Robust Domain Adaptation for Semantic Role Labeling

Chandan Reddy Akiti, Sai Ajay Modukuri
{cra5302, svm6277} @ psu.edu

1 Introduction

Semantic Role Labeling(SRL) is an important supervised NLP task with applications in question answering, information extraction. However, SRL models are highly domain-dependent. Pradhan et al. (2008) showed that the out-of-domain SRL F1-score on Brown corpus is found to be worse. In the real world, it's impractical to assume that the data are independent and identically distributed (i.i.d.) to train the SRL model for each domain. Therefore, much recent literature proposed domain adaptation methods for SRL models. For example, Gal and Ghahramani (2016) showed a domain adaptation approach to perform SRL on clinical corpora showing the importance of domain-robustness in SRL tasks. In this paper, we propose a novel approach to train a cross-domain robust SRL model based on Distributionally Robust Language Modeling proposed by (Oren et al., 2019).

2 Problem Statement

Frame-semantic parsing (Gildea and Jurafsky, 2002) is the task of identifying the semantic frames evoked in text, along with their arguments, formalized in the FrameNet project (Baker et al., 2007). An example of Frame-semantic parsing is shown in figure 1. Target words and phrases are highlighted in the sentence, and their lexical units are shown italicized below. Frames are shown in colored blocks, and frame element segments are shown horizontally alongside the frame.

Domain dependency of the SRL task is a disadvantage (Gal and Ghahramani, 2016) when we adapt the trained models to a new task or domain. This is especially evident when the test set is from a completely different distribution when compared to the training dataset distribution. The 9, illustrates how a supervised learning model like Maximum Likelihood Estimation(MLE) emphasizes heavy-

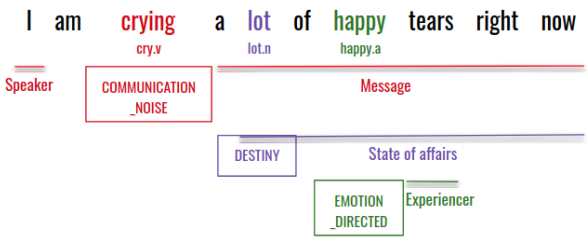

Figure 1: An example of Semantic Role Labeling

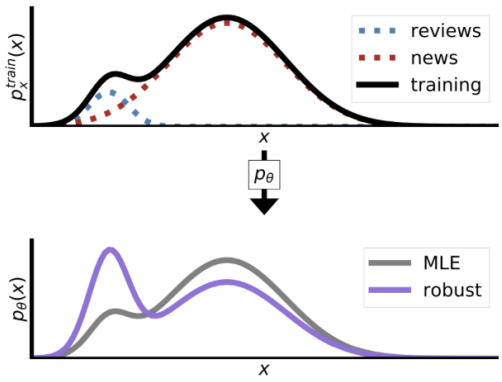

Figure 2: Figure showing that supervised learning such as MLE cannot help train robust model when training data is not in the same distribution as test set and an ideal robust case (Oren et al., 2019).

weight on the distributions with more data while ignoring distributions with fewer data. We try to make the model more robust under the cross-domain setting.

Formally, the model receives sequential words $\boldsymbol{x}$, as input to predict a sequence of labels $\boldsymbol{y}$, where $\boldsymbol{x} \in \mathcal{X}$ is set of vocab, $\boldsymbol{y} \in \mathcal{Y}$ is a discrete set of tag. Consider $\mathcal{Q}$ with a mixture of K distribution. In the traditional supervised machine learning setting, the goal is to minimize the following term

$$\inf_{\theta \in \Theta} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{Q}}[-log(P_\theta(\boldsymbol{x})]$$

Where Θ is a model family. The above expectation

can be approximated by $\inf_{\theta \in \Theta} -\frac{1}{N} \sum_{i=1}^{N} P(\boldsymbol{x}_i|\theta)$ when $N \to \infty$, under the assumption that the training data is from the same distribution as the test set. However, in our setting, the adversary will impose an unknown K dimensional weight vector $\boldsymbol{q}$ on test data distribution (Oren et al., 2019). For the robustness purpose, we aim to minimize the worst-case through sampling $\mathcal{G} \triangleq \widehat{\boldsymbol{q}}\mathcal{Q}$ as:

$$\inf_{\theta \in \Theta} \sup_{\mathcal{G} \in \mathcal{Q}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{G}}[-log(P_\theta(\boldsymbol{x})]$$

3 Data

For our SRL task, we will be using FrameNet (Baker et al., 2007) dataset. The dataset is part of CoNLL shared tasks for SRL. The CoNLL-2005 task is bundled with software to compute precision, recall, and the F1 measure for recognized arguments, which can be used for task evaluation of both the datasets. These datasets provide gold predicates as part of the input.

