How Far Can We Go with Data Selection? A Case Study on Semantic Sequence Tagging Tasks

Samuel Louvan
University of Trento
Fondazione Bruno Kessler
slouvan@fbk.eu

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

Abstract

Although several works have addressed the role of data selection to improve transfer learning for various NLP tasks, there is no consensus about its real benefits and, more generally, there is a lack of shared practices on how it can be best applied. We propose a systematic approach aimed at evaluating data selection in scenarios of increasing complexity. Specifically, we compare the case in which source and target tasks are the same while source and target domains are different, against the more challenging scenario where both tasks and domains are different. We run a number of experiments on semantic sequence tagging tasks, which are relatively less investigated in data selection, and conclude that data selection has more benefit on the scenario when the tasks are the same, while in case of different (although related) tasks from distant domains, a combination of data selection and multi-task learning is ineffective for most cases.

1 Introduction

Transfer learning is a common approach for training NLP models that scale across different tasks, domains, and languages. One of the challenges in transfer learning is to deal with the data distribution mismatch between the source ($\mathcal{D}_S$) and the target data ($\mathcal{D}_T$) (Rosenstein et al., 2005). One solution to alleviate the impact of the mismatch is using data selection, a process for selecting relevant training instances from the source data. Data selection (DS) has been applied in the context of domain adaptation to address changes in the data distribution for various NLP tasks, such as sentiment analysis and POS Tagging (Ruder and Plank, 2017; Liu et al., 2019; Blitzer et al., 2007; Remus, 2012), machine translation (Axelrod et al., 2011), dependency parsing (Søgaard, 2011) and Named Entity Recognition (NER) (Murthy et al., 2018; Zhao et al., 2018). To our knowledge, all existing previous works apply data selection to *different* domains, while maintaining the *same* task.

In this work we aim to investigate the benefit of data selection in a more complex setting, where we have not only different domains ($\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$), but also different tasks ($\mathcal{T}_\mathcal{S} \neq \mathcal{T}_\mathcal{T}$). Intuitively, such setting may bring advantage in situations where large training data are available for a source task $\mathcal{T}_\mathcal{S}$, and we want to exploit such data for a different (although related) target task $\mathcal{T}_\mathcal{T}$, where much less training is available. We experiment with the situation where $\mathcal{T}_\mathcal{S}$ is Named Entity Recognition (NER) on a general domain, where several datasets are available, and $\mathcal{T}_\mathcal{T}$ is slot tagging (ST) in the context of utterance interpretation for dialogue systems, where much less data is available. Both of the tasks are rarely investigated in data selection and there is no consensus about the benefit of data selection for them.

We propose an experimental framework where we can compare data selection settings with an increasing level of complexity. First, we consider data selection where NER is both the source and target task, and apply transfer learning from different domains: we call this setting **S**ame **T**asks from **D**ifferent **D**omains (**STDD**), $\mathcal{T}_\mathcal{S} = \mathcal{T}_\mathcal{T}$ and $\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$. In a second, more complex setting, we consider NER as the source task and ST as the target: this is called $\mathcal{T}_\mathcal{S} \neq \mathcal{T}_\mathcal{T}$ and $\mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$, **D**ifferent **T**asks from **D**ifferent **D**omains (**DTDD**). In this scenario, as we have disjoint label space between the source and the target task, we combine the data selection process with multi-task learning (MTL). To our knowledge, this combination has received very little attention in the literature.

We base our work on the data selection framework proposed by Ruder and Plank (2017), and apply it to our experimental settings. Their framework is model-agnostic and has shown significant advantage in sentiment analysis, POS tagging, and

Proceedings of the First Workshop on Insights from Negative Results in NLP, pages 15–21

ISBN 978-1-952148-66-8

parsing. However, it is not obvious to what extent the selection process can actually help on semantic sequence tagging tasks on STDD and DTDD scenarios. The contributions of the paper are the following: (i) we apply previous work to multi-task learning setup to evaluate the effectiveness of data selection in DTDD scenarios; (ii) we systematically compare data selection on settings of increasing complexity, and observe that existing selection metrics do not show clear advantages over baselines in most cases. Nevertheless, data selection has more potential in STDD when source and target are more similar, while *combining* MTL and data selection for DTDD is *ineffective* for most cases in our experimental settings in which we have different but related tasks (NER and ST) from relatively distant domains (news and conversational domains).

