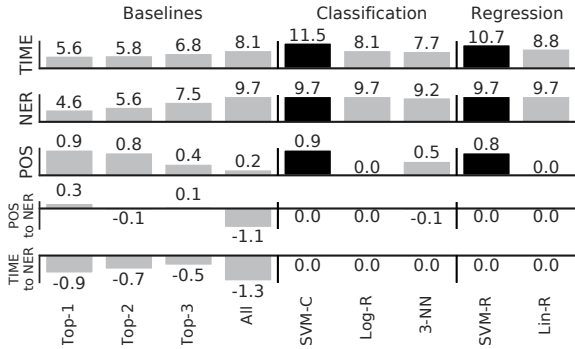|	Baselines				Classification			Regression	
	Top-1	Top-2	Top-3	All	SVM-C	Log-R	3-NN	SVM-R	Lin-R
TIME	5.6	5.8	6.8	8.1	11.5	8.1	7.7	10.7	8.8
NER	4.6	5.6	7.5	9.7	9.7	9.7	9.2	9.7	9.7
POS	0.9	0.8	0.4	0.2	0.9	0.0	0.5	0.8	0.0
POS to NER	0.3	0.1	-0.1	-1.1	0.0	0.0	-0.1	0.0	0.0
TIME to NER	-0.9	-0.7	-0.5	-1.3	0.0	0.0	0.0	0.0	0.0

Figure 2: Average transfer gains using different classifiers for predicting the set of most promising sources.

tion on whether to expect positive transfer. Thus, we now explore methods to automatically predict sets of promising sources.

5.3 Prediction of Sets of Sources

We use the methods introduced in Section 3.3 to predict the set of most promising sources. Then, we train a model on the combination of the selected sources and adapt it to the target.[7]

The results averaged across the different settings are visualized in Figure 2. While NER and TIME targets benefit from training on many sources, POS tagging targets gain the most from using only one or two of the most related source domains. We find that our methods based on SVMs are able to predict this behavior and assign fewer sources for POS targets, and more sources for TIME and NER settings. In particular, for TIME settings, our methods SVM-C and -R result in much higher transfer gains compared to the static ranking-based methods and other classifiers or regression models.

For example, transferring multiple sources using our SVM classifier to the ACE-UN target (see

Table 3) increases performance from 60.5 F_1 for single-task training to 84.5 F_1 (+24.0), which is much higher than the 10.9 points increase when using the single best source or 10.2 points using all available sources.

For the cross-task experiments in the lower part of Figure 2, we find that even the inclusion of the single best-ranked model results in a transfer loss of -0.9 points on average for TIME→NER. In this setting, our models correctly adapt to this new challenge and predict an empty set of sources, indicating that no transfer should be performed.

6 Conclusion

We explored different transfer settings across three sequence-labeling tasks and various domains. Our new model similarity measure based on feature mappings outperforms currently used similarity measures as it is able to capture both task and domain similarity at the same time. We further addressed the automatic selection of sets of sources as well as the challenge of potential negative transfer by proposing a selection method based on support vector machines. Our method results in performance gains of up to 24 F_1 points.

---

[7]We do not explore the NER to NER setting, as we restrict the sources to have the same set of labels. For the other tasks, we trained source combinations which were predicted by at least one model (SVM-R/C, Log-/Lin-R, k-NN) or baseline method (Top-1, 2, .., All). Training all possible combinations would be be infeasible.

References

Lars Ahrenberg. 2015. Converting an English-Swedish parallel treebank to Universal Dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 10–19, Uppsala, Sweden. Uppsala University.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 5012–5019, New Orleans, Louisiana, USA. AAAI Press.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources?