In addition to the gold predicates and role labeling, these datasets are annotated to provide various other useful syntactic and semantic information such as parts of speech, chunks, clauses, named entities, and parse trees.

The FrameNet 1.7 dataset has with 10,147 sentences with corresponding semantic role labeling. The data consists of sentences from seven sources, namely LUCorpus-v0.3, WikiTexts, ANC, PropBank, NTI, KBEval, and Miscellaneous. The data distribution from each of these sources is shown in figure 9.

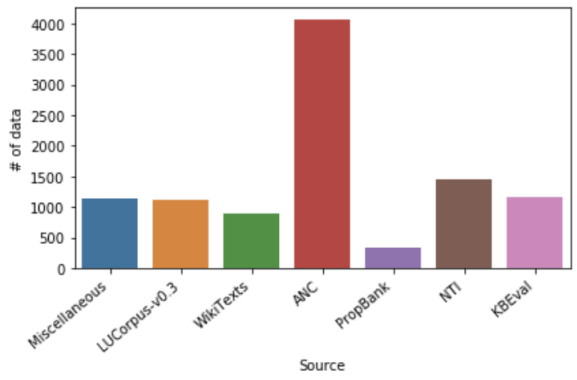

Figure 3: Data distribution from different sources

4 Approach

Our task is to increase the robustness of the SRL task across domains. We measure the robustness of the model by the average performance gain across domains. Since the domains are not explicitly mentioned in the dataset, we perform a topic clustering on the documents in the dataset and analyze the distribution of topics on the dataset. We use the distribution of topics along with the domains to model an online learning mechanism, as described in (Oren et al., 2019). This method improves the robustness at the domain-level and topic level clusters in the data.

Our baseline training architecture is the same as Swayamdipta et al. (2017). The two major contributions of this approach are a pipelined approach, incorporating features from automatic dependency or phrase-structure parsers, and a syntactic scaffolding approach, discarding the need for a syntactic parser altogether.

Formally, the first part of the pipeline detects the predicate set $\mathbf{v}$ using a bidirectional LSTM. The predictions come from a softmax on the output of bidirectional lstm trained to maximize the likelihood of gold labels. The second part of the pipeline is given a sentence-predicate pair $(\mathbf{w}, \mathbf{v})$ as input. The model detects the Semantic-Frame invoked by the corresponding predicate in the sentence. These Semantic-Frames are input to the third part of the pipeline. This model outputs the frame elements for each Semantic-Frame detected in the previous step.

The clustering of the dataset based on topics is useful in a better understanding of distribution of a dataset. It can also be used to sample data from different distributions for online training purpose. For the clustering, we experimented with both LDA (Blei et al., 2003) and Kmeans Clustering ().

K-means helps us better understand the distribution of dataset. Different strategies to find optimal number of clusters are experimented as per (Kodinariya and Makwana, 2013). However, for better modeling of topic clusters, we adopt Latent Dirichlet Analysis (LDA).

LDA is used for topic modeling and clustering documents based on latent topic representation of documents. To find the optimal number of topics, we use the coherence score as an anchor.

We use the online learning approach to minimize the worst-case error on each topic for the weight vector imposed by the adversary. This method of online learning was first proposed by (Oren et al., 2019) for NLP tasks. We will then measure the algorithm performance through standard regret. In

particular, the online learning algorithm can be viewed as a method to approximate $\widehat{q}$ through sampling training batch in each round.

5 Baseline

Our baseline model is the Segmental RNN proposed by Swayamdipta et al. (2017) for frame element detection.