2 Data Selection Framework

In general, the goal of data selection is to select an optimal subset of training instances, X_S^*, from all the available data X_S in $\mathcal{T}_S$, to be used for training the model for the target task $\mathcal{M}_{\mathcal{T}_T}$. Given the source data $X_S = \{x_1^S, x_2^S, ..., x_n^S\}$, each instance is ranked according to a score $\mathcal{S}$ and the top m examples are then used to train $\mathcal{M}_{\mathcal{T}_T}$.

We apply the data selection approach from Ruder and Plank (2017), based on Bayesian Optimization (BO) (Brochu et al., 2010), to evaluate the effectiveness of data selection on both the STDD and DTDD scenarios. Specifically, for DTDD we *combine* data selection and multi-task learning. Given X_S, the framework performs data selection based on a score $\mathcal{S}$ derived from a set of features. The top m examples are then used to train $\mathcal{M}_{\mathcal{T}_T}$. In case of STDD, the $\mathcal{M}_{\mathcal{T}_T}$ is a single task sequence tagging model, where we use a biLSTM-CRF model (Lample et al., 2016). As for DTDD, $\mathcal{M}_{\mathcal{T}_T}$ is a *hard parameter sharing* MTL model, which has been applied to many NLP tasks (Søgaard and Goldberg, 2016; Plank et al., 2016; Changpinyo et al., 2018; Schulz et al., 2018). The performance on the validation set of the target task is then used by the BO optimizer to update the weight of the scoring features.

Following Ruder and Plank (2017), the selection process is based on a score $\mathcal{S}$ computed as the linear combination of weighted features, which include both similarity and diversity features: $\mathcal{S}_\theta(x) = \theta^\top \cdot \phi(x)$, where θ represents the weight for each feature and $\phi(x)$ denotes the feature values of each instance x. The features are calculated between the representation of X_S instances and X_T. We use term distribution as the representation of the instances. We use the same similarity and diversity measures as Ruder and Plank (2017). The weights θ are learned through BO by taking into account the performance on the validation set when selecting a particular subset of X_S. The score $\mathcal{S}$ is computed for each x in X_S, and then the top m examples are selected for training the $\mathcal{M}_{\mathcal{T}_T}$ model. The loss value $\mathcal{L}$ from the $\mathcal{M}_{\mathcal{T}_T}$ in the validation set is used by BO as a feedback to select the next points for θ.

3 Experiments

We systematically investigate how data selection is effective when applied on both the STDD and DTDD scenarios. We address two semantic sequence labeling tasks: Named Entity Recognition (NER) and slot tagging (ST).

3.1 Datasets

For NER we use the OntoNotes 5.0 (Pradhan et al., 2012) dataset, which consists of several sections: newswire (NW), talkshows broadcast (BC), telephone conversation (TC), news broadcast (BN), articles from web sources (WB), and articles from magazines (MZ). We use different OntoNotes sections as different domains in our experiments.

As for ST we use three datasets: ATIS (Price, 1990), MIT-R, and MIT-M (Liu et al., 2013), that are widely used as benchmarks for spoken language understanding. Each dataset contains utterances annotated with domain-specific slot labels, which are typically more fine-grained than NER labels. For example, in the utterance "*show me all **Delta** flights from **Milan** to **New York***", the bold words are tagged as *airline_name*, *fromloc*, and *toloc* respectively. The overall statistics of each dataset are shown in Table 1.

3.2 Data Selection Configurations

We make use of the selection framework described in Section 2, and apply three Bayesian Optimization data selection (*BODS*) configurations, according to whether we use features both for similarity and diversity ($\text{DS}_{\text{sim,div}}$), similarity features only (DS_{sim}), or diversity features only (DS_{div}). We compare the three configurations with the following baselines:

- All source, which uses all the data from $\mathcal{T}_S$.
- Random, which selects random data from $\mathcal{T}_S$.

Dataset	#train	#dev	#test	#label
Slot Tagging				
ATIS	4478	500	893	79
MIT Restaurant	6128	1532	3385	8
MIT Movie	7820	1955	2443	12
NER				
OntoNotes NW	34970	5896	2327	18
OntoNotes BC	11879	2117	2211	18
OntoNotes TC	12891	1634	1366	18
OntoNotes BN	10683	1295	1357	18
OntoNotes WB	16598	2316	2307	18
OntoNotes MZ	6911	642	780	18

Table 1: Statistics about the datasets used in the experiments. The language of the datasets is English.

- $\text{DS}_{\text{map,full}}$. We provide a manual mapping from NER labels to ST labels (Appendix A). A sentence from $\mathcal{T}_\mathcal{S}$ is selected is if *all* the NER occurrences have a mapping to a slot in $\mathcal{T}_\mathcal{T}$.
- $\text{DS}_{\text{map,partial}}$. A sentence from $\mathcal{T}_\mathcal{S}$ is selected if *at least* one of the NER occurrences in the sentence has a mapping to a slot label in $\mathcal{T}_\mathcal{T}$.

