Adapting to the Long Tail: A Meta-Analysis of Transfer Learning Research for Language Understanding Tasks

Aakanksha Naik[1,3] **Jill Lehman**[2] **Carolyn Rosé**[1,3]

[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Human-Computer Interaction Institute, Carnegie Mellon University, USA
[3]Rehabilitation Medicine Department, Clinical Center, National Institutes of Health, USA
{anaik,jfl,cp3a}@andrew.cmu.edu

Abstract

Natural language understanding (NLU) has made massive progress driven by large benchmarks, but benchmarks often leave a long tail of infrequent phenomena underrepresented. We reflect on the question: *Have transfer learning methods sufficiently addressed the poor performance of benchmark-trained models on the long tail?* We conceptualize the long tail using macro-level dimensions (underrepresented genres, topics, etc.), and perform a qualitative meta-analysis of 100 representative papers on transfer learning research for NLU. Our analysis asks three questions: (i) Which long tail dimensions do transfer learning studies target? (ii) Which properties of adaptation methods help improve performance on the long tail? (iii) Which methodological gaps have greatest negative impact on long tail performance? Our answers highlight major avenues for future research in transfer learning for the long tail. Lastly, using our meta-analysis framework, we perform a case study comparing the performance of various adaptation methods on clinical narratives, which provides interesting insights that may enable us to make progress along these future avenues.

1 Introduction

''There is a growing consensus that significant, rapid progress can be made in both text understanding and spoken language understanding by investigating those phenomena that occur most centrally in naturally occurring unconstrained materials and by attempting to automatically extract information about language from very large corpora.'' (Marcus et al., 1993)

Since the creation of the Penn Treebank, using shared benchmarks to measure and drive progress in model development has been instrumental for accumulation of knowledge in the field of natural language processing, and has become a dominant practice. Ideally, we would like shared benchmark corpora to be diverse and comprehensive, which can be addressed at two levels: (i) macro-level dimensions such as language, genre, topic, and so forth, and (ii) micro-level dimensions such as specific language phenomena like negation, deixis, causal reasoning, and so on. However, diversity and comprehensiveness are not straightforward to achieve.

According to Zipf's law, many micro-level language phenomena naturally occur infrequently and will be relegated to the *long tail*, except in cases of intentional over-sampling. Moreover, the advantages of restricting community focus to a specific set of benchmark corpora and limitations in resources lead to portions of the macro-level space being under-explored, which can further cause certain micro-level phenomena to be under-represented. For example, since most popular coreference benchmarks focus on English narratives, they do not contain many instances of zero anaphora, a phenomenon quite common in other languages (e.g., Japanese, Chinese). In such situations, model performance on benchmark corpora may not be truly reflective of expected performance on micro-level long tail phenomena, raising questions about the ability of state-of-the-art models to generalize to the long tail.

Most benchmarks do not explicitly catalogue the list of micro-level language phenomena that are included or excluded in the sample, which makes it non-trivial to construct a list of long tail micro-level language phenomena. Hence, we formalize an alternate conceptualization of the long tail: undersampled portions of the macro-level space that can be treated as proxies for long tail micro-level phenomena. These undersampled *long tail* macro-level dimensions highlight gaps and present potential new challenging directions

Transactions of the Association for Computational Linguistics, vol. 10, pp. 956–980, 2022. https://doi.org/10.1162/tacl_a_00500
Action Editor: Benjamin van Durme. Submission batch: 12/2021; Revision batch: 3/2022; Published 9/2022.

for the field. Therefore, periodically taking stock of research to identify long tail macro-level dimensions can help in highlighting opportunities for progress that have not yet been tackled. This idea has been gaining prominence recently; for example, Joshi et al. (2020) survey languages studied by NLP papers, providing statistical support for the existence of a macro-level long tail of low-resource languages.

In this work, our goal is to attempt to characterize the macro-level long tail in natural language understanding (NLU) and efforts that have tried to address it from research on transfer learning. Large benchmarks have driven much of the recent methodological progress on NLU (Bowman et al., 2015; Rajpurkar et al., 2016; McCann et al., 2018; Talmor et al., 2019; Wang et al., 2019a,b), but the generalization abilities of benchmark-trained models to the long tail have been unclear. In tandem, the NLP community has been successfully developing transfer learning methods to improve generalization of models trained on NLU benchmarks (Ruder et al., 2019). The goal of transfer learning research is to tackle the macro-level long tail in NLU, leading to the question: *How far has transfer learning addressed performance of benchmark models on the NLU long tail, and where do we still fall behind?*

Probing further, we perform a qualitative meta-analysis of a representative sample of 100 papers on domain adaptation and transfer learning in NLU. We sample these papers based on citation counts and publication venues (§2.1), and document seven facets for each paper such as tasks and domains studied, adaptation settings evaluated, and so on. (§2.2). Adaptation methods proposed (or applied) are documented using a hierarchical categorization described in §2.3, which we develop by extending the hierarchy from Ramponi and Plank (2020). With this information, our analysis focuses on three questions:

- **Q1:** What long tail macro-level dimensions do transfer learning studies target? Dimensions include tasks, domains, languages and adaptation settings covered in transfer learning research.
- **Q2:** Which properties of adaptation methods help improve performance on long tail dimensions?
- **Q3:** Which methodological gaps have greatest negative impact on long tail performance?

The rest of the paper presents thorough answers to these questions, laying out avenues for future research on transfer learning that more effectively address the macro-level long tail in NLU. We also present a case study[1] to demonstrate how our meta-analysis framework can be use to systematically design experiments that provide insights that enable us to make progress along these avenues.

2 Meta-Analysis Framework

2.1 Sample Curation

We gather a representative sample of work on domain adaptation or transfer learning in NLU from the December 2020 dump of the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020). First, we extract all abstracts published at 9 prestigious *CL venues: ACL, EMNLP, NAACL, EACL, COLING, CoNLL, SemEval, TACL, and CL. This results in 25,141 abstracts, which are filtered to retain those containing the terms ‘‘domain adaptation’’ or ‘‘transfer learning’’ in the title or abstract,[2] producing a set of 382 abstracts after duplicate removal. Figure 2 shows the distribution of these retrieved abstracts across search terms and years. From this graph we can see that interest in this field has increased tremendously in recent years, and that there has been a slight terminology shift with recent work preferring the term ‘‘transfer learning’’ over ‘‘domain adaptation’’.

We manually screen this subset and remove abstracts that are not eligible for our NLU-focused analysis (e.g., papers on generation-focused tasks like machine translation), leaving us with a set of 266 abstracts. From this, we construct a final meta-analysis sample of 100 abstracts via application of two inclusion criteria. Per the first criterion, all abstracts with 100 or more citations are included because they are likely to describe landmark advances.[3] Then, remaining abstracts (to bring our meta-analysis sample to 100) are randomly chosen, after discarding ones with no citations.[4] The random sampling criterion ensures

[1]The codebase for our case study experiment is available at: https://github.com/CC-RMD-EpiBio/Domain-Adaptation-Meta-Analysis.

[2]Search scope is limited to title and abstract in order to prefer papers that focus on transfer learning rather than ones including a brief discussion or experiment on transfer learning as part of an investigation of something else.

[3]This makes up 23% of the final meta-analysis sample.

[4]Mean citation count for randomly sampled set is 28.4.

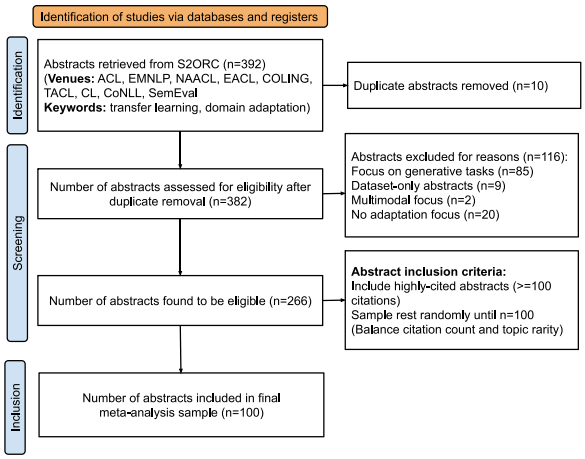


Figure 1: PRISMA diagram explaining our sample curation process.