6 Results

It is critical to identify and model the dataset distribution, as we will be adopting to different sampling strategies for training based on the dataset distribution. We used K-means clustering for better understanding and also to get a ballpark estimate for the optimal number of topics. K-means clustering has been performed of term frequency-inverse document frequency (TFIDF) of the sentences. From the graph 4, using the elbow method, we can see that the optimal number of clusters is around four or eight.

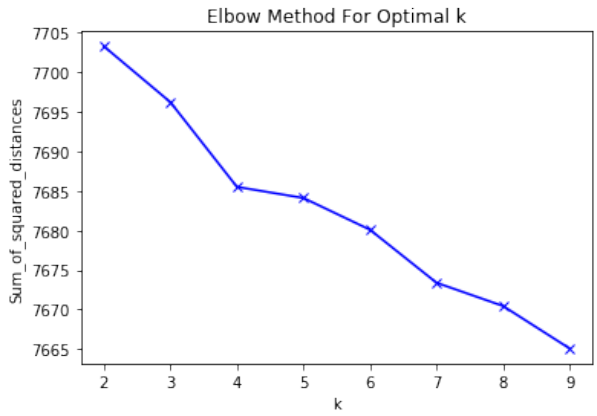

Figure 4: Number of clusters vs sum of squared distances graph for finding optimal clusters using K-means clustering. There are two significant clusters, optimal number of clusters is ambiguous.

LDA provides better topic modelling and clustering. There are two key sampling approaches available in LDA, (i) variational bayes sampling (Blei et al., 2003) implemented in gensim library (Řehůřek and Sojka, 2010), (ii) Gibbs sampling (Porteous et al., 2008) available in mallet library (McCallum, 2002).

The coherence score is used as an anchor to compare different implementations of LDA (i.e., variational Bayes, Gibbs sampling) and also for the estimation of the optimal number of clusters. A better coherence score implies well-modeled topics, and the score generally improves with an increasing number of topics. However, the key is finding the right balance between a number of topics and a coherence score. The coherence score, starting from two topics up to thirty topics, is calculated and plotted on a graph. We found that a model of eight topics provides the optimal balance between the number of topics and the coherence score for our dataset. The 5 shows the performance of gensim LDA and mallet LDA.

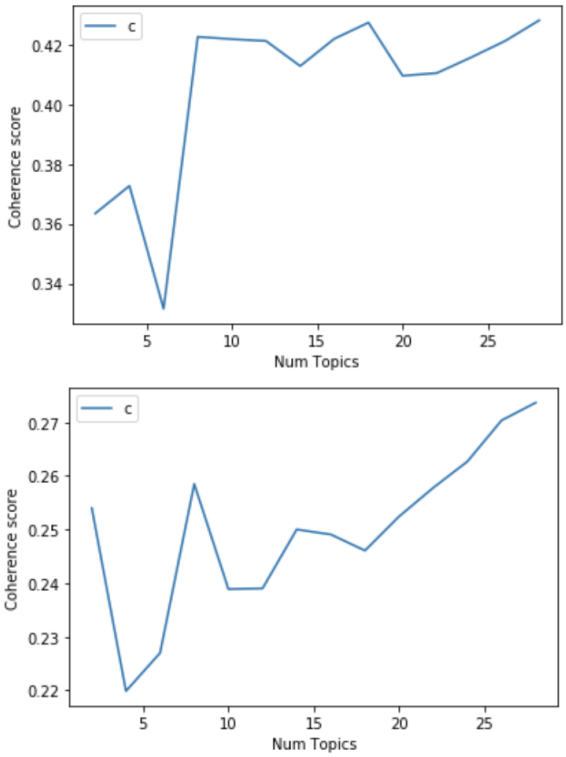

Figure 5: Coherence score of mallet LDA (top) and gensim LDA

After the topic modeling and clustering steps, we experimented our Distributionally Robust Optimization(DRO) for the SRL tasks under two settings: (i) A setting in which, we know the distribution of both training set and testing set beforehand, (ii) a setting where we only know the distribution of training data and have no information about the test set distribution. The second setting is more common in real-world situations where test data distribution is unknown. We will discuss in detail about the results of both the settings in the upcoming sub-sections.

