Domain adaptation in practice: Lessons from a real-world information extraction pipeline

Timothy Miller
Computational Health Informatics Program
Boston Children's Hospital
Department of Pediatrics, Harvard Medical School
timothy.miller@childrens.harvard.edu

Egoitz Laparra and **Steven Bethard**
School of Information
University of Arizona
{laparra,bethard}@arizona.edu

Abstract

Advances in transfer learning and domain adaptation have raised hopes that once-challenging NLP tasks are ready to be put to use for sophisticated information extraction needs. In this work, we describe an effort to do just that – combining state-of-the-art neural methods for negation detection, document time relation extraction, and aspectual link prediction, with the eventual goal of extracting drug timelines from electronic health record text. We train on the THYME colon cancer corpus and test on both the THYME brain cancer corpus and an internal corpus, and show that performance of the combined systems is unacceptable despite good performance of individual systems. Although domain adaptation shows improvements on each individual system, the model selection problem is a barrier to improving overall pipeline performance.

1 Introduction

Advances in machine learning methods and the release of annotated datasets of clinical texts (Uzuner et al., 2011; Styler IV et al., 2014) in the past decade has led to an increase of available clinical NLP systems for interesting tasks. Recent advances in pre-trained models (Devlin et al., 2019; Liu et al., 2019) have made ever more accurate clinical NLP systems possible. Unsupervised domain adaptation algorithms (e.g., Ziser and Reichart (2019)) have made it possible to reduce performance degradation when applying trained models to new domains. The great promise of these developments is that these methods can be combined into pipelines that allow for sophisticated information extraction capabilities for downstream clinical use cases. Rather than building one-off datasets for each complex downstream task that arises, standard NLP components could potentially be used as "Lego"-style building blocks that allow for flexibly approaching new tasks as they arise.

However, the existence of the building blocks alone does not solve this problem. Combining individual components into NLP pipelines can lead to cascading errors (Finkel et al., 2006). The true error rate for structured extraction tasks is potentially as high as the sum of the component tasks' errors. For example, if the goal is to extract normalized concepts with assertion status, the concept error can come from normalization error, negation detection error, uncertainty detection error, etc, and the errors may not be correlated. These problems are exacerbated in the common case where individual components are trained on data from different domains, and tested on data from yet another domain.

In this work, we quantitatively examine the issues described above in the context of extracting drug temporality signatures, with the goal of understanding drug start and stop events. We approach this task with the combination of three sub-tasks: 1) the temporal relation of these mentions to the document creation time (DocTimeRel), 2) negation status of the mention, and 3) aspectual link relations of the mention (e.g., is it being described as starting or stopping). Figure 1 shows an example sentence with a drug mention, that demonstrates how the three tasks work together to establish the status of that drug in that patient. Successfully solving this task is beneficial for understanding patient treatment course, and enabling more causal understanding in important tasks such as adverse drug event detection or relating medication courses to outcomes.

We first set state-of-the-art benchmarks for three tasks on the THYME corpus by fine-tuning large pre-trained transformer models (Devlin et al., 2019; Liu et al., 2019). We then examine how the performance of individual systems degrades when moving from the training data to our target data (a pediatric cardiology cohort), and how the overall system performs when combining multiple systems

Proceedings of the Second Workshop on Domain Adaptation for NLP, pages 105–110
April 20, 2021. ©2021 Association for Computational Linguistics

Additionally , as the patient has preserved ejection fraction , no prior history of embolic phenomena , and no significant valvular disease , it would be acceptable for him to remain off ⟨e⟩ ***Coumadin*** ⟨/e⟩ *for the interim .*

Figure 1: An example sentence with highlighted drug name *Coumadin* to be classified for all three tasks. The gold standard has this drug mention classified as *negated*, with DocTimeRel=OVERLAP, and ALINK=CONTINUES. These three facts can be used to understand that the patient is not on the drug now or going forward, and was likely not on the drug prior to the note as well.

with imperfect performance. Despite strong individual results, we find that performance suffers immensely due to both out-of-domain performance losses and the basic combinatorial math of integrating outputs from multiple systems. This is the case even though we use a metric, accuracy, that is forgiving to the worst-performing individual model.

2 Background

It is both formally and empirically understood that classifiers can suffer performance loss when the test data is drawn from a different distribution than the training data (sometimes called *domain shift*). This presents a difficult challenge in clinical NLP because data-sharing limitations make it difficult to create large and diverse training corpora. As a result, domain adaptation approaches have been applied to multiple tasks in clinical NLP (Miller et al., 2017; Liu et al., 2018; Hur et al., 2020).