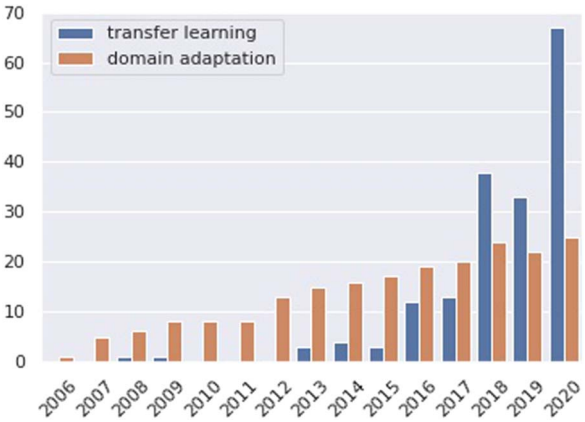


Figure 2: Distribution of papers retrieved by our search strategy across search terms and years.

that we do not neglect studies that study less mainstream topics by focusing solely on highly cited work. This produces a final representative sample of transfer learning work for our meta-analysis. Figure 1 describes our sample curation process via a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram (Page et al., 2021).

Characterizing Limitations of Our Curation Process: Since our sample curation process primarily relies on a keyword-based search, it might miss relevant work that does not use any of these keywords. To characterize the limitations of our curation process, we use two additional strategies for relevant literature identification:

- **Citation graph retrieval:** Following Blodgett et al. (2020), we include all abstracts that cite or are cited by abstracts included in our keyword-retrieved set of 382 abstracts. This retrieves 3727 additional abstracts, but many of these works are cited for their description or introduction of new tasks, datasets, evaluation metrics, etc. Therefore, we discard all works that do not have the words ‘‘adaptation’’ or ‘‘transfer’’, leaving 282 new abstracts.

- **Nearest neighbor retrieval:** We use SPECTER (Cohan et al., 2020) to compute embeddings for all abstracts included in our keyword-retrieved set, as well as all abstracts in the ACL anthology. Then we retrieve the nearest neighbor for every abstract in our keyword-retrieved set, which results in the retrieval of 262 new abstracts.

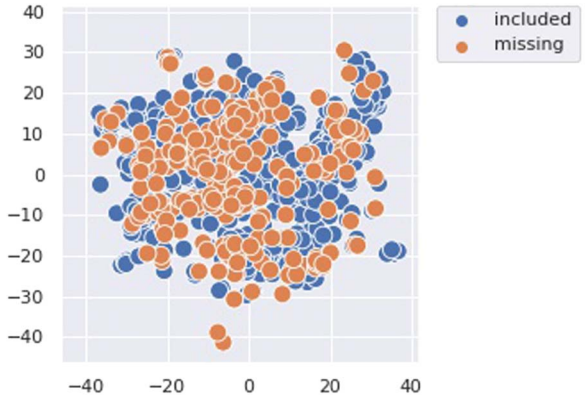


Figure 3: t-SNE visualization of our meta-analysis sample alongside additional transfer learning papers missed by our keyword search.

Combining abstracts returned by both strategies, we are able to identify 510 additional works. However, while going over them manually, we notice that despite our noise reduction efforts, not all abstracts describe transfer learning work. We perform an additional manual screening step to discard such work, which leaves us with a final set of 232 additional papers.[5] To identify whether the exclusion of these papers from the initial sample may have led to visible gaps or blind spots in our meta-analysis, we perform a TSNE visualization of SPECTER embeddings for both keyword-retrieved papers and this additional set of papers. Figure 3 presents the results of this visualization and indicates that there aren’t visible distributional differences between the two subsets.

[5]We make this subset of papers available at: `http://www.shorturl.at/uFGIY`.

Hence, though our sample curation strategy is imperfect, this seems to indicate that our final observations from the meta-analysis might not have been very different. We note that this conclusion comes with two caveats: (i) t-SNE embeddings are not always reliable, and (ii) embedding overlap does not necessarily confirm that annotations for overlapping papers are similar/correlated. Keeping these caveats in mind, we perform a spot check for additional validation. For this spot check, we consider the following highly cited large language models that have been considered to be major recent advances in transfer learning: ELMo, BERT, RoBERTa, BART, T5, ERNIE, DeBERTa, and ELECTRA. Note that we do not consider any few-shot models (GPT3, PET, etc.) since our sample only consists of work that was *accepted* to a *CL venue by December 2020. Of these major language models, RoBERTa, DeBERTa, T5, and ELECTRA were published at non-*CL venues (JMLR and ICLR), which excludes them from our sample. The remaining works (ELMo, BERT, BART, and ERNIE) are all present in the set of additional works we identified in this section, lending further support to our conclusion that our sampling strategy and subsequent analyses have not overlooked influential work.

2.2 Meta-Analysis Facets

For every paper from our meta-analysis sample, we document the following key facets:

Task(s): NLP task(s) studied in the work. Tasks are grouped into 12 categories based on task formalization and linguistic level (e.g., lexical, syntactic), as shown in Table 1.

Domain(s): Source and target domains and/or languages studied, along with datasets used for each.

Task Model: Base model used for the task, to which domain adaptation algorithms are applied.

Adaptation Method(s): Domain adaptation method(s) proposed or used in the work. Adaptation methods are grouped according to the categorization showed in Figure 4 (details in §2.3).

Adaptation Baseline(s): Baseline domain adaptation method(s) to compare new methods against.

Cat	Tasks Included
TC	Text classification tasks like sentiment analysis, hate speech detection, propaganda detection, etc.
NER	Semantic sequence labeling tasks like NER, event extraction, etc.
POS	Syntactic sequence labeling tasks like POS tagging, chunking, etc.
NLI	Natural language inference, NLU Tasks recast as NLI (e.g., GLUE)
SP	Structured prediction tasks such as entity and event coreference
WSD	Word sense disambiguation
TRN	Text ranking tasks (e.g., search)
TRG	Text regression tasks
RC	Reading comprehension
MF	Matrix factorization
LI	Lexicon induction
SLU	Spoken language understanding

Table 1: Categorization of tasks studied. Note that the matrix factorization (MF) category includes text-based recommender systems.

Adaptation Settings: Source-target transfer settings explored in the work (unsupervised adaptation, multi-source adaptation, etc.).

Result Summary: Performance improvements (if any), performance differences across multiple source-target pairs or methods, and so forth.

2.3 Adaptation Method Categorization

For adaptation methods proposed or used in each study, we assign type labels according to the categorization presented in Figure 4. This categorization is an extension of the one proposed by Ramponi and Plank (2020).[6] Broadly, methods are divided into three *coarse* categories: (i) model-centric, (ii) data-centric, and (iii) hybrid approaches. Model-centric approaches perform adaptation by modifying the structure of the model, which may include editing the feature representation, loss function, or parameters. Data-centric approaches perform adaptation by modifying or leveraging labeled/unlabeled data from the source and target domains to bridge the

[6]Since our meta-analysis is not limited to neural unsupervised domain adaptation, we need to add additional classes.

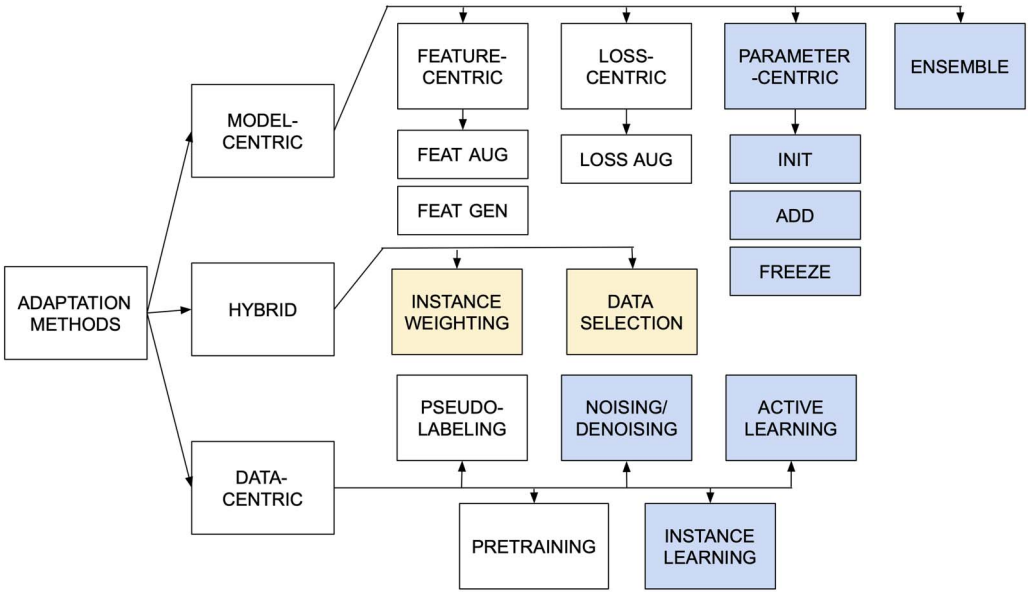


Figure 4: Categorization of adaptation methods proposed, extended, or used in all studies. This categorization is an extension of the one proposed by Ramponi and Plank (2020), with blue blocks indicating newly added categories, and yellow blocks indicating categories that have been moved to a different coarse category.

domain gap. Finally, hybrid approaches are ones that cannot be clearly classified as model-centric or data-centric. Each coarse category is divided into *fine* subcategories.