6.1 Model 1: Known test distribution

We assume that both training and test set distributions are known. We thus model our online learning

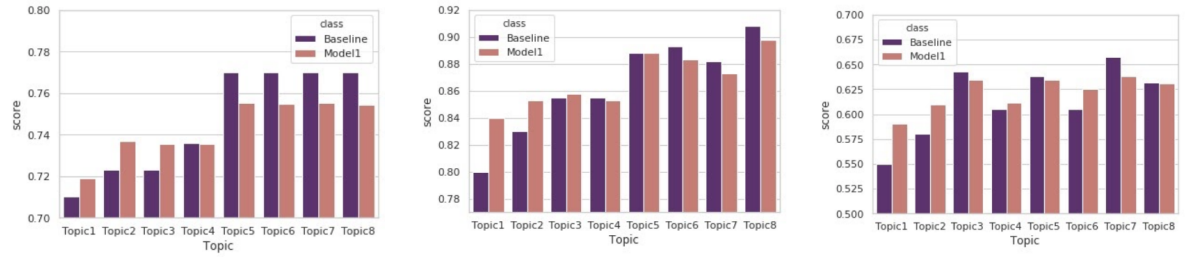

Figure 6: Graphs showing the results over baseline across eight topics when distributions of both train set and test ser are known. We have acheived a 1.1% increase on absolute F1 score in target prediction 0.87% absolute F1 increase in frame prediction and 0.02% absolute F1 increase in frame element prediction steps. However, we can see that there is a huge increse in F1 score for under represented distributions.

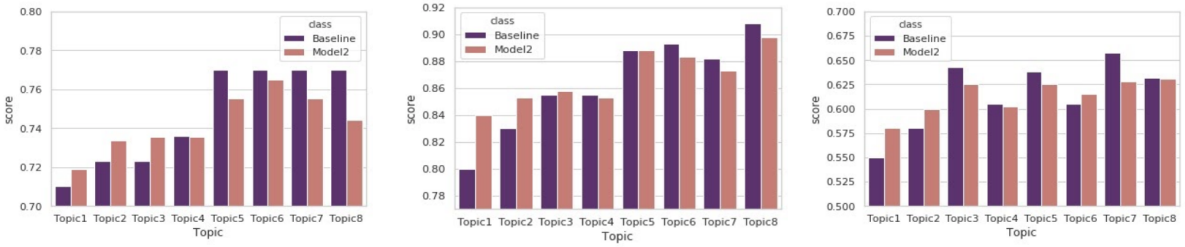

Figure 7: Graphs showing the results over baseline across eight topics when distribution of only training set is known. We have acheived a 1.0% increase on absolute F1 score in target prediction 0.86% absolute F1 increase in frame prediction and 0.54% absolute F1 decrease in frame element prediction steps.

algorithm to minimize the maximum worst case error on each batch of training set based on the prior distribution.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \max_{z \in \mathcal{Z}_{all}} \mathbb{E}_{(x,y)\in z} l(\theta; (x, y))$$

where Θ is the online learning parameter space, $\mathcal{Z}_{all}$ is the train+test domain space.

The online algorithm updates the θ parameter for sampling the higher error domain samples at a higher rate. This improves the model performance on worst case domains while improving the overall performance. This is a min-max model of robust optimization.

We update the domain sampling weight using the above objective function in every iteration. The domain space $\mathcal{Z}_{all}$ refers to the training and test data domains. In this experiment, we assume that the test data distribution is available for training. However, we do not do any optimization of the training process specifically for test distribution.

We achieve a 1.1% increase on absolute F1 score in target prediction with 83% precision and 68% recall, 0.87% absolute F1 increase in frame prediction with 90% precision and 85% recall, and 0.02% absolute F1 increase with 74% precision and 0.55% recall in frame element prediction steps. However, we can see that there is a huge increase in the F1 score for underrepresented distributions.

The main observation is that the F1-score for the underrepresented domains in the dataset has increased at the cost of a small decrease in performance on dominant domains. Thus the model is more robust than the baseline.

6.2 Model 2: Unknown test distribution

We do not know the test distribution in the real-world scenario. Thus there is a need for the model to be robust on the unknown prior test distribution.

