On Domain Transfer for Intent Predicting in Text

Petar Stojanov *
Carnegie Mellon University
Pittsburgh, PA, USA
pstojano@cs.cmu.edu

Ahmed Awadallah, Saghar Hosseini, Paul N. Bennett
Microsoft Research AI
Redmond, WA, USA
{hassanam, sahoss, pauben}@microsoft.com

Abstract

In many domains, especially enterprise text analysis, there is an abundance of data which can be used for the development of new AI-powered intelligent experiences to improve people's productivity. However, there are strong-guarantees of privacy which prevent broad sampling and labeling of personal text data to learn or evaluate models of interest. Fortunately, in some cases like enterprise email, manual annotation is possible on certain public datasets. The hope is that models trained on these public datasets would perform well on the target private datasets of interest. In this paper, we study the challenges of transferring information from one email dataset to another, for predicting user intent. In particular, we present approaches to characterizing the transfer gap in text corpora from both an intrinsic and extrinsic point-of-view, and evaluate several methods for bridging this gap. We conclude by raising issues for further discussion in this arena.

1 Introduction

Using publicly available text data to train predictive models for use in privacy-aware enterprise settings is a very fruitful direction in the area of document understanding. However, when the labeled training dataset (source domain) is different from the unlabeled test (target domain), the two datasets likely follow different distributions. This application setting violates the *i.i.d.* assumption made by classic supervised learning methods and calls for domain adaptation techniques to properly account for this difference. State of the art domain adaptation techniques are generally developed and evaluated using a limited number of benchmark datasets and under constrained settings. The extent to which these methods are applicable for predictive settings over enterprise text data has neither been explored nor characterized in detail.

To explore the effectiveness of state of the art domain adaptation methodology in enterprise text data, we focus on communication intent prediction in enterprise email. In particular, we use two *public* enterprise email datasets (Avocado, an IT company, and Enron, an oil company) to systematically analyze the transfer problem in enterprise email. The two intent prediction tasks that we focus on in this study are Meeting Intent (the email expresses an intent to meet with the recipient) and Commitment Intent (the email expresses an action the sender intends to take in the future) – both of which are binary classification tasks.

1.1 Setting - Domain Adaptation

Denote the joint distributions of the source and the target domain as $P^S(X, Y)$ and $P^T(X, Y)$ respectively. In our setting, we *assume* $P^S(X, Y) \neq P^T(X, Y)$ and seek to establish measures that quantify these differences. This difference can be measured: (1) directly (*intrinsically*) in terms of observed distributional differences of words, n-grams, or more generally the representation conditioned on class; (2) or in terms of the downstream (extrinsic) impact on models learned to predict the class when used in a different domain. We refer to these measurable differences as the **transfer gap**. To understand the nature and extent to which these challenges persist in enterprise email, we first present intrinsic analyses of email text and features. We then proceed to extrinsic analyses to determine the extent to which state of the art methodology can account for the aforementioned

*Work done primarily while at Microsoft Research.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

challenges. Our contributions are: (1) we provide ways of measuring the transfer gap intrinsically over both classical bag-of-words and bag-of-n-grams representations as well as distributed representations; (2) provide evidence that these distributional differences lead to downstream measurable differences specifically in *enterprise* domains across different companies; (3) and, provide a first evaluation of proposed transfer methods from elsewhere in the literature to enterprise settings.

2 Intrinsic Analysis

We first begin by analyzing the intrinsic differences in distribution (i.e. the transfer gap) across the two datasets. The way text is processed into features has evolved over the years from BoW vectors, to dense feature vectors from Skip-Gram word embeddings [5], and more recently feature vectors that take into account contextual information such as ELMO and BERT [6, 3]. We seek to establish distributional differences exist across a variety of representation choices. To this end, we first analyze the difference by first comparing most frequent words when conditioned on positive class (i.e. comparing the head of $P(w|+)$) and then comparing the contexts in which these most frequent words in the positive class are used. Then we proceed to examine the transfer gap in terms of distributional measures of sentence-level encodings using a generic CNN encoder and state of the art word embeddings.

2.1 Overlap of Most Commonly Used Words

To analyze the most frequent words in positive intent, we compared the overlap between the top 30 n-grams (where n ranges from 1 to 3) in each dataset, when considering only sentences with positive labels, for each task. The results for 1-grams show that the overlap for the most frequent words in the two enterprises is 53.3% and 70.0% for meeting and commitment intents respectively. Similar counts hold for bi-grams and tri-grams too. From this analysis, it is apparent that for both tasks there is a significant number of words (nearly half in Meeting and close to a third in Commitment) that are frequently used to express positive intent, which do not overlap across domains. Furthermore, the most frequent words in positive-intent sentences overlap more in Commitment than in Meeting Intent. This means that in Meeting Intent the difference in distribution may come down to different word usage to express intent, whereas in Commitment if differences exist they may be more subtle and due to differences in context.