Model-centric approaches are divided into four categories, based on which portion of the model they modify: (i) feature-centric, (ii) loss-centric, (iii) parameter-centric, and (iv) ensemble. Feature-centric approaches are further divided into two fine subcategories: (i) feature augmentation, and (ii) feature generalization. Feature augmentation includes techniques that learn an alignment between source and target feature spaces using shared features called *pivots* (Blitzer et al., 2006). Feature generalization includes methods that learn a joint representation space using autoencoders, motivated by Glorot et al. (2011) and Chen et al. (2012). Loss-centric approaches contain one fine subcategory: loss augmentation. This includes techniques that augment task loss with adversarial loss (Ganin and Lempitsky, 2015; Ganin et al., 2016), multi-task loss (Liu et al., 2019), or regularization terms. Parameter-centric approaches include three fine subcategories: (i) parameter initialization, (ii) new parameter addition, and (iii) parameter freezing. Finally ensemble, used in settings with multiple source domains, includes techniques that learn to combine predictions of multiple models trained on source and target domains.

Data-centric approaches are divided into five fine subcategories. Pseudo-labeling approaches train classifiers that then produce ‘‘gold’’ labels for unlabeled target data. This includes semi-supervised learning methods such as bootstrapping, co-training, self-training, and so on. (e.g., McClosky et al., 2006). Active learning approaches use a human-in-the-loop setting to annotate a subset of target data that the model can learn most from (Settles, 2009). Instance learning approaches leverage neighborhood structure in joint source-target feature spaces to make target predictions (e.g., nearest neighbor learning). Noising/denoising approaches include data corruption/pre-processing that increase surface similarity between source and target examples. Finally, pretraining includes approaches that train large-scale language models on unlabeled data to learn better source and target representations, a strategy that has gained popularity in recent years (Gururangan et al., 2020).

Hybrid approaches contain two fine subcategories that cannot be classified as model-centric or data-centric because they involve manipulation of the data distribution, but can also be viewed as loss-centric approaches that edit the training loss. Instance weighting approaches assign weights to target examples based on similarity to source data. Conversely, data selection approaches filter target data based on similarity to source data. Table 2

Category	Example Methods	Example Studies
Feat Aug (FA)	Structural correspondence learning, Frustratingly easy domain adaptation	(Blitzer et al., 2006; Daumé III, 2007)
Feat Gen (FG)	Marginalized stacked denoising autoencoders, Deep belief networks	(Jochim and Schütze, 2014; Ji et al., 2015; Yang et al., 2015)
Loss Aug (LA)	Multi-task learning, Adversarial learning, Regularization-based methods	(Zhang et al., 2017; Liu et al., 2019; Chen et al., 2020)
Init (PI)	Prior estimation, Parameter matrix initialization	(Chan and Ng, 2006; Al Boni et al., 2015)
Add (PA)	Adapter networks	(Lin and Lu, 2018)
Freeze (FR)	Embedding freezing, Layerwise freezing	(Yin et al., 2015; Tourille et al., 2017)
Ensemble (EN)	Mixture of experts, Weighted averaging	(McClosky et al., 2010; Nguyen et al., 2014)
Instance Weighting (IW)	Classifier based weighting	(Jiang and Zhai, 2007; Jeong et al., 2009)
Data Selection (DS)	Confidence-based sample selection	(Scheible and Schütze, 2013; Braud and Denis, 2014)
Pseudo-Labeling (PL)	Semi-supervised learning, Self-training	(Umansky-Pesin et al., 2010; Lison et al., 2020)
Noising/Denoising (NO)	Token dropout	(Pilán et al., 2016)
Active Learning (AL)	Sample selection via active learning	(Rai et al., 2010; Wu et al., 2017)
Pretraining (PT)	Language model pretraining, Supervised pretraining	(Conneau et al., 2017; Howard and Ruder, 2018)
Instance Learning (IL)	Nearest neighbor learning	(Gong et al., 2016)

Table 2: Examples of methods from each category, and papers studying these methods. These lists are non-exhaustive. In the interest of replicability, we have made our coding for all papers publicly available at: `http://www.shorturl.at/stuAT`.

lists example adaptation methods for each fine category and example studies from our meta-analysis subset that use these methods.

3 Which Long Tail Macro-Level Dimensions Do Transfer Learning Studies Target?

The first goal of our meta-analysis is to document long tail macro-level dimensions that transfer learning studies have tested their methods on. We look at distributions of tasks, domains, languages, and adaptation settings studied in all papers in our sample. Ten studies are surveys, position papers or meta-experiments, and so excluded from these statistics. Studies can cover multiple tasks, domains, languages, or settings so counts may be higher than 90.

Task Distribution: Figure 5 gives a brief overview of the distribution of tasks studied across papers. Text classification tasks clearly dominate, followed by semantic and syntactic tagging. Text classification covers a variety of tasks, but sentiment analysis is the most well-studied, with research driven by the multi-domain sentiment detection (MDSD) dataset (Blitzer et al., 2007). Conversely, structured prediction is under-studied ($<$10% studies from our sample evaluate on structured prediction tasks), despite covering a variety of tasks such as coreference resolution, syntactic parsing, dependency parsing, semantic parsing, etc. This indicates that tasks with complex formulations/objectives are under-explored. We speculate that there may be two reasons for this: (i) difficulty of collecting annotated data in

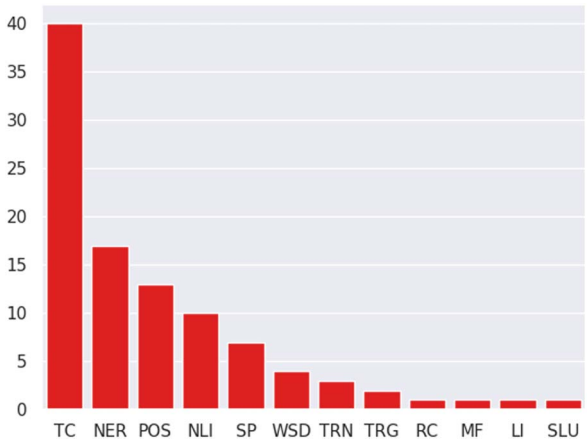


Figure 5: Distribution of papers according to tasks studied. Top three task categories are text classification (TC), semantic sequence labeling (NER), and syntactic sequence labeling (POS). Table 1 describes the remaining task categories.

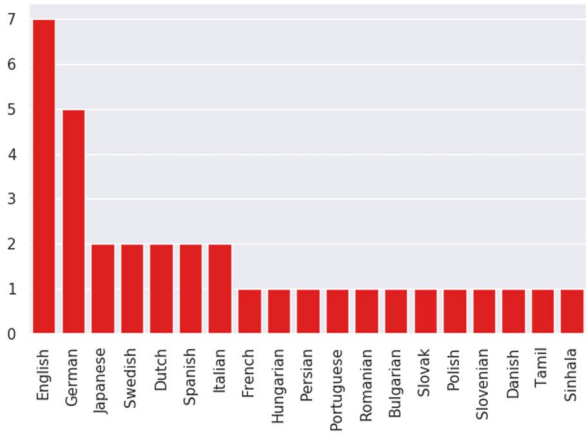


Figure 6: Distribution of multi-lingual studies according to languages included.

HE	#P	NN	#P
Clinical	10	Twitter	12
Biomedical	9	Conversations	10
Science	3	Forums	8
Finance	3	Emails	6
Literature	3		
DefSec	1		

Table 3: Counts of papers (#P) studying high-expertise (HE) and non-narrative (NN) domains (DefSec refers to security and defense reports).

multiple domains/languages for such tasks,[7] and (ii) shift in output structures (e.g., different named entity types in source and target domains) making adaptation harder.

Languages Studied: Despite a focus on generalization, most studies in our sample rarely evaluate on other languages aside from English. As stated by Bender (2011), this is problematic because the ability to apply a technique to other languages does not necessarily guarantee comparable performance. Some studies do cover multi-lingual evaluation or focus on cross-linguality. Figure 6 shows the distribution of languages included in these studies, which is a limited subset. For a more comprehensive discussion of linguistic diversity in NLP research not limited to transfer learning, we refer interested readers to Joshi et al. (2020).