In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL 2019)*, pages 2138–2144, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (EACL 2017)*, pages 164–169, Valencia, Spain. Association for Computational Linguistics.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, pages 160–167, New York, New York, USA. Association for Computing Machinery.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*, pages 758–759, New York, New York, USA. Association for Computing Machinery.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using similarity measures to select pre-training data for NER. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL 2019)*, pages 1460–1470, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. Cost-effective selection of pretraining data: A case study of pretraining BERT on social media. In *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, Online. 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 (ACL 2007)*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text (WNUT 2017)*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL 2019)*, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. 2005. TIDES 2005 standard for the annotation of temporal expressions. *The MITRE Corporation*.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Maruszczyk, and Lukas Lange. 2020. The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 1255–1268, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 8342–8360, Online. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. pages 2545–2568.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Zhengbao Jiang, Wei Xu, Jun Araki, and Graham Neubig. 2020. Generalizing natural language analysis through span-relation representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 2120–2133, Online. Association for Computational Linguistics.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2020. From zero to hero: Human-in-the-loop entity linking in low resource domains. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 6982–6993, Online. Association for Computational Linguistics.

Lukas Lange, Heike Adel, and Jannik Strötgen. 2021. Boosting transformers for job expression extraction and classification in a low-resource setting. In *Proceedings of The Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

Lukas Lange, Anastasiia Iurshina, Heike Adel, and Jannik Strötgen. 2020. Adversarial alignment of multilingual models for extracting temporal expressions from text. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 103–109, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, (ICLR 2019)*, New Orleans, Louisiana, USA. OpenReview.net.

Pawel Mazur and Robert Dale. 2010. WikiWars: A new corpus for research on temporal expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, Cambridge, Massachusetts, USA.

Sara Meftah, Nasredine Semmar, Mohamed-Ayoub Tahiri, Youssef Tamaazousti, Hassane Essafi, and Fatiha Sadat. 2020. Multi-task supervised pretraining for neural domain adaptation. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media (SocialNLP 2020)*, pages 61–71, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations (ICLR 2013)*, Scottsdale, Arizona, USA.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 20h16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Md Rizwan Parvez and Kai-Wei Chang. 2021. Evaluating the values of sources in transfer learning. *arXiv preprint arXiv:2104.12567*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, RON Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Nanyun Peng and Mark Dredze. 2017. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP (RepL4NLP 2017)*, pages 91–100, Vancouver, Canada. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 5231–5247, Online. Association for Computational Linguistics.

James Pustejovsky, Robert Ingria, Roser Saurí, José Castaño, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language TimeML. In *The language of time: a reader*, pages 545–557. Oxford University Press.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4933–4941, Marseille, France. European Language Resources Association (ELRA).

Sebastian Ruder. 2019. *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis, National University of Ireland, Galway.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 372–382, Copenhagen, Denmark. Association for Computational Linguistics.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In

Proceedings of the Australasian Language Technology Association Workshop (ALTA 2015), pages 84–90, Parramatta, Australia.

Manuela Sanguinetti and Cristina Bosco. 2014. Converting the parallel treebank ParTUT in Universal Stanford Dependencies. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014)*, pages 316–321, Pisa, Italy. Pisa University Press.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Fynn Schröder and Chris Biemann. 2020. Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 2971–2985, Online. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, page 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016a. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016b. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT 2016)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.

Jannik Strötgen, Thomas Bögel, Julian Zell, Ayser Armiti, Tran Van Canh, and Michael Gertz. 2014. Extending HeidelTime for temporal expressions referring to historic dates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2390–2397, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Jannik Strötgen and Michael Gertz. 2016. Domain-sensitive temporal tagging. *Synthesis Lectures on Human Language Technologies*, 9(3):1–151.

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.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Jeniya Tabassum, Sydney Lee, Wei Xu, and Alan Ritter. 2020. WNUT-2020 task 1 overview: Extracting entities and relations from wet lab protocols. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT 2020)*, pages 260–267, Online. 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 (CoNLL 2003)*, pages 142–147.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Adam Trischler Alessandro Sordoni, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across nlp tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 7882–7926, Online. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A Datasets

All datasets are listed in Table 4 with information on their domain and size with respect to the label set and number of sentences. We take the last 20% and 10% of the training data as test or development data whenever no split was provided.