2.2 Difference in Contextual Use of Common Positive-Intent Words

To explore whether the top-30 positive-intent associated words were used in different contexts, we embedded each sentence using contextual word embeddings. For each domain and task, we considered each word in the top-30 most frequent list for positive intent (obtained as described above), and we retrieved a distribution of its contextual word embeddings from each of the sentences in which the word was present. To measure whether there is a difference in terms of the contexts in which a word was used in one domain vs. another, we compared the distribution of its contextual word embeddings in the two domains by maximum mean discrepancy (MMD) [4]: Avocado (source) vs. Enron (target). MMD is a kernel-based symmetric measure of distribution difference that can be applied to a variety of representation types including embeddings; we use a Gaussian kernel. As a baseline reference since any sample even from the same distribution may show differences, we also performed this comparison with two disjoint subsets of the Avocado dataset. The results are shown in Table 1, One can appreciate that there is a much larger relative difference in the distribution of contextual word embeddings when comparing across domains, as opposed to within domain. This indicates that the transfer gap includes not only different vocabularies, but also difference in context in the most relevant words associated with positive intent.

	MMD - Uni	MMD - Bigram	Rel. Uni	Rel Bigram
Avocado-Avocado (Meeting)	0.00001	0.001	1×	1×
Avocado-Enron (Meeting)	0.003	0.007	300×	7×
Avocado-Avocado (Commitment)	∼0.0	0.004	1×	1×
Avocado-Enron (Commitment)	0.008	0.01	∞	2.5×

Table 1: Comparisons of distributions of contextual word embeddings of the top-30 most common words in positive-labeled sentences, within domain and across domains. The MMD values displayed are averages over all top-30 terms (unigrams and bi-grams).

2.3 Analysis of Cross-Domain Differences of Encoder Representations

The previous intrinsic analyses consider lists of individual most frequent words. To capture sentence-level structure, sentences in modern methodologies are represented as sequences of dense vectors,

often obtained with a supervised deep encoder. In order to evaluate the extent to which there is a distribution difference in this dense sentence-level representation across domains, for each task and domain we trained a CNN encoder/classifier in that particular domain (for example Avocado). We then compared two distributions (in terms of MMD) of the CNN dense encodings of two sets of sentences from: (1) the domain in which the encodings were trained; (2) across domains. The results can be found in Table 2. From these results, one can appreciate that in all cases except one (Avocado-Enron for Commitment), the difference between distributions of two sets of sentences in the same vs. different domains is different by an order of magnitude. In the dense encoding representation, we see a large difference in the distribution across domains, and can expect a gap in the performance of predictive models when trained within vs. across domains. The main takeaways from the intrinsic analyses are: **(1)** the transfer gap can be observed in individual words (n-grams) used in sentences that express intent, **(2)** the transfer gap can be observed in terms of the context distribution in the top words that express intent, **(3)** the transfer gap can be observed in terms of the sentence encodings of the sentences in which they occur. Therefore, an ideal transfer learning method to bridge the transfer gap would need to address these word-level and sentence-level differences.

Meeting	MMD	Rel. to in Domain
Avo-Avo	0.00037	1×
Avo-Enron	0.0071	19.2×
Enron-Enron	0.00038	1×
Enron-Avo	0.002	5.3×

Commitment	MMD	Rel. to in Domain
Avo-Avo	0.0013	1×
Avo-Enron	0.0011	0.85×
Enron-Enron	0.0003	1×
Enron-Avo	0.004	13.3×

Table 2: Maximum Mean Discrepancy (MMD) of the distributions of encodings for meeting intent (left) and commitment (right). "Avo-Avo" compares the encodings of two disjoint sets of sentence encodings from Avocado. "Avo-Enron" compares sentence encodings from Avocado and Enron, obtained by first training the CNN on Avocado labeled data.