Domains Studied: Many popular transfer benchmarks (Blitzer et al., 2007; Wang et al., 2019a,b) are homogeneous. They focus on narrative English, drawn from plentiful sources such as news articles, reviews, blogs, essays, and Wikipedia. This sidelines some categories of domains[8] that fall into the long tail: (i) non-narrative text (social media, conversations etc.), and (ii) texts from high-expertise domains that use specialized vocabulary and knowledge (e.g., clinical text). Table 3 shows the number of papers focusing on high-expertise and non-narrative domains, highlighting the lack of focus on these areas.

Adaptation Settings Studied: Most studies evaluate methods in a supervised adaptation setting, that is, labeled data is available from both source and target domains. This assumption may not always hold. Often adaptation must be performed in harder settings such as unsupervised adaptation (no labeled data from target domain), adaptation from multiple source domains, online adaptation, and so forth, and we refer to all such settings aside from supervised adaptation as unconventional adaptation settings. Figure 7 shows the distribution of unconventional settings across papers, indicating that these settings are understudied in literature.

[7]Note that despite these difficulties, efforts to collect data for structured prediction tasks are underway, such as the massive Universal Dependencies project, which has collected consistent grammar annotations for over 100 languages: `https://universaldependencies.org`.

[8]*Domain* is an overloaded term covering genres, styles, registers, etc., but we use it for consistency with prior work.

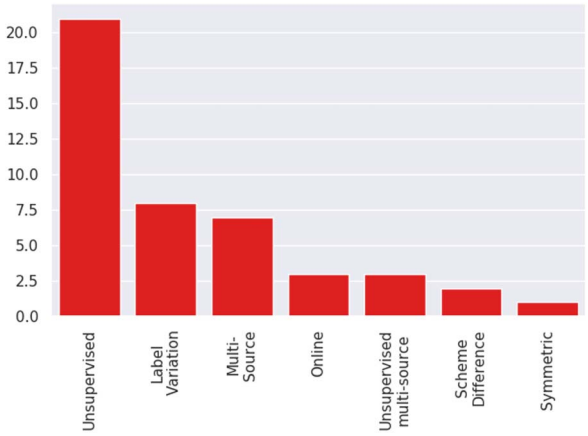


Figure 7: Distribution of papers according to unconventional (non-supervised) adaptation settings.

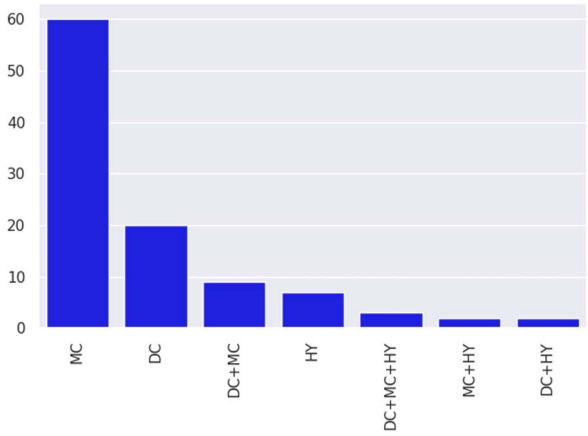


Figure 8: Distribution of transfer learning studies according to coarse method categories. DC, MC, and HY refer to data-centric, model-centric, and hybrid coarse categories, respectively.

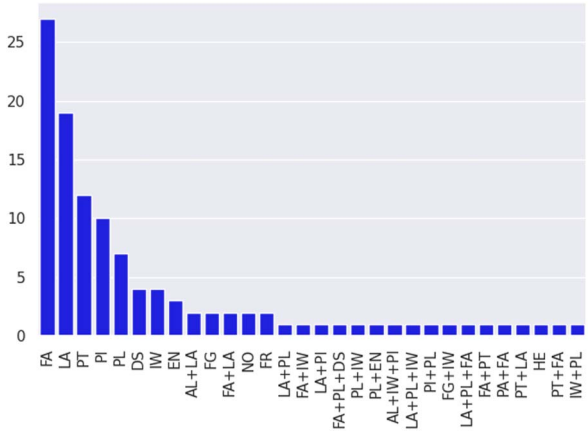
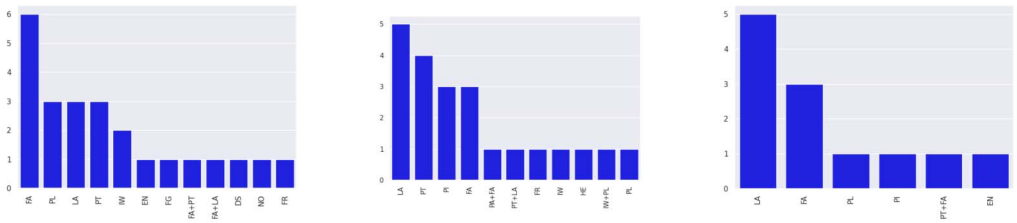


Figure 9: Distribution of transfer learning studies according to fine method categories. The top five fine categories are feature augmentation (FA), loss augmentation (LA), pretraining (PT), parameter initialization (PI), and pseudo-labeling (PL). Table 2 describes the remaining categories in more detail.

Open Issues: We can see that there is much ground to cover in testing adaptation methods on the macro long tail. Two research directions may be key to achieving this: (i) development of and evaluation on diverse benchmarks, and (ii) incentivizing publication of research on long tail domains at NLP venues. Diverse benchmark development has gained momentum, with the creation of benchmarks such as BLUE (Peng et al., 2019) and BLURB (Gu et al., 2020) for biomedical and clinical NLP, XTREME (Hu et al., 2020) for cross-lingual NLP, and GLUECoS (Khanuja et al., 2020) for code-switched NLP. However, newly proposed adaptation methods are often not evaluated on them, which is imperative to test the methods' limitations and generalization abilities. On the other hand, application-specific or domain-specific evaluations of adaptation methods are sidelined at NLP venues and may be viewed as limited in terms of bringing broader insights. But applied research can unearth significant opportunities for advances in transfer learning, and should be viewed from a *translational* perspective (Newman-Griffis et al., 2021). For example, source-free domain adaptation in which only a trained source model is available with no access to source data (Liang et al., 2020), was conceptualized partly due to data sharing restrictions on Twitter or clinical data. Though this issue is limited to certain domains, source-free adaptation may be of broader interest since it has implications for reducing models' reliance on large amounts of data. Therefore, encouraging closer ties with applied transfer learning research can help us gain more insight into limitations of existing techniques on the macro long tail.

4 Which Properties Of Adaptation Methods Help Improve Performance On Long Tail Dimensions?

The second goal of our meta-analysis is to identify which categories of adaptation methods have been tested extensively and have exhibited good performance on various long tail macro-level dimensions. Figures 8 and 9 provide an overview of categories of methods tested across all



(a) Fine method categories evaluated on high-expertise domains. The top five fine categories are FA, PL, LA, PT, and IW.

(b) Fine method categories evaluated on non-narrative domains. The top four fine categories are LA, PT, PI, and FA.

(c) Fine method categories evaluated on both domain types. The top four fine categories are LA, FA, PL, and PI.

Figure 10: Distribution of fine method categories from studies evaluating on long tail domains. Note that FA stands for feature augmentation, LA for loss augmentation, PL for pseudo-labeling, PT for pretraining, IW for instance weighting, and PI for parameter initialization. Table 2 describes the remaining categories in more detail.

papers in our subset. We can see that studies overwhelmingly develop or use model-centric methods. Within this coarse category, feature augmentation (FA) and loss augmentation (LA) are the top two categories, followed by pretraining (PT), which is data-centric. Parameter initialization (PI) and pseudo labeling (PL) round out the top five. Feature augmentation being the most explored category is no surprise, given that much pioneering early domain adaptation work in NLP (Blitzer et al., 2006, 2007; Daumé III, 2007) developed methods to learn shared feature spaces between source and target domains. Loss augmentation methods have gained prominence recently, with multi-task learning providing large improvements (Liu et al., 2015, 2019). Pretraining methods, both unsupervised (Howard and Ruder, 2018) and supervised (Conneau et al., 2017), have also gained popularity with large transformer-based language models (Peters et al., 2018; Devlin et al., 2019, etc.) achieving huge gains across tasks.

To specifically identify techniques that work on long tail domains, we look at categories of methods evaluated on high-expertise domains or non-narrative domains (or both). Figures 10a, 10b, and 10c present the distributions of fine method categories tested on high-expertise domains, non-narrative domains, and both domain types, respectively. While feature augmentation techniques remain the most explored category for high-expertise domains, we see a change in trend for non-narrative domains. Loss augmentation and pretraining are more commonly explored categories. The difference in dominant model categories can be partly attributed to easy availability of large-scale unlabeled data and weak signals (e.g., likes, shares), particularly for social media. Such user-generated content (called ‘‘fortuitous data’’ by Plank [2016]) is leveraged well by pretraining or multi-task learning techniques, making them popular choices for non-narrative domains. In contrast, high-expertise domains (security and defense reports, finance, etc.) often lack fortuitous data, with methods developed for them focusing on learning shared feature spaces.