In this model, we assume that both training data distributions is known, but test data distribution is unknown. We thus model our online learning algorithm to minimize the maximum worst case error on each batch of training set based on the prior distribution.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \max_{z \in \mathcal{Z}_{train}} \mathbb{E}_{(x,y)\in z} l(\theta; (x, y))$$

where Θ is the online learning parameter space, $\mathcal{Z}_{train}$ is the train-only domain space.

We update the domain sampling weight using the above objective function in every iteration. The domain space $\mathcal{Z}_{train}$ refers to the training data domain. In this experiment, we assume that the test data distribution is available for training. However, we do not do any optimization of the training process specifically for test distribution.

Similar to model 1, we observe that this model also has robust performance compared to the baseline. The test data distribution knowledge increases the performance, as seen in model 1. However, the training data distribution itself is sufficient to make the model more robust.

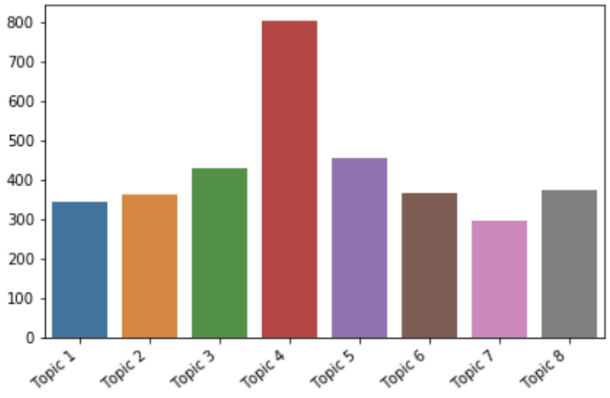

Figure 8: Graph showing a combined training and test distribution.

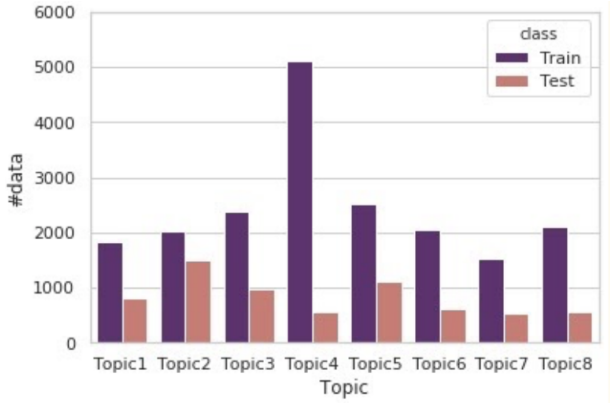

Figure 9: Graph showing the distributions of training set and test set separately.

7 Conclusion

We have used the online learning approach of (Oren et al., 2019) to train a robust Semantic Role Labelling model on the baseline (Swayamdipta et al., 2017). We show that training data distribution and test data distribution of domains in the dataset can be leverages to make the model robust across the domain.

As future work, we plan to use contextual language models like BERT, and it's robust version RoBERT to improve the model accuracy and robustness.

8 Acknowledgements

The project is divided into five modules.

- dataset preprocessing, analyses, topics modeling, and clustering.
- Selection and running of Baselines.
- Implementation of the online learning framework.
- Implementation of sampling strategy for training.
- Ablation study without domain distribution.

Ajay has initially worked on dataset preprocessing, analysis, and topic clustering using LDA and K-means clustering to produce the topic and domain distributions necessary for online learning. He also worked on baseline selection by running the implementation of (He et al., 2017).

Chandan also worked on the selection of baseline by running the implementation of (Swayamdipta et al., 2017), which has been used as the baseline. He also worked on the deep learning framework for online learning on top of baseline (Swayamdipta et al., 2017). He also worked on the implementation of sampling strategy from topic distributions.

Ajay has done ablation studies of robustness using our novel approach.

References

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, Czech Republic. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing*

Systems, NIPS'16, page 1027–1035, Red Hook, NY, USA. Curran Associates Inc.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483.

Trupti M Kodinariya and Prashant R Makwana. 2013. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally robust language modeling.

Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577.

Sameer S. Pradhan, Wayne Ward, and James H. Martin. 2008. Towards robust semantic role labeling. *Computational Linguistics*, 34(2):289–310.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *CoRR*, abs/1706.09528.