3 Extrinsic Analyses

Now that we have established through the intrinsic analyses that there is a distributional difference, we analyze how that difference is reflected in terms of predictive performance of state of the art text classification and domain adaptation methods. We use two classes of methods: non-transfer text classification methods without accounting for the difference in distribution, and methods which perform domain adaptation. To provide a measure of the extrinsic transfer gap, we fix the training set and train each model on the training set and apply it in-domain and out-of-domain. This is equivalent to training on one enterprise's data and deploying the model both to that enterprise (in-domain) and to another enterprise (out-of-domain). Because the ratio of positive/negatives is different in the in-domain test set versus the out-of-domain test set, we use AUC as our performance metric which is a ranking-based metric that is invariant to class-skew. Thus it is reasonable to expect a similar performance on the out-of-domain test set on AUC if it really is from the same distribution as the in-domain. We prefer AUC over other performance metrics (like average precision and F1) since they are sensitive to class skew making it hard to use them for comparisons across test sets where the class skew changes.

Transfer Methods The two domain adaptation methods that we use are: **(1)** an mSDA autoencoder [1] which combines unlabeled data from the source and target domains to learn a joint feature representation that can then be used with a regular classifier like logistic regression, and **(2)** a domain adversarial deep learning method using a CNN architecture [2], which combines the labeled source domain data and the unlabeled target domain data, to extract a latent representation that is invariant across domains, and still useful for performing classification in the source domain – we call this CNN+ADV. **Non-transfer baselines**: as baselines, we apply models which are not designed to perform any transfer across domains, and inherently assume that the source and the target domains follows the same distribution. The non-transfer text classification methods we use are: **(1)** L1-regularized logistic regression, which uses sparse BoW representation; **(2)** A CNN , which encodes each sentence in a dense encoding using word embeddings as input – to provide a controlled comparison this is the same CNN as in CNN+ADV essentially eliminating the adversarial training; **(3)** mSDA-NT where unlabeled data from only the source domain is used when training the autoencoder.

Examining the results in Table 3, it is clear comparing the upper in-domain line in each pair of rows to the out-of-domain line that regardless of the model there is a transfer gap (higher in-domain performance) in all cases except one. That one case is when transferring from Avocado to Enron in

		No Transfer			Transfer	
Task	Train → Test	LR	CNN	mSDA-NT	CNN+ADV	mSDA
Meeting	Avocado → Avocado	0.91	0.93	0.93	–	–
	Avocado → Enron	0.89	0.92	0.91	0.92	0.90
	Enron → Enron	0.92	0.94	0.93	–	–
	Enron → Avocado	0.89	0.90	0.90	0.91	0.91
Commitment	Avocado → Avocado	0.95	0.97	0.96	–	–
	Avocado → Enron	0.98	0.99	0.99	0.99	0.98
	Enron → Enron	0.99	0.99	0.99	–	–
	Enron → Avocado	0.93	0.96	0.95	0.95	0.94

Table 3: AUC Scores for both Prediction Tasks. Each pair of adjacent rows in a block compares the model applied in-domain (upper row) and the model applied out-of-domain (lower row). The transfer out-of-domain results (e.g. CNN+ADV, mSDA) can be compared to the corresponding no Transfer in-domain (e.g. CNN, mSDA-NT).

Commitments. Referring back to the MMD results on right of Table 2, note that MMD actually shows Enron is closer to Avocado than a sample of Avocado is! This can be interpreted as the variation in the language in commitments in Enron is smaller than that seen in Avocado; therefore a random sample of Enron commitments has less variability. Given that this predicts the direction of better out-of-domain performance, understanding how to leverage this better in learning is an interesting future direction. Now, note that both of the transfer methods generally improve over the logistic regression baseline but not over their most similar no-transfer counterpart. In much of the transfer learning literature a new model and training approach is introduced jointly and only compared to a baseline. Here we see by comparing to a no-transfer version, the gains on out-of-domain relative to a simple baseline is due to improved modeling but that it has not closed the size of the gap between in-domain and out-of-domain performance. In fact, many transfer papers do not compute in-domain performance nor compare on a performance metric that is skew invariant.

4 Conclusion

We presented several ways of demonstrating an intrinsic transfer gap in the distributions of words across different text corpora. Using contextual embeddings, MMD based distances on the positive class were especially accurate in predicting how models would perform on out-of-domain datasets. While methods proposed for transfer in the literature do provide improved gains compared to simple baselines, we found the change in model architecture to be the primary explanation for improved absolute performance. Furthermore, techniques like adversarial training did not add additional improvement – nor did the gap between in-domain and out-of-domain performance decrease with transfer based methods. Investigating the reasons for this in detail may be a fruitful direction in order to improve cross-domain prediction in enterprise text data.

References

[1] M. Chen, Z. Xu, K. Weinberger, and F. Sha. Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683*, 2012.

[2] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570, 2018.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[6] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.