Ten studies in our meta-analysis sample evaluate on both domain types. Five of these studies (described in Table 4) operationalize two key ideas that seem to improve adaptation performance but have remained relatively under-explored in the context of recent methods like pretraining:

- **Incorporating Source-target Distance:** Several methods explicitly incorporate distance between source and target domain (e.g., Xing et al., 2018; Wang et al., 2018). Aside from allowing flexible adaptation based on the specific domain pairs being considered, adding source-target distance provides two benefits. It offers an additional avenue to analyze generalizability by monitoring source-target distance during adaptation. It also allows performance to be estimated in advance using source-target distance, which can be helpful when choosing an adaptation technique for a new target domain. Kashyap et al. (2020) provide a comprehensive overview of source-target distance metrics and discuss their utility in analysis and performance prediction. Despite these benefits, very little work has tried to incorporate source-target distance into newer adaptation method categories such as pretraining.

Study	Method	Performance
(Arnold et al., 2008)	Manually constructed feature hierarchy across domains, allowing back off to more general features (FA)	Positive transfer from 5 corpora (biomedical, news, email) to email
(McClosky et al., 2010)	Mixture of domain-specific models chosen via source-target similarity features (e.g., cosine similarity) (EN)	Positive transfer to biomedical, literature and conversation domains
(Yang and Eisenstein, 2015)	Dense embeddings induced from template features and manually defined domain attribute embeddings (FA)	Positive transfer to 4/5 web domains and 10/11 literary periods
(Xing et al., 2018)	Multi-task learning method with source-target distance minimization as additional loss term (LA)	Positive transfer on 4/6 intra-medical settings (EHRs, forums) and 5/9 narrative to medical settings
(Wang et al., 2018)	Source-target distance minimized using two loss penalties (LA)	Positive transfer to medical and Twitter data

Table 4: Model and performance details for studies testing on high-expertise and non-narrative domains. Fine method categories used in these studies include feature augmentation (FA), loss augmentation (LA), ensembling (EN), pretraining (PT), parameter initialization (PI), and pseudo-labeling (PL).

- **Incorporating Nuanced Domain Variation:** Despite NLP treating domain variation as a dichotomy (source vs. target), domains vary along a multitude of dimensions (topic, genre, medium of communication, etc.) (Plank, 2016). Some methods acknowledge this nuance and treat domain variation as multi-dimensional, either in a discrete feature space (Arnold et al., 2008) or in a continuous embedding space (Yang and Eisenstein, 2015). This allows knowledge sharing across dimensions common to both source and target, improving transfer. This idea has also remained under-explored, though recent work such as the development of domain expert mixture (DEMix) layers (Gururangan et al., 2021) has attempted to incorporate nuanced domain variation into pretraining.

Open Issues: Interestingly many studies from our sample do not analyze failures, namely, source-target pairs on which adaptation methods do not improve performance. For some studies in Table 4, adaptation methods do not improve performance on all source-target pairs. But failures are not investigated, presenting the question: *Do we know blind spots for current adaptation methods?* Answering this is essential to develop a complete picture of the generalization capabilities of adaptation methods. Studies that present negative transfer results (e.g., Plank et al., 2014) are rare, but should be encouraged to develop a sound understanding of adaptation techniques. Analyses should also study ties between datasets used and methods applied, highlighting dimensions of variation between source-target domains and how adaptation methods bridge them (Kashyap et al., 2020; Naik et al., 2021). Such analyses can uncover important lessons about generalizability of adaptation methods and the kinds of source-target settings they can be expected to improve performance on.

Identifying Under-explored and Promising Methods: Annotating long tail macro-level dimensions and adaptation method categories studied by all works included in our representative sample has the additional benefit of providing a framework to identify both the most under-explored, as well as most promising methods, under various settings. Tables 5 and 6 provide evidence gap maps presenting the number of works from our sample that study the utility of various method categories on different tasks and domains respectively.[9] The first thing we note is that both maps are highly sparse, indicating that there is little to no evidence for several combinations, many of which are worth exploring. In particular, given recent state-of-the-art advances, the following settings seem ripe for exploration:

- **Parameter Addition and Freezing:** Though there are only four studies in our sample (providing positive evidence) that study parameter addition and freezing methods, we believe that given the advent of large-scale language models, these categories merit further exploration for popular task categories (TC, POS, NER, NLI, SP). Both methods attempt to improve generalization by reducing overfitting which is likely to be more prevalent with large language models, and are additionally *efficient* methods that do not require a large number of extra parameters.

[9]We do not include languages because our meta-analysis does not solely focus on multilingual and cross-lingual work.

M \ T	TC	POS	NER	NLI	SP	WSD	TRN	TRG	RC	MF	LI	SLU
Feature Augmentation	25	6	13	3	4	2	1	2	1		1	1
Feature Generalization	2	1	1	1	1							
Loss Augmentation	21	5	7	4	3		1		2		1	
Parameter Initialization	7	1	4	2		2	1			1		
Parameter Addition			1									
Parameter Freezing	1	1	1									
Ensemble	2		1	1	2							
Instance Weighting	9	3	1	1		1						
Data Selection	3		1	1	1							
Pretraining	13		2	9	3			1			1	
Pseudo-Labeling	9	3	3	1	4				1			
Noising/Denoising	2	1	1									
Active Learning	4					1						
Instance Learning	2											

Table 5: Evidence gap map showing which method categories have not been explored sufficiently for various task categories. Please refer to Tables 1 and 2 for task abbreviations.

M \ D	HE	NN
Feature Augmentation	12	8
Feature Generalization	1	
Loss Augmentation	9	11
Parameter Initialization	1	4
Parameter Addition		1
Parameter Freezing	1	1
Ensemble	2	1
Instance Weighting	2	2
Data Selection	1	
Pretraining	5	6
Pseudo-Labeling	4	3
Noising/Denoising	1	
Active Learning		
Instance Learning		

Table 6: Evidence gap map showing which method categories have not been explored sufficiently for various long tail domain categories. Note that HE and NN refer to high-expertise and non-narrative domains. Please refer to Table 2 for model abbreviations.

- **Active Learning:** Studies included in our sample provide positive evidence for the use of active learning in an adaptation setting, but they have mainly evaluated on text classification (primarily sentiment analysis). We hypothesize that active learning during adaptation might also prove to be beneficial for task categories POS, NER, and SP, which require more complex, linguistically informed annotation.
- **Data Selection:** Despite being similar in nature to instance weighting methods for which several studies provide positive evidence, data selection methods seem to have been under-explored. We believe that these methods might be useful for POS, NER, and SP tasks for which large-scale fortuitous data is not as easily available, and adaptation must also take into account shifts in output structure.

Despite the scarcity of both maps, there are certain method-task and method-domain combinations for which our meta-analysis sample includes a reasonable number of studies ($\geq$10%). For these combinations, we provide a quick performance summary below:

- **Feature Augmentation:** On text classification, 12 of 25 studies use FA methods as baselines. Of the remaining 13 studies, 6 provide strong positive evidence—that is, the FA method outperforms all methods across all settings/domains tested. The remaining 7 provide mixed results—that is, there are certain domains on which this method category doesn't work best. On semantic sequence labeling tasks like NER, 4 of 13 studies use FA methods as baselines, 5 show strong positive results, and 4 show mixed results. Finally, on high-expertise domains, 1 study uses FA methods as baselines, 5 show strong positive results, and 6 show mixed results. These observations indicate that despite their popularity, feature augmentation methods are not as strong as other method categories.
- **Loss Augmentation:** For text classification, 8 of 21 studies use LA methods as

baselines. Of the remaining 13 studies, 11 provide strong positive evidence, while only 2 provide mixed results. On non-narrative domains, 9 of 11 studies provide strong positive evidence, while 2 provide mixed results. Based on their performance, loss augmentation methods seem to be extremely promising, especially for text classification and non-narrative domains.

- **Pretraining:** For text classification, 4 of 13 studies use pretraining as a baseline. Of the remaining 9 studies, 8 provide strong positive evidence and only one provides mixed results. Despite their relatively recent emergence, pretraining methods also seem to be extremely promising based on performance.

5 Which Methodological Gaps Have Greatest Negative Impact On Long Tail Performance?

The final goal of our meta-analysis is to identify methodological gaps in developing adaptation methods for long tail domains, which provide avenues for future research. Our observations highlight three areas: (i) combining adaptation methods, (ii) incorporating external knowledge, and (iii) application to data-scarce settings.