Task	Corpus	Domain	# Labels	# Train / Dev / Test sentences
NER	CONLL (Tjong Kim Sang and De Meulder, 2003) News		9	14,987 / 3,466 / 3,684
	I2B2-CLIN (Uzuner et al., 2011)	Clinical concepts	7	13,052 / 3,263 / 27,625
	I2B2-ANON (Stubbs and Uzuner, 2015)	Clinical anonymization	47	45,443 / 5,439 / 32,587
	WNUT-16 (Strauss et al., 2016b)	Twitter posts	21	2,394 / 1,000 / 3,850
	WNUT-17 (Derczynski et al., 2017)	Social media	13	3.394 / 1,009 / 1.287
	WNUT-20 (Tabassum et al., 2020)	Wetlab protocols	37	8.444 / 2,862 / 2,813
	LITBANK (Bamman et al., 2019)	Literature	13	5.549 / 1.388 / 2.973
	SEC (Salinas Alvarado et al., 2015)	Financial	9	825 / 207 / 443
	SOFC (Friedrich et al., 2020)	Materials science	9	490 / 123 / 263
NER& POS	GUM-ALL (Zeldes, 2017)	All (GUM)	23/17	3,883 / 960 / 2,060
	GUM-ACAD	Academic	23/17	321 / 81 / 173
	GUM-BIO	Biography	23/17	434 / 106 / 233
	GUM-FICT	Fiction	23/17	576 / 144 / 309
	GUM-INT	Interview	23/17	599 / 150 / 321
	GUM-NEWS	News	23/17	360 / 91 / 194
	GUM-RED	Reddit	23/17	500 / 126 / 269
	GUM-TRAV	Travel	23/17	431 / 108 / 232
	GUM-WHOW	Wikihow	23/17	612 / 154 / 329
POS	EWT (Silveira et al., 2014)	Blog. Email. Social	17	12,514 / 1,998 / 2.074
	LINES (Ahrenberg, 2015)	(non-)Fiction. spoken	17	2,738 / 912 / 914
	PARTUT (Sanguinetti and Bosco, 2014)	Legal. News. Wikipedia	17	1,781 / 156 / 153
Temporal Expressions	TIMEBANK (UzZaman et al., 2013)	News	9	2,557 / 640 / 303
	AQUAINT (UzZaman et al., 2013)	News	9	972 / 243 / 522
	ANCIENT (Strötgen et al., 2014)	Historical Wikipedia	9	456 / 114 / 245
	WWARS (Mazur and Dale, 2010)	Wikipedia	9	2,788 / 697 / 1,494
	TIME4SMS (Strötgen and Gertz, 2013)	SMS	9	1,674 / 419 / 898
	TIME4SCI (Strötgen and Gertz, 2013)	Clinical	9	461 / 116 / 248
	I2B2-TIME (Sun et al., 2013)	Clinical	9	5,943 / 1,486 / 5,665
	ACE-ALL (Walker et al., 2006)	All (ACE-05)	9	8,958 / 2,241 / 4,802
	ACE-BC	Broadcast conversations	9	1,655 / 414 / 887
	ACE-BN	Broadcast news	9	2,087 / 522 / 1,119
	ACE-CTS	Conversational telephony	9	1,756 / 440 / 942
	ACE-NW	Newswire	9	1,172 / 293 / 628
	ACE-UN	Usenet	9	1,168 / 292 / 626
	ACE-WB	Webblog	9	1,120 / 280 / 600

Table 4: Overview of dataset domains and their sizes used in the transfer experiments.