3.3 Settings

We follow most of the hyperparameters[1] as recommended by Reimers and Gurevych (2018). We train the model for $\mathcal{T}_\mathcal{S}$ and $\mathcal{T}_\mathcal{T}$ in an alternating fashion. We use early stopping on the dev. performance of $\mathcal{T}_\mathcal{T}$. For the model performance evaluation, we calculate the F1-score using the standard CoNLL script[2]. For all experiments, we report the average F1 score results from 10 runs with different seeds.

We follow Ruder and Plank (2017) for most configurations of the optimizer, and run 50 iterations. For both the STDD and DTDD scenarios, we select top 50%[3] examples from X_S. For MTL we adapt the implementation from Reimers and Gurevych (2017), extending the Bayesian Optimization data selection framework from Ruder and Plank (2017) to support MTL.

4 STDD Scenario: $\mathcal{T}_\mathcal{S} = \mathcal{T}_\mathcal{T}, \mathcal{D}_\mathcal{S} \neq \mathcal{D}_\mathcal{T}$

This is scenario is the same setup as Ruder and Plank (2017), where we use the same tasks both for the source and the target task from different domains, except that we apply the data selection to a semantic sequence tagging task namely NER. In this scenario, we use NER both for the source and the target task. The target domain is one three OntoNotes sections namely NW (news), TC (telephone conversation) and BC (mixed of conversation and broadcast) while as source domain ($\mathcal{D}_\mathcal{S}$) we use all available sections in OntoNotes except the one used as the target domain. We only use 10% of training data for the target domain to simulate limited data settings. At the end of the data selection process, we select the top 50% sentences from $\mathcal{D}_\mathcal{S}$ using the best feature weights learned with the Bayesian Optimizer.

Table 2(a) compares the performance of the baselines with the selection-based approaches. In general, we do not observe clear advantages of data selection methods over the baselines, especially the all source data baseline. Using all source data yields the most competitive results almost in all cases. The only case in which DS surpasses the all source baseline is on the BC domain but only for a tiny gain. For NW and BC domains, some DS methods show clear advantages over the random baseline, but still worse than using all source data.

We want to see whether the distance between domains may characterize the performance of the data selection. For this purpose we quantify the domain similarity between each pair $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$ with Jensen Shannon Divergence (JSD) (Lin, 1991). We compute the JSD between the term distribution of $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$. The average JSD of each target task with respect to the source tasks are 0.80 (TC), 0.86 (NW), and 0.87 (BC)[4]. We observe that the higher the JSD is, the more beneficial is the data selection for the target task. BC, which has the highest JSD average, benefits the most from the data selection. On the other hand, TC with the lowest average similarity, has the largest gap between the baseline and the best DS methods (-1.7 F1 point).

Based on our experiments, for the STDD scenario we observe that:

1. In most of the cases, DS methods are inferior to the all source baseline. Yet, it is clear that each domain has a different selection metric configuration that performs the best. This observation suggests that the hypothesis from Ruder and Plank (2017) i.e., different tasks or even different domains demand a different notion of selection metric, is also applicable to semantic sequence tagging tasks such as NER.
2. The gap between the best DS method and the baseline for each $\mathcal{D}_\mathcal{T}$ can be characterized from the average JSD similarity to its $\mathcal{D}_\mathcal{S}$. Being

[1] Appendix C reports all used hyperparameters.
[2] https://www.clips.uantwerpen.be/conll2000.
[3] We tune from 10% to 50% on the dev set.
[4] Complete pairwise JSD values are listed in Appendix B.

Method	TC	NW	BC
Baseline			
All source	$\mathbf{63.17}_{4.75}$	$\mathbf{79.08}^{\dagger}{}_{0.42}$	$\underline{73.42}_{2.13}$
Random	$\underline{62.02}_{4.47}$	$77.93_{0.54}$	$71.39_{2.12}$
BODS			
$\text{DS}_{\text{sim,div}}$	$61.71_{4.57}$	$76.99_{0.40}$	$72.60_{1.14}$
DS_{sim}	$61.45_{3.80}$	$78.30_{0.41}$	$\mathbf{73.44}_{\mathbf{1.12}}$
DS_{div}	$61.65_{3.77}$	$\underline{78.32}_{0.53}$	$71.89_{1.53}$