Recent work in the general domain has made use of transfer learning, which can attack the problem of domain shift, but by a different mechanism than domain adaptation; by training on massive corpora, large pre-trained models both learn general features, and are able to learn from smaller new datasets without overfitting. The most prominent of these models are based on the transformer architecture (Vaswani et al., 2017).

BERT (Devlin et al., 2019) uses a transformer encoder, and has shown that pre-training with massive amounts of text on a language modeling task, then fine-tuning on a supervised task of interest, achieves large performance gains in multiple NLP tasks.[1] During fine-tuning for sentence classification tasks, a classification head with randomly initialized weights is attached to a special sentence-initial token. Fine-tuning then proceeds in a standard supervised learning paradigm, with the goal of learning the weights of the classification head, but where the weights of all of the transformer encoder layers can also be updated. We use RoBERTa-base, a 12-layer transformer encoder that provides excellent performance but manageable memory utilization for our hardware (Liu et al., 2019).

The bigger vision of our current work is extracting temporally-aware medication mentions from electronic health records data. This would enable important downstream tasks including automatically extracting drug timelines to correlate with treatments, or extracting better causal information about drugs and potential adverse events. Some other recent work has also examined this topic (Ramirez et al., 2019), but focused on a single drug class (proton pump inhibitors), was limited to the problem list section, and made the assumption that missing drug implied drug stoppage.

3 Methods

We began this work by developing several NLP components necessary to extract drug temporality signatures, including negation detection, relation to document creation time (DocTimeRel), and aspectual link extraction (ALINK), all detailed below. Detecting negation helps us avoid false positives from mentions corresponding to, for example, decisions to not use a drug. DocTimeRel helps us distinguish mentions of drugs that are current from those that predate the current time period, or are being speculated about for future use. ALINK can model drug start, stop, and continuation events, which can help to distinguish whether a missing mention in the middle of a record corresponds to a stop and restart, or an incidentally omitted mention. Figure 1 shows an example instance of a drug mention to be classified for all three tasks.

The THYME dataset (Styler IV et al., 2014), released as part of Clinical TempEval (Bethard et al., 2017), contains all three of these annotation types, on 1200 notes of patients with colon and brain cancer. We train all models on the colon cancer section (details on data are in Section 4). While our bigger

[1]The RoBERTa system that followed (Liu et al., 2019) found further gains by pre-training on even larger datasets and for more iterations.

project is specific to drug mentions, the problem is not limited to drug mentions, so we train and evaluate on all annotated events in the THYME corpus. We also assume that events are given, to allow a straightforward metric of how many events we "get right" when combining all property predictions. In the real world, events will have to be automatically detected, so our metric will be an upper bound on how often the combined models get everything correct.

Negation detection

This is the task of finding whether a given event is being negated (e.g., *statins* is negated in *not currently on statins*). We model this as a span-in-context classification – given a sentence in a document with a marked event span, classify that span as being negated or not negated. We experiment with two different machine learning models. The first is a classical feature-based support vector machine that is the default model of Apache cTAKES (Savova et al., 2010). Features include bag of words and part of speech tags in and around the event, negation cue words from lists and their relation to the event, and dependency parse features that relate negation cue words to events. Details of this system were presented by Wu et al. (2014). For comparison we train a RoBERTa-based system, where the input representation is the sentence with special tokens indicating the event to be classified. We put a binary sigmoid layer as the output, with the "[CLS]" token representation from the final layer as the classifier input, and fine-tune the entire model. Hyperparameters such as learning rate and number of training epochs are optimized on the THYME colon development set. Our implementation uses the Huggingface Transformers library (Wolf et al., 2019).

DocTimeRel

DocTimeRel classification is the task of relating an event to the document creation time. The categories are BEFORE, OVERLAP, AFTER, and BEFORE/OVERLAP. As above, we model this as a span-in-context classification, and we again compare a feature-based approach with a RoBERTa-based approach.

The feature-based approach again uses the default cTAKES SVM-based implementation (Lin et al., 2016), with features based on bags of words in and around the event, and verb tense information for verbs on either side of the event. We train a separate RoBERTa-based model with the same architecture as the negation model, with the only difference being that the output layer is a softmax over the four categories rather than a sigmoid.