Corpus	BERT		
	Pre	Rec	F_1
CONLL	90.5	91.9	91.2
WNUT-20	78.2	81.0	79.6
WNUT-17	60.1	35.9	44.9
WNUT-16	46.8	44.7	45.7
I2B2-CLIN	82.0	85.8	83.9
I2B2-ANON	94.7	93.2	94.0
SEC	76.7	87.9	81.9
LITBANK	66.1	74.5	70.0
SOFC	73.3	82.8	77.8
GUM-ACAD	46.3	58.8	51.8
GUM-BIO	61.0	72.1	66.1
GUM-FICT	62.8	72.0	67.1
GUM-INT	48.9	58.7	53.4
GUM-NEWS	43.7	52.7	47.8
GUM-RED	50.5	61.9	55.6
GUM-TRAV	37.7	51.0	43.3
GUM-WHOW	40.0	49.0	44.0
GUM-ALL	55.1	64.3	59.4

(a) Named Entity Recognition

Corpus	BERT		
	Pre	Rec	F_1
TIMEBANK	75.2	76.3	75.7
AQUAINT	77.6	77.6	77.6
ANCIENT	71.8	79.6	75.5
WWARS	87.1	90.7	88.9
TIME4SMS	63.8	68.4	66.0
TIME4SCI	55.9	51.6	53.7
I2B2-TIME	72.2	76.7	74.4
ACE-BC	60.5	64.0	62.2
ACE-BN	60.3	71.5	65.4
ACE-CTS	39.0	55.6	45.8
ACE-NW	76.8	81.9	79.2
ACE-UN	56.5	65.2	60.5
ACE-WB	65.6	69.4	67.5
ACE-ALL	66.9	77.6	71.8

(b) Temporal Expression Extraction

Corpus	BERT
	F_1 =Acc.
PARTUT	96.9
EWT	97.0
LINES	97.5
GUM-ACAD	95.1
GUM-BIO	96.3
GUM-FICT	96.8
GUM-INT	95.5
GUM-NEWS	95.9
GUM-RED	94.6
GUM-TRAV	94.5
GUM-WHOW	94.9
GUM-ALL	96.5

(c) POS Tagging

Table 5: Single task learning performance.

Task	Min.	Avg.	Max.	# Pos.	Pos. Avg.	# Neg.	Neg. Avg.
Zero-Shot Model Transfer							
NER	-57.3 (-37.9)	-17.7 (-10.1)	18.1 (8.0)	7 /64	8.6 (3.8)	56 /64	-21.3 (-12.0)
POS	-8.7 (-8.4)	-2.8 (-2.7)	1.6 (1.5)	13 /144	0.9 (0.8)	127 /144	-3.3 (-3.2)
TIME	-100.0 (-83.2)	-42.7 (-29.6)	38.6 (17.7)	13 /196	10.3 (5.6)	183 /196	-46.5 (-31.7)
Supervised Domain Adaptation							
NER	-5.2 (-2.7)	3.8 (1.9)	14.5 (6.3)	55 /64	4.7 (2.3)	8 /64	-1.7 (-0.9)
POS	-0.3 (-0.3)	0.4 (0.4)	1.8 (1.7)	116 /144	0.5 (0.5)	9 /144	-0.1 (0.1)
TIME	-15.3 (-10.1)	3.4 (2.0)	32.7 (15.1)	133 /196	6.0 (3.7)	62 /196	-2.1 (-1.4)
Cross-Task Transfer							
NER→NER	-9.1 (-4.1)	-0.2 (-0.2)	6.8 (3.1)	39 /90	1.1 (0.6)	46 /90	-1.4 (-0.9)
POS→NER	-5.9 (-4.8)	-0.5 (-0.3)	2.6 (1.2)	42 /120	2.6 (1.2)	65 /120	-1.3 (-0.9)
TIME→NER	-7.2 (-3.3)	-0.9 (-0.5)	0.9 (0.6)	35 /150	0.4 (0.3)	100 /150	-1.5 (-0.9)

Table 6: Transfer gains (F_1 increase) and the number of positive and negative transfer scenarios for each setting.