(a) STDD

Method	ATIS	MIT-R	MIT-M
STL			
biLSTM-CRF	$85.46_{0.25}$	$63.99_{0.77}$	$76.39_{0.57}$
Baseline (MTL)			
All source	$\mathbf{90.05}_{\mathbf{0.34}}$	$69.28_{0.40}$	$81.28_{0.23}$
Random	$89.93_{0.26}$	$\mathbf{69.54}_{\mathbf{0.35}}$	$\mathbf{81.35}_{\mathbf{0.31}}$
$\text{DS}_{\text{map,full}}$	$89.97_{0.25}$	$68.82_{0.50}$	$79.27_{0.36}$
$\text{DS}_{\text{map,partial}}$	$89.85_{0.29}$	$69.24_{0.40}$	$80.76_{0.30}$
MTL+BODS			
$\text{DS}_{\text{sim,div}}$	$89.78_{0.39}$	$69.29_{0.37}$	$81.07_{0.29}$
DS_{sim}	$89.83_{0.31}$	$69.25_{0.41}$	$81.17_{0.25}$
DS_{div}	$89.95_{0.41}$	$69.09_{0.24}$	$81.10_{0.28}$

(b) DTDD

Table 2: Average F1-score and standard deviation on the test set. † indicates significant differences ($p < 0.05$) between the best BODS approach and the best baseline.

more similar to other $\mathcal{D}_S$ is a more suitable situation to get benefit from data selection.

5 DTDD Scenario: $\mathcal{T}_S \neq \mathcal{T}_T, \mathcal{D}_S \neq \mathcal{D}_T$

In this scenario we intend to observe whether data selection adds benefit to MTL. As in the STDD case, data selection is performed on the auxiliary task, where data is assumed to be abundant, and we only use a small portion of data for the target task. We use NER as the auxiliary task and ST as the target task. Prior work from Louvan and Magnini (2019) shows that NER is helpful for ST through MTL, although it is not clear whether adding data selection is beneficial. We follow the setup in Louvan and Magnini (2019), where OntoNotes NW is used as the auxiliary task, and the target task is one of the ST datasets with only 10% of available training data.

Observing the results in Table 2(b), in all the cases the baselines, namely all source data and random selection, perform better than MTL with DS methods. The selection methods based on manual label mapping, DS_{map}, do not bring advantage over all source data. Therefore, given two distant $\mathcal{D}_S$ and $\mathcal{D}_T$, selecting sentences based on the label mapping does not help. Moreover, as random selection gives good results as well for most scenarios, this indicates that data selection is not beneficial in our experimental setting that combines data selection and MTL.

Our findings and lessons learned for DTDD are the following:

1. We observe that MTL performs better than single-task learning (STL) for low-resource slot tagging, confirming the finding from Louvan and Magnini (2019). However, adding data selection for MTL is *ineffective* in our DTDD experimental setup. We hypothesize that MTL learns good common feature representations across tasks, this way inherently helping the model to focus on relevant features even from noisy data in $\mathcal{T}_S$. In addition to that, due to data sparsity in limited training, using all the training data works better because the model may learn a better text representation (sentence encoder). Recent similar work from Schröder and Biemann (2020) which uses information theoretic based for estimating the usefulness of an auxiliary task for MTL also found that for semantic sequence tagging tasks such as NER and argument mining, it is less clear when a particular dataset is useful as an auxiliary task.
2. Data selection typically produces selected sentences with concentrated similarity distribution[5]. Therefore, it is probably ineffective when the sentence similarity distribution between $\mathcal{T}_S$ and $\mathcal{T}_T$ is already concentrated on a very narrow range.

6 Conclusion

In this paper we investigated the benefit of data selection for transfer learning in several scenarios of increasing complexity. We apply an existing model-agnostic state of the art data selection framework, and carried on experiments on two semantic sequence tagging tasks, NER and Slot Tagging, and two transfer learning scenarios, STDD (Same

[5]We embed the sentence in source and target with InferSent (Conneau et al., 2017) and compute cosine similarity between the centroid of the target and each of the sentence in source.

Tasks Different Domains), and DTDD (Different Tasks Different Domains).

For the STDD scenario, selection methods show potential when the target domain has the highest similarity to the source domains, based on Jensen Shannon Divergence. As for the DTDD scenario in which we use related tasks (NER and ST) from distant domains (news and conversational domains), using selection does not bring advantage over using all the source data. A possible cause is that, because of data sparsity on the target task, it is only by injecting more source data that we can improve the model. Finally, MTL does not benefit from data selection, as it may already effectively help the model to focus on relevant features even though in the presence of noisy data from distant domains.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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.