5.1 Combining Adaptation Methods

The potential of combining multiple adaptation methods has not been systematically and extensively studied. Combining methods may be useful in two scenarios. The first one is when source and target domains differ along multiple dimensions (topic, language, etc.) and different methods are known to work well for each. The second one is when methods focus on resolving issues in specific portions of the model such as feature space misalignment, task level differences, and so forth. Combining model-centric adaptation methods[10], that tackle each issue separately may improve performance over individual approaches. Despite its utility, method combination has only been systematically explored by one meta-study from 2010. On the other hand, 23 studies apply a particular combination of methods to their tasks/domains, but do not analyze when these combinations do/do not work. We summarize both sources of evidence and highlight open questions.

Method Combination Meta-study: Chang et al. (2010) observe that most adaptation methods either tackle shift in feature space ($P(X)$) or shift in how features are linked to labels ($P(Y|X)$). They call the former category ''unlabeled adaptation methods'' since feature space alignment can be done using unlabeled data alone. Methods from the latter category require some labeled target data and are called ''labeled adaptation methods.''[11] Through theoretical analysis, simulated experiments and experiments with real-world data on two tasks (named entity recognition and preposition sense disambiguation), they observe: (i) combination generally improves performance, (ii) combining best-performing individual methods may not provide best combination performance, and (iii) simpler labeled adaptation algorithms (e.g., jointly training on source and target data) interface better with strong unlabeled adaptation algorithms.

Applying Particular Combinations: Table 7 lists all studies that apply method combinations and fine-grained category labels from our hierarchy for the methods used. Combining methods from different coarse categories is the most popular strategy, utilized by 15 out of 23 studies. Five studies combine methods from the same coarse category, but different fine categories. They combine model-centric methods that edit different parts of the model (e.g., a feature-centric and a loss-centric method). The last 3 studies combine methods from the same fine category. Only 7 studies evaluate on at least one long tail domain.

Several studies observe performance improvements (Yu and Kübler, 2011; Mohit et al., 2012; Scheible and Schütze, 2013; Kim et al., 2017; Yang et al., 2017; Alam et al., 2018), mirroring the observation by Chang et al. (2010) that method combination helps. However, this is not consistent across all studies. For example, Jochim and Schütze (2014) find that combining marginalized stacked denoising autoencoders (mSDA) (Chen et al., 2012) and frustratingly easy domain adaptation (FEDA) (Daumé III, 2007) performs worse than individual methods in preliminary experiments on citation polarity classification. Both methods are feature-centric, though mSDA is a generalization method (FG) while FEDA is an augmentation method (FA). Additionally, mSDA is an unlabeled adaptation method while FEDA is

[10] As per our categorization presented in §2.3.

[11] These categories do not map cleanly to our hierarchy.

Study	Method	LT
Different Coarse Categories		
(Jeong et al., 2009)	IW+PL	✔
(Hangya et al., 2018)	PT+FA	✔
(Cer et al., 2018)	PT+LA	✔
(Dereli and Saraclar, 2019)	FA+PT	✔
(Ji et al., 2015)	FG+IW	
(Huang et al., 2019)	PI+PL	
(Li et al., 2012)	LA+PL+IW	
(Chan and Ng, 2007)	AL+PI+IW	
(Nguyen et al., 2014)	PL+EN	
(Yu and Kübler, 2011)	PL+IW	
(Scheible and Schütze, 2013)	FA+PL+DS	
(Tan and Cheng, 2009)	FA+IW	
(Mohit et al., 2012)	LA+PL	
(Rai et al., 2010)	AL+LA	
(Wu et al., 2017)	AL+LA	
Same Coarse Categories		
(Lin and Lu, 2018)	PA+FA	✔
(Zhang et al., 2017)	FA+LA	✔
(Yan et al., 2020)	FA+LA	
(Yang et al., 2017)	LA+PL+FA	
(Gong et al., 2016)	LA+PI	
Same Fine Categories		
(Alam et al., 2018)	LA+LA	✔
(Lee et al., 2020)	PL+PL	
(Kim et al., 2017)	LA+LA	

Table 7: Category combinations explored by studies that combine multiple methods. LT indicates whether long tail domains were evaluated on. Method categories explored include feature augmentation (FA), feature generalization (FG), loss augmentation (LA), parameter initialization (PI), ensembling (EN), pseudo-labeling (PL), pretraining (PT), active learning (AL), instance weighting (IW), and data selection (DS).

a labeled adaptation method. Owing to negative results, Jochim and Schütze (2014) do not experiment further to find a combination that might have worked. Wright and Augenstein (2020) show that combining adversarial domain adaptation (ADA) (Ganin and Lempitsky, 2015) with pretraining does not improve performance, but combining mixture of experts (MoE) with pretraining does. This indicates that methods from the same coarse category (model-centric) may react differently in combination settings. Similarly, studies achieving positive results do not analyze which properties of chosen methods allow them to combine well, and whether this success extends to other methods with similar properties.

Open Questions: To understand method combination, we must examine the following questions:

- Is it possible to draw general conclusions about the potential of combining methods from various coarse or fine categories?
- Which properties of adaptation methods are indicative of their ability to interface well with other methods?
- Do task and/or domain of interest influence the abilities of methods to combine successfully?

5.2 Incorporating External Knowledge

Most methods leverage labeled/unlabeled text to learn generalizable representations. However, knowledge from sources beyond text such as ontologies, human understanding of domain/task variation, and so on, can also improve adaptation performance. This is especially true for domains with expert-curated ontologies (e.g., UMLS for biomedical/clinical text [Bodenreider, 2004]). From our study sample, we observe some exploration of the following knowledge sources:

Ontological Knowledge: Romanov and Shivade (2018) use UMLS for clinical natural language inference via two techniques: (i) retrofitting word vectors as per UMLS (Faruqui et al., 2015), and (ii) using UMLS concept distance-based attention. Retrofitting hurts performance, whereas concept distance provides modest improvements.

Domain Variation: Arnold et al. (2008) and Yang and Eisenstein (2015) incorporate human understanding of domain variation in discrete and continuous feature spaces, respectively, with some success (Table 4). Structural correspondence learning (Blitzer et al., 2006) relies on manually defined pivot features common to source and target domains, and shows performance improvements.

Task Variation: Zarrella and Marsh (2016) incorporate human understanding of knowledge required for stance detection to define an auxiliary hashtag prediction task, which improves target task performance.

Manual Adaptation: Chiticariu et al. (2010) manually customize rule-based NER models, matching scores achieved by supervised models.

Another source that is not explored by studies in our sample, but has gained popularity, is providing task descriptions for sample-efficient transfer learning (Schick and Schütze, 2021). Despite initial explorations, the potential of external knowledge sources is largely under-explored.

Open Questions: Given varying availability of knowledge sources across tasks/domains, comparing their performance across domains may be impractical. But studies experimenting with a specific source can still probe the following questions:

- Can reliance on labeled/unlabeled data be reduced while maintaining the same performance?
- Does incorporating the knowledge source improve interpretability of the adaptation method?
- Can we preemptively identify a subset of samples which may benefit from the knowledge?

5.3 Application to Data-scarce Settings

§3 shows that most studies test methods in a supervised setting in which labeled and/or unlabeled data is available from both source and target domains. But availability of labeled or unlabeled data is often limited for long tail domains and languages. Hence, methods should also be developed for and applied to settings that reflect real-world criteria like data availability. Data-scarce adaptation settings might be harder, but are extremely important since they closely resemble contexts in which transfer learning is likely to be used. In particular, more evaluation should be carried out in the following data-scarce settings:

Unsupervised Adaptation: No labeled target data is available. Methods can use unlabeled target data or obtain distantly supervised target data from auxiliary resources (gazetteers) and user-generated signals (likes, shares, etc.).

Multi-source Adaptation: Instead of a single large-scale source dataset, smaller datasets from several source domains are available.

Online Adaptation: Especially pertinent for productionizing models, in this setting, adaptation methods must learn to adapt to new domains on-the-fly. Often information about the target domain beyond the current sample may not be available.

Source-free Adaptation: A trained model must be adapted to a target domain without source domain data, either labeled or unlabeled. This setting is especially useful for domains that have strong data-sharing restrictions such as clinical data.

Some of these settings have attracted attention in recent years. Ramponi and Plank (2020) comprehensively survey neural methods for unsupervised adaptation. In their survey on low-resource NLP, Hedderich et al. (2020) cover transfer learning techniques that reduce need for supervised target data. Wang et al. (2021) list human-in-the-loop data augmentation and model update techniques that can be used for data-scarce adaptation. However, there is room to further study application of adaptation methods in data-scarce settings.