Aspectual Link Extraction

Aspectual link extraction (ALINK) is the task of classifying whether an event mention is related to an aspectual temporal modifier, for example, *discontinued*. This is annotated as a relation between an event and a modifier, but we model it as an event property classification task since each event can only participate in one type of relation. The set of possible labels is INITIATES, CONTINUES, TERMINATES, and REINITIATES.

We are not aware of any existing open-source models for this task, so for our feature-based baseline we train a model with the same SVM classification approach and feature set as the DocTimeRel model in cTAKES. We did not perform extensive feature engineering for this task, so further gains in the SVM system are probably possible. For the RoBERTa-based model, we used the same architecture as both systems above, with a softmax over the 5 categories – the 4 ALINK categories above as well as NONE, indicating a drug mention does not participate in any ALINK relation. NONE is by far the most common category.

3.1 Domain Adaptation Methods

The tasks described above are trained on a single *source* dataset, and must be combined into a pipeline that will run on data from a different *target* distribution. To adapt to the target domain, we use *unsupervised* domain adaptation methods, where we have access to only unlabeled target examples.

Since large pre-trained transformer models have arrived, they have been shown to be quite robust to out-of-distribution examples (Hendrycks et al., 2020), including on clinical tasks (Lin et al., 2020), where it was shown that adding domain adaptation layers on top of BERT was no better than BERT itself for negation detection. One of the few effective methods for improving the out-of-distribution performance of pre-trained transformer models has been to continue to pre-train the language modeling objective on the target domain data, before any fine-tuning is done on the source data (Han and Eisenstein, 2019; Gururangan et al., 2020). In this work, we focus on that method, since this is currently the most promising direction for adapting large pre-trained transformers. Specifically, to use

this method, we run additional masked language model training steps on the target training data from the RoBERTa-base checkpoint, before fine-tuning on the labeled colon cancer data, and then testing on target test data. We tune the learning rate for the language model pre-training on target development set data, optimizing for perplexity.

4 Evaluation

For the three tasks of interest, we evaluate in-domain (THYME colon cancer corpus), as well as one closely related out-of-domain corpus (THYME brain cancer corpus). We also use a second out-of-domain corpus, an internal data set we annotated for all three tasks (pulmonary hypertension [PH] notes). This annotation was performed by an experienced annotator who has worked on clinical annotation projects in the past.

We measure performance on negation with F1-score, on DocTimeRel with accuracy (because the classes are relatively balanced), and on ALINK extraction with the average F1 score of all categories, macro-F1 (because the high frequency NONE label makes accuracy uninformative). In addition to system-level performance, we report an evaluation of mention-level accuracy: an event is counted as correct if all three systems made the correct prediction, and we report the percentage of events that were correct. This setting estimates how usable the entire pipeline is, given different system settings.

The "Colon" columns of Table 1 show results on the THYME colon cancer data (in-domain). RoBERTa performance is stronger than the SVM on all three tasks. Negation performance is particularly strong, though we are not aware of any reported results on this dataset to compare against. DocTimeRel performance is 3 points better than the best result of Clinical TempEval 2016 (Bethard et al., 2017). ALINK scores are lower than the other tasks, though again there are no published comparisons. It is likely this is a more difficult task, in particular because the RE-INITIATES category has relatively few examples and whose low performance skews the averaging of the macro-F1.

The "Brain" and "PH" columns of Table 1 show out-of-domain performance of the same systems on the THYME brain cancer and our internal pulmonary hypertension data, respectively. On THYME brain cancer data, RoBERTa again out-performs SVM substantially on all sub-tasks, but surprisingly the SVM performs better on PH data for negation and DocTimeRel. Adapting the RoBERTa model (RoBERTa+LM) by performing additional language modeling in the target domain before fine-tuning on colon cancer data leads to gains only on DocTimeRel for the PH data and on ALINK for both corpora. However, the improvement to DocTimeRel from adapting RoBERTa still leaves it worse off than the SVM.

Mention level accuracy ("All" column) is good for the in-domain data (THYME colon cancer), but drops off substantially even for the THYME brain cancer corpus from the same institution, created with the same guidelines and using the same annotators. The mention level accuracy for our internal PH data is unusable at an accuracy of 0.506 with RoBERTa+LM. This accuracy means that roughly one of every two drug mentions will have at least one of its attributes classified incorrectly.