Eric Brochu, Vlad M Cora, and Nando De Freitas. 2010. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Soravit Changpinyo, Hexiang Hu, and Fei Sha. 2018. Multi-task learning for sequence tagging: An empirical study. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2965–2977, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Trans. Information Theory*, 37:145–151.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013. Asgard: A Portable Architecture for Multilingual Dialogue Systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8386–8390. IEEE.

Miaofeng Liu, Yan Song, Hongbin Zou, and Tong Zhang. 2019. Reinforced training data selection for domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1957–1968, Florence, Italy. Association for Computational Linguistics.

Samuel Louvan and Bernardo Magnini. 2019. Leveraging non-conversational tasks for low resource slot filling: Does it help? In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 85–91, Stockholm, Sweden. Association for Computational Linguistics.

Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2018. Judicious selection of training data in assisting language for multilingual neural NER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–406, Melbourne, Australia. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Patti J Price. 1990. Evaluation of Spoken Language Systems: The ATIS Domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 338–348, Copenhagen, Denmark.

Nils Reimers and Iryna Gurevych. 2018. Why Comparing Single Performance Scores Does Not Allow to Draw Conclusions About Machine Learning Approaches. *CoRR*, abs/1803.09578.

Robert Remus. 2012. Domain adaptation using domain similarity-and domain complexity-based instance selection for cross-domain sentiment analysis. In *2012 IEEE 12th international conference on data mining workshops*, pages 717–723. IEEE.

Michael T. Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G. Dietterich. 2005. To transfer or not to transfer. In *In NIPS'05 Workshop, Inductive Transfer: 10 Years Later*.

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

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

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 682–686, Portland, Oregon, USA. Association for Computational Linguistics.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, Berlin, Germany. Association for Computational Linguistics.

Huasha Zhao, Yi Yang, Qiong Zhang, and Luo Si. 2018. Improve neural entity recognition via multi-task data selection and constrained decoding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 346–351, New Orleans, Louisiana. Association for Computational Linguistics.

A Label Mapping

ATIS Slot	**OntoNotes Label**
AIRLINE_NAME	ORG
AIRPORT_NAME	FAC
ARRIVE_DATE, DAY_NAME, DAY_NUMBER, DEPART_DATE, DEPART_TIME, FLIGHT_DAYS, TIME_RELATIVE, TODAY_RELATIVE	DATE
ARRIVE_TIME, MONTH_NAME, PERIOD_OF_DAY, RETURN_TIME, TIME	TIME
CITY_NAME, FROM_LOC, STATE_CODE, STATE_NAME, STOP_LOC, TO_LOC	GPE
COST_RELATIVE, FARE_AMOUNT	MONEY
DAYS_CODE, ECONOMY, FARE_BASIS_CODE, FLIGHT_MOD, MEAL, MEAL_CODE, MEAL_DESCRIPTION, MOD, FLIGHT_STOP, FLIGHT_MOD, OR, RESTRICTION_CODE, ROUNDTRIP, TRANSPORT_TYPE	O
FLIGHT_NUMBER	CARDINAL

Table 3: Label Mapping from ATIS to OntoNotes.

MIT Movie Slot	**OntoNotes Label**
CHARACTER, ACTOR, DIRECTOR	PER
YEAR	DATE
PLOT, RATING, TITLE, REVIEW, SONG, RATINGS_AVERAGE, GENRE, TRAILER	O

Table 4: Label Mapping from MIT Movie to OntoNotes.

B Domain Similarity

$\mathcal{D}_\mathcal{T}$	$\mathcal{D}_\mathcal{S}$						Avg	Δ
	TC	NW	BC	BN	WB	MZ		
TC	-	0.74	0.84	0.80	0.83	0.77	0.80	1.7
NW	0.74	-	0.85	0.91	0.91	0.90	0.86	0.7
BC	0.84	0.85	-	0.90	0.90	0.86	0.87	0.02

Table 5: Domain Similarity (JSD) for each $\mathcal{D}_\mathcal{T}$ and $\mathcal{D}_\mathcal{S}$

C Hyperparameters

Hyperparameter	**Value**
LSTM cell size	100
Dropout	0.5
Word embedding dimension	300
Character embedding dimension	100
Mini-batch size	128
Clip norm	1
Optimizer	Adam
Number of epoch	20
Early stopping	10

Table 6: Neural model hyperparameters

Parameter	**Adopted value**
Surrogate model	Gaussian Processes with MCMC sampling
Acquisition function	Expected Logarithmic Improvement
Number of initial evaluation points	3
Search space upper bound	1
Search space lower bound	-1
Number of iterations	50

Table 7: Parameters used by the Bayesian Optimizer.