Open Questions: Broadly, two main questions in this area still remain unanswered:

- At different levels of data scarcity (no labeled target data, no unlabeled target data, etc.), which adaptation methods perform best?
- Can we correlate source-target domain distance and data-reliance of adaptation methods?

6 Case Study: Evaluating Adaptation Methods on Clinical Narratives

Finally, we attempt to demonstrate how our meta-analysis framework and observations can be used to systematically design case studies that can provide answers to the prevailing open questions laid out previously. As an example, we conduct a case study to evaluate the effectiveness of popularly used adaptation methods on high-expertise domains in an unsupervised adaptation setting, a burgeoning area of interest (Ramponi and Plank, 2020). Specifically, our study focuses on the question: Which method categories perform best for semantic sequence labeling tasks when transferring from news to clinical narratives, given an unsupervised setting (i.e., no labeled clinical

data available)? We focus on two semantic sequence labeling tasks: entity extraction and event extraction.

6.1 Datasets

We use the following entity extraction datasets:

- **CoNLL 2003** (Tjong Kim Sang and De Meulder, 2003): News stories annotated with four entity types: persons, organizations, locations, and miscellaneous names.
- **i2b2 2006** (Uzuner et al., 2007): Medical discharge summaries annotated with PHI (private health information) entities of eight types: patients, doctors, locations, hospitals, dates, IDs, phone numbers, and ages.
- **i2b2 2010** (Uzuner et al., 2011): Discharge summaries annotated with three entity types: medical problems, tests, and treatments.
- **i2b2 2014** (Stubbs and Uzuner, 2015): Longitudinal medical records annotated with PHI entities of eight broad types: name, profession, location, age, date, contact, IDs, and other.

All entities are annotated in IOB format. For event extraction, we use the following datasets:

- **TimeBank** (Pustejovsky et al., 2003): News articles annotated with events.
- **i2b2 2012** (Sun et al., 2013): Discharge summaries annotated with events.
- **MTSamples** (Naik et al., 2021): Medical records annotated with events (test-only).

CoNLL 2003 and TimeBank are the source datasets for all entity and event extraction experiments, respectively, while the remaining are target datasets. We focus on English narratives only. Among the NER datasets, the label sets for i2b22006 and i2b22014 can be mapped to the label set for CoNLL2003, however the label set for i2b22010 is quite distinct and cannot be mapped. Therefore, we evaluate NER in two settings: *coarse* and *fine*. In the coarse setting, the model only detects entities, but does not predict entity type, whereas in the fine setting, the model detects entities and predicts types.

6.2 Adaptation Methods

The baseline model for both tasks is a BERT-based sequence labeling model that computes token-level representations using BERT, followed by a linear layer that predicts entity/event labels. We compare the performance of adaptation methods from the top five fine categories most frequently applied to high-expertise domains as per our analysis (Figure 10a), on top of this BERT baseline. Since feature augmentation (FA) methods require some target labeled data to train target-specific weights and our focus is on an unsupervised setting, our study tests the remaining four categories:

- **PL:** From pseudo-labeling, we test the self-training method. Self-training first trains a sequence labeling model on the source dataset (news), then uses this model to generate labels for unlabeled target data (clinical narratives). High-confidence predictions from the ‘‘pseudo-labeled’’ clinical data are combined with source data to train a new sequence labeling model. This process can be repeated iteratively.
- **LA:** From loss augmentation, we test adversarial domain adaptation (Ganin and Lempitsky, 2015). This method learns domain-invariant representations by adding an adversary that predicts an example’s domain and subtracting the loss from this adversary from the overall model loss. This setup is trained in a two-stage alternating optimization process (complete details in Ganin and Lempitsky [2015]).
- **PT:** From pretraining, we test domain-adaptive pretraining as described by Gururangan et al. (2020). This method tries to improve target domain performance of BERT-based models by continual masked language modeling pretraining on unlabeled text from the target domain.
- **IW:** From instance weighting, we test classifier-based instance weighting. This method trains a classifier on the task of predicting an example’s domain, then runs the classifier on all source domain examples and uses target domain probabilities as weights. Source examples that ‘‘look’’ more like the target domain get higher weights, improving performance on the target domain. We perform interleaved training, recomputing source weights after each model training pass.

	i2b22006			i2b22010			i2b22014		
AM	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
ZS	18.7	21.8	20.1	35.2	10.1	15.7	21.2	**32.8**	25.7
LA	16.1	21.2	18.3	36.6	**15.4**	**21.7**	27.5	28.6	28.0
PL	**23.2**	22.0	**22.6**	23.3	5.0	8.3	**47.4**	23.6	**31.5**
PT	19.5	**22.1**	20.7	**38.1**	12.8	19.1	27.3	27.4	27.3
IW	21.0	19.5	20.2	34.3	12.1	17.9	21.0	29.2	24.4

Table 8: Performance of all adaptation methods on NER in the coarse setting. Recall that the fine adaptation method categories we evaluate are loss augmentation (LA), pseudo-labeling (PL), pretraining (PT), and instance weighting (IW).

	i2b22006			i2b22014		
AM	**P**	**R**	**F1**	**P**	**R**	**F1**
ZS	12.6	14.1	13.3	24.0	**28.3**	25.9
LA	16.1	**15.8**	**16.0**	22.8	25.7	24.2
PL	**17.5**	11.4	13.8	**39.5**	21.4	**27.7**
PT	10.0	12.3	11.1	17.1	22.3	19.4
IW	14.4	14.1	14.2	21.8	25.6	23.6

Table 9: Performance of all adaptation methods on NER in the fine setting. Recall that the fine adaptation method categories we evaluate are loss augmentation (LA), pseudo-labeling (PL), pretraining (PT), and instance weighting (IW).

	i2b22012			MTSamples		
AM	**P**	**R**	**F1**	**P**	**R**	**F1**
ZS	48.8	15.3	23.3	91.4	48.0	63.0
LA	**51.7**	**19.0**	**27.8**	88.1	**58.5**	**70.3**
PL	44.1	11.4	18.2	**91.8**	39.3	55.1
PT	41.5	10.4	16.6	90.2	46.3	61.2
IW	50.5	18.1	26.6	90.6	48.4	63.1

Table 10: Performance of all adaptation methods on event extraction. Recall that the fine adaptation method categories we evaluate are loss augmentation (LA), pseudo-labeling (PL), pretraining (PT), and instance weighting (IW).

6.3 Results

Tables 8 and 9 show the results of all adaptation methods on coarse and fine NER, while Table 10 shows results on event extraction. ZS indicates baseline model scores in a zero-shot setting, that is, training on source and testing on target with no adaptation. From these tables, we can see that the best-performing method categories are loss augmentation and pseudo-labeling across different settings. Loss augmentation methods work best for event extraction. For coarse NER, pseudo-labeling methods work better on target datasets whose labels can be mapped to the source (i.e., *closer* transfer tasks). For i2b22010, which is more distant transfer, loss augmentation works best. The effectiveness of pseudo-labeling is interesting because it often suffers from the pitfall of propagating errors made by the source-trained model, which may in part explain its poor performance on i2b22010. Early work on applying these methods to parsing showed negative results or minor improvements (Charniak, 1997; Steedman et al., 2003), but these methods have shown more promise in recent years with advances in embedding representations. Finally, for fine NER, loss augmentation and pseudo-labeling do better on i2b22006 and i2b22014, respectively. Pretraining is not the best-performing method in any setting, which may be a side effect of continual pretraining leading to some forgetting, negatively impacting an unsupervised setting. This highlights the need to systematically compare adaptation methods under data-scarce settings because the ranking of methods can change based on the availability and quality of domain-specific data.

7 Conclusion

This work presents a qualitative meta-analysis of 100 representative papers on domain adaptation and transfer learning in NLU, with the aim of understanding performance of adaptation methods on the long tail. Through this analysis, we assess current trends and highlight methodological gaps that we consider to be major avenues for future research in transfer learning for the long tail. We observe that long tail coverage in current research is far from comprehensive, and identify two properties of adaptation methods that may improve long tail performance, but have been under-explored: (i) incorporating source-target distance, and (ii) incorporating nuanced domain variation. Additionally, we identify three major gaps that must be addressed to improve long tail performance: (i) combining adaptation methods, (ii) incorporating external knowledge, and (iii) application to data-scarce adaptation settings. Finally, we demonstrate the utility of our meta-analysis framework and share

observations in guiding the design of systematic meta-experiments to address prevailing open questions by conducting a systematic evaluation of popular adaptation methods for high-expertise domains in a data-scarce setting. This case study reveals interesting insights about the adaptation methods evaluated and shows that significant progress can be made towards developing a better understanding of adaptation for the long tail by conducting such experiments.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health, Clinical Research Center, and through an Inter-Agency Agreement with the US Social Security Administration. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NIH, or the US Government. The authors would like to thank Emma Strubell, Matt Gormley, Luke Zettlemoyer, Abhilasha Ravichander and Khyathi Chandu for their feedback on early drafts of this work. The authors are also extremely grateful for the valuable reviews provided by the anonymous reviewers and the action editor, Benjamin van Durme, which significantly improved our final version.