5 Discussion and Conclusion

The results also show that combining NLP systems for new, complex, information needs is likely to run into issues even when individual systems perform well. In particular, our experiments raise questions about real-world use of domain adaptation. If we treated THYME colon and brain sets as representative in-domain and out-of-domain datasets we would select RoBERTa or RoBERTa+LM for everything. But an oracle optimizing PH performance would tell us to use the SVM for negation and DocTimeRel and RoBERTa+LM for ALINK. One of the difficulties in even *studying* domain adaptation is model selection – if labeled target data is not available, standard practices like tuning on held out data are impossible. But the reality our results suggest is that different algorithms work well on different tasks and datasets, and selecting the best model for each task is an unsolved and under-studied problem.

One direction of research that may address these concerns is on better modeling of domains themselves. The problem has been exacerbated with the move from feature-based classifiers to pre-trained black box models, as it is now even more difficult to understand the cause of errors in new domains without interpretable features. Domain adaptation should leverage "BERTology" and interpretability research to help understand how different aspects of domains contribute to performance differences. For example, in clinical notes, variation in institutions, specialties, note types, authors, etc., all

System	Negation (F1)			DocTimeRel (Acc)			ALINK (MacroF)			All (Acc)		
	Colon	Brain	PH	Colon	Brain	PH	Colon	Brain	PH	Colon	Brain	PH
SVM	0.924	0.705	0.625	0.842	0.703	0.694	0.502	0.338	0.345	–	–	–
RoBERTa	0.950	0.833	0.583	0.874	0.757	0.542	0.684	0.633	0.674	0.860	0.732	0.454
RoBERTa+LM	–	0.831	0.582	–	0.758	0.615	–	0.660	0.694	–	0.736	0.506

Table 1: Performance on both the individual sub-tasks (Negation, DocTimeRel, and ALINK) and the complete task (All) for systems trained on the THYME colon cancer training set and tested on the in-domain THYME colon test set, the out-of-domain THYME brain test set, and the out-of-domain pulmonary hypertension (PH) test set.

probably contribute differently to domain shift, and these sources of variation should be empirically explored. Future work will explore this direction to develop unsupervised model selection algorithms that better predict target domain performance.

Acknowledgments

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Numbers R01LM012918 and R01LM012973. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

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

Jenny Rose Finkel, Christopher D. Manning, and Andrew Y. Ng. 2006. Solving the problem of cascading errors: Approximate Bayesian inference for linguistic annotation pipelines. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 618–626, Sydney, Australia. Association for Computational Linguistics.

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

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Brian Hur, Timothy Baldwin, Karin Verspoor, Laura Hardefeldt, and James Gilkerson. 2020. Domain adaptation and instance selection for disease syndrome classification over veterinary clinical notes. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 156–166, Online. Association for Computational Linguistics.

Chen Lin, Steven Bethard, Dmitriy Dligach, Farig Sadeque, Guergana Savova, and Timothy A Miller. 2020. Does BERT need domain adaptation for clinical negation detection? *Journal of the American Medical Informatics Association*, 27(4):584–591.

Chen Lin, Dmitriy Dligach, Timothy A Miller, Steven Bethard, and Guergana K Savova. 2016. Multilayered temporal modeling for the clinical domain. *Journal of the American Medical Informatics Association*, 23(2):387–395.

Miaofeng Liu, Jialong Han, Haisong Zhang, and Yan Song. 2018. Domain adaptation for disease phrase matching with adversarial networks. In *Proceedings of the BioNLP 2018 workshop*, pages 137–141, Melbourne, Australia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Timothy Miller, Steven Bethard, Hadi Amiri, and Guergana Savova. 2017. Unsupervised domain adaptation for clinical negation detection. In *BioNLP 2017*, pages 165–170, Vancouver, Canada,. Association for Computational Linguistics.

Andrea H. Ramirez, Yaping Shi, Elliot Fielstein, Jonathan Schildcrout, Henry H. Ong, Josh C. Denny, and Josh F. Peterson. 2019. Extracting Drug Exposure Epochs and Drug Response Outcomes from Electronic Health Records. In *Proceedings of AMIA Annual Symposium (Podium Abstract)*, Washington, DC.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PloS one*, 9(11):e112774.

Yftah Ziser and Roi Reichart. 2019. Task refinement learning for improved accuracy and stability of unsupervised domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5895–5906, Florence, Italy. Association for Computational Linguistics.