	Zero-Shot Model Transfer						Supervised Domain Adaptation						Cross-Task Transfer						Average	
	NER		POS		TIME		NER		POS		TIME		NER→NER		POS→NER		TIME→NER			
Distance	ρ	N	ρ	N	ρ	N	ρ	N	ρ	N	ρ	N	ρ	N	ρ	N	ρ	N	ρ	N
Vocabulary	3.1	90.8	1.3	91.7	2.9	**93.9**	3.6	83.7	2.3	92.4	**2.6**	90.7	5.3	82.8	5.4	**87.7**	8.4	84.3	3.9	88.7
Annotation	3.5	89.7	1.3	91.7	**2.4**	93.8	3.9	85.1	2.3	92.4	3.1	90.5	5.3	86.2	**5.2**	**87.7**	7.9	84.0	3.9	**89.1**
Datasize	5.0	81.0	2.1	88.3	3.6	90.0	4.6	81.2	2.8	89.9	4.1	86.7	6.9	75.0	5.5	87.3	9.1	84.6	4.9	84.9
Term Dist.	4.4	86.1	**1.0**	**92.8**	2.9	92.5	5.0	83.4	3.5	90.4	4.1	88.6	4.9	87.1	6.3	85.0	8.8	83.5	4.5	87.7
LM Perp.	4.9	83.8	2.1	91.2	4.6	81.8	3.9	**86.8**	1.8	**93.5**	4.6	84.4	4.8	84.8	6.8	86.5	6.1	81.9	4.4	86.1
Text Emb.	3.5	88.8	4.0	86.1	4.4	89.4	4.0	84.3	5.0	84.1	4.9	86.5	4.7	84.3	9.6	84.7	7.0	**84.8**	5.2	85.9
Task Emb.	3.4	89.4	4.0	86.3	5.0	89.8	3.8	83.7	5.0	84.3	5.2	86.3	4.9	84.5	7.2	82.8	7.8	83.3	5.1	85.6
Model Sim.	**2.6**	**91.8**	2.7	88.7	3.2	91.9	**3.5**	85.9	3.3	88.4	3.1	**91.8**	**3.0**	**87.2**	6.6	84.5	**5.7**	84.6	**3.7**	88.3

Table 7: Ranking results for different similarity measures in the three transfer settings. The values displayed are the average rank of the best model (ρ) and the NDCG-score (N) compared to the observed performance.

B Model Performance

We list the performance for all single-task models in Table 5 with precision, recall and micro-F_1 for NER and TIME corpora and accuracy for POS.

C Fine-grained Results

In addition to Table 1 and Table 2 that display task-wise averages, we report more fine-grained results in Table 6 and Table 7.

D Similarity Measures

This section provides a more detailed overview of the similarity measures introduced in Section 3.2.

Target vocabulary overlap is the percentage of unique words from the target corpus covered in the source corpus. In contrast to vocabulary overlap, this is an asymmetric measure. *Annotation overlap* is a special case considering only annotated words.

We also experiment with the *Language model perplexity* (Baldwin et al., 2013) between two datasets. For this, a language model, in our case a 5-gram LM with Kneser–Ney smoothing (Heafield, 2011) as used by Dai et al. (2019), is trained for each source domain and tested against the target domain. The resulting perplexity gives hints how similar these domains are, i.e., a lower perplexity indicates similarity between domains.

Jensen-Shannon divergence (Ruder and Plank, 2017) compares the term distributions between two texts, which are probability distributions that capture the frequency of words. It is similar to vocabulary overlap, as it describes the textual overlap, but based on distributions instead of sets of terms.

A *Text embedding* (Vu et al., 2020) can be computed by extracting the feature vectors of a neural model. For this, the output of the last layer is averaged over all words in the dataset. This vector then represents the text domain. The distance between two vectors is computed by using cosine similarity.

The *Task embedding* (Vu et al., 2020) takes a labeled source dataset and computes a representation based on the Fisher Information Matrix, which captures the change of model parameters w.r.t. the computed loss. This method assumes that similar tasks require similar parameters changes. We use the code released by Vu et al. (2020) to compute task embeddings from the different components of our BERT models and similarly use reciprocal rank fusion (Cormack et al., 2009) to combine these.