References

Mohammad Al Boni, Keira Zhou, Hongning Wang, and Matthew S. Gerber. 2015. Model adaptation for personalized opinion analysis. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 769–774, Beijing, China. Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-2126

Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain adaptation with adversarial training and graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1077–1087, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1099

Andrew Arnold, Ramesh Nallapati, and William W. Cohen. 2008. Exploiting feature hierarchy for transfer learning in named entity recognition. In *Proceedings of ACL-08: HLT*, pages 245–253, Columbus, Ohio. Association for Computational Linguistics.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26. https://doi.org/10.33011/lilt.v6i.1239

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. Association for Computational Linguistics.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics. https://doi.org/10.3115/1610075.1610094

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of ‘‘bias’’ in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Olivier Bodenreider. 2004. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(1):D267–D270. https://doi.org/10.1093/nar/gkh061

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015*

Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1075

Chloé Braud and Pascal Denis. 2014. Combining natural and artificial examples to improve implicit discourse relation identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1694–1705, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Yee Seng Chan and Hwee Tou Ng. 2006. Estimating class priors in domain adaptation for word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 89–96, Sydney, Australia. Association for Computational Linguistics. https://doi.org/10.3115/1220175.1220187

Yee Seng Chan and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics.

Ming-Wei Chang, Michael Connor, and Dan Roth. 2010. The necessity of combining adaptation methods. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 767–777, Cambridge, MA. Association for Computational Linguistics.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. *AAAI/IAAI*, 2005(598–603):18.

Minmin Chen, Zhixiang Eddie Xu, Kilian Q. Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *ICML*.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.634

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1012, Cambridge, MA. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.207

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1070

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. Association for Computational Linguistics.

Neşat Dereli and Murat Saraclar. 2019. Convolutional neural networks for financial text

regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 331–337, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-2046

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1184

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 513–520.

Lin Gong, Mohammad Al Boni, and Hongning Wang. 2016. Modeling social norms evolution for personalized sentiment classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 855–865, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1081

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A. Smith, and Luke Zettlemoyer. 2021. Demix layers: Disentangling domains for modular language modeling. *arXiv preprint arXiv:2108.05036*.

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*, pages 8342–8360, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

Viktor Hangya, Fabienne Braune, Alexander Fraser, and Hinrich Schütze. 2018. Two methods for domain adaptation of bilingual tasks: Delightfully simple and broadly applicable. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 810–820, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1075

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*. https://doi.org/10.18653/v1/2021.naacl-main.201

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Yin Jou Huang, Jing Lu, Sadao Kurohashi, and Vincent Ng. 2019. Improving event coreference resolution by learning argument compatibility from unlabeled data. 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)*, pages 785–795, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1085

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1250–1259, Singapore. Association for Computational Linguistics. https://doi.org/10.3115/1699648.1699671

Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. 2015. Closing the gap: Domain adaptation from explicit to implicit discourse relations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2219–2224, Lisbon, Portugal. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Charles Jochim and Hinrich Schütze. 2014. Improving citation polarity classification with product reviews. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 42–48, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2008

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2020. Domain divergences: A survey and empirical analysis. *arXiv preprint arXiv:2010.12198*. https://doi.org/10.18653/v1/2021.naacl-main.147

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.329

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838, Copenhagen, Denmark. Association for Computational Linguistics.

Young-Suk Lee, Ramón Fernandez Astudillo, Tahira Naseem, Revanth Gangi Reddy, Radu Florian, and Salim Roukos. 2020. Pushing the limits of AMR parsing with self-learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3208–3214, Online. Association for Computational Linguistics.

Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 410–419, Jeju Island, Korea. Association for Computational Linguistics.

Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do we really need to access the source data? Source hypothesis transfer for unsupervised

domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2012–2022, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1518–1533, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.139

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921, Denver, Colorado. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330. https://doi.org/10.21236/ADA273556

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA. Association for Computational Linguistics. https://doi.org/10.3115/1220835.1220855

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, California. Association for Computational Linguistics.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 2012. Recall-oriented learning of named entities in Arabic Wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173, Avignon, France. Association for Computational Linguistics.

Aakanksha Naik, Jill Fain Lehman, and Carolyn Rose. 2021. Adapting event extractors to medical data: Bridging the covariate shift. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2963–2975, Online. Association for Computational Linguistics.

Denis Newman-Griffis, Jill Fain Lehman, Carolyn Rosé, and Harry Hochheiser. 2021. Translational NLP: A new paradigm and general principles for natural language processing research. *arXiv preprint arXiv:2104.07874*. https://doi.org/10.18653/v1/2021.naacl-main.325

Minh Luan Nguyen, Ivor W. Tsang, Kian Ming A. Chai, and Hai Leong Chieu. 2014. Robust domain adaptation for relation extraction via clustering consistency. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 807–817, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1076

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5006

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan. The COLING 2016 Organizing Committee.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19-21, 2016*, volume 16 of *Bochumer Linguistische Arbeitsberichte*.

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014. Importance weighting and unsupervised domain adaptation of POS taggers: A negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1104

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus Linguistics*, volume 2003, page 40. Lancaster, UK.

Piyush Rai, Avishek Saha, Hal Daumé, and Suresh Venkatasubramanian. 2010. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, Los Angeles, California. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1264

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.603

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1187

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-5004

Christian Scheible and Hinrich Schütze. 2013. Sentiment relevance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 954–963, Sofia, Bulgaria. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.20

Burr Settles. 2009. Active learning literature survey. http://digital.library.wisc.edu/1793/60660

Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Bootstrapping statistical parsers from small datasets. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics. https://doi.org/10.3115/1067807.1067851

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. https://doi.org/10.1016/j.jbi.2015.07.020

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. https://doi.org/10.1136/amiajnl-2013-001628

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Songbo Tan and Xueqi Cheng. 2009. Improving SCL model for sentiment-transfer learning. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 181–184, Boulder, Colorado. 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*, pages 142–147. https://doi.org/10.3115/1119176.1119195

Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017. LIMSI-COT at SemEval-2017 task 12: Neural architecture for temporal information extraction from clinical narratives. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 597–602, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2098

Shulamit Umansky-Pesin, Roi Reichart, and Ari Rappoport. 2010. A multi-domain web-based algorithm for POS tagging of unknown words. In *COLING 2010: Posters*, pages 1274–1282, Beijing, China. COLING 2010 Organizing Committee.

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. https://doi.org/10.1197/jamia.M2444

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. https://doi.org/10.1136/amiajnl-2011-000203

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*. https://doi.org/10.18653/v1/W18-5446

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1–15, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1001

Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2020. Transformer based multi-source domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7963–7974, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.639

Fangzhao Wu, Yongfeng Huang, and Jun Yan. 2017. Active sentiment domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1711, Vancouver, Canada. Association for Computational Linguistics.

Junjie Xing, Kenny Zhu, and Shaodian Zhang. 2018. Adaptive multi-task transfer learning for Chinese word segmentation in medical text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3619–3630, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ming Yan, Hao Zhang, Di Jin, and Joey Tianyi Zhou. 2020. Multi-source meta transfer for low resource multiple-choice question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7331–7341, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.654

Haitong Yang, Tao Zhuang, and Chengqing Zong. 2015. Domain adaptation for syntactic and semantic dependency parsing using deep belief networks. *Transactions of the Association for Computational Linguistics*, 3:271–282. https://doi.org/10.1162/tacl_a_00138

Yi Yang and Jacob Eisenstein. 2015. Unsupervised multi-domain adaptation with feature embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 672–682, Denver, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1069

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1040–1050, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1096

Wenpeng Yin, Tobias Schnabel, and Hinrich Schütze. 2015. Online updating of word representations for part-of-speech tagging. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1329–1334, Lisbon, Portugal. Association for Computational Linguistics.

Ning Yu and Sandra Kübler. 2011. Filling the gap: Semi-supervised learning for opinion detection across domains. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 200–209, Portland, Oregon, USA. Association for Computational Linguistics.

Guido Zarrella and Amy Marsh. 2016. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1074

Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2017. Aspect-augmented adversarial networks for domain adaptation. *Transactions of the Association for Computational Linguistics*, 5:515–528. https://doi.org/10.1162/tacl_a_00077