Robust Semantic Parsing with Adversarial Learning for Domain Generalization

Gabriel Marzinotto[1,2], **Géraldine Damnati**[1], **Frédéric Béchet**[2], **Benoît Favre**[2]
(1) Orange Labs / Lannion France
(2) Aix Marseille Univ, CNRS, LIS / Marseille France
gabriel.marzinotto@orange.com, geraldine.damnati@orange.com
frederic.bechet@lis-lab.fr, benoit.favre@lis-lab.fr

Abstract

This paper addresses the issue of generalization for Semantic Parsing in an adversarial framework. Building models that are more robust to inter-document variability is crucial for the integration of Semantic Parsing technologies in real applications. The underlying question throughout this study is whether adversarial learning can be used to train models on a higher level of abstraction in order to increase their robustness to lexical and stylistic variations. We propose to perform Semantic Parsing with a domain classification adversarial task without explicit knowledge of the domain. The strategy is first evaluated on a French corpus of encyclopedic documents, annotated with FrameNet, in an information retrieval perspective, then on PropBank Semantic Role Labeling task on the CoNLL-2005 benchmark. We show that adversarial learning increases all models generalization capabilities both on in and out-of-domain data.

1 Introduction

For many NLP applications, models that perform well on multiple domains and data sources are essential. As data labeling is expensive and time consuming, especially when it requires specific expertise (FrameNet, Universal Dependencies, etc.), annotations for every domain and data source are not feasible. On the other hand, domain biases are a major problem in almost every supervised NLP task. Models learn these biases as useful information and experience a significant performance drop whenever they are applied on data from a different source or domain. A recent approach attempting to tackle domain biases and build robust systems consists in using neural networks and adversarial learning to build domain independent representations. In the NLP community, this method has been mostly used in crosslingual models to transfer information from English to low resource languages in problems and recently in various monolingual tasks in order to alleviate domain bias of trained models.

In the context of Semantic Frame Parsing, we address in this paper the generalization issue of models trained on one or several domains and applied to a new domain. We show that adversarial learning can be used to improve the generalization capacities of semantic parsing models to out of domain data. We propose an adversarial framework based on a domain classification task that we use as a regularization technique on state-of-the-art semantic parsing systems. We use unsupervised domain inference to obtain labels for the classification task.

Firstly we perform experiments on a large multi-domain frame corpus (Marzinotto et al., 2018a) where only a relatively small number of frames where annotated, corresponding to possible targets in an Information Extraction applicative framework. We evaluate our adversarial framework with a semantic frame parser we developed on this corpus and presented in (Marzinotto et al., 2018b). Secondly we checked the genericity of our approach on the standard PropBank Semantic Role Labeling task on the CoNLL-2005 benchmark, with a tagging model proposed by (He et al., 2017). We show that in both cases adversarial learning increases all models generalization capabilities both on in and out-of-domain data.

2 Related Work

2.1 Domain-Adversarial Training

Domain Independence can be approached from different perspectives. A popular approach that emerged in image processing (Ganin et al., 2016) consists in optimizing a double objective composed of a task-specific classifier and an adversarial domain classifier. The latter is called adver-

Proceedings of NAACL-HLT 2019, pages 166–173
Minneapolis, Minnesota, June 2 - June 7, 2019.

sarial because it is connected to the task-specific classifier using a gradient reversal layer. During training a saddle point is searched where the task-specific classifier is good and the domain classifier is bad. It has been shown in (Ganin and Lempitsky, 2015) that this implicitly optimizes the hidden representation towards domain independence.

In NLP problems this approach has successfully been used to train cross-lingual word representations (Conneau et al., 2017) and to transfer learning from English to low resource languages for POS tagging (Kim et al., 2017) and sentiment analysis (Chen et al., 2016). These approaches introduce language classifiers with an adversarial objective to train task-specific but language agnostic representations. Besides the cross-lingual transfer problem, there are few studies of the impact of domain-adversarial training in a monolingual setup. For instance, (Liu et al., 2017) successfully uses this technique to improve generalization in a document classification task. It has also been used recently for varied tasks such as transfer learning on Q&A systems (Yu et al., 2018) or duplicate question detection (Shah et al., 2018) and removal of protected attributes from social media textual data (Elazar and Goldberg, 2018).

2.2 Robustness in Semantic Frame Parsing

In Frame Semantic Parsing, data is scarce and classic evaluation settings seldom propose out-of-domain test data. Despite the existence of out-of-domain corpora such MASC (Passonneau et al., 2012) and YAGS (Hartmann et al., 2017) the domain adaptation problem has been widely reported (Johansson and Nugues, 2008; Søgaard et al., 2015) but not extensively studied. Recently, (Hartmann et al., 2017) presented the first in depth study of the domain adaptation problem using the YAGS frame corpus. They show that the main problem in domain adaptation for frame semantic parsing is the frame identification step and propose a more robust classifier using predicate and context embeddings to perform frame identification. This approach is suitable for cascade systems such as SEMAFOR (Das et al., 2014), (Hermann et al., 2014) and (Yang and Mitchell, 2017). In this paper we propose to study the generalization issue within the framework of a sequence tagging semantic frame parser that performs frame selection and argument classification in one step.

3 Semantic parsing model with an adversarial training scheme

3.1 Semantic parsing model: `biGRU`

We use in this study a sequence tagging semantic frame parser that performs frame selection and argument classification in one step based on a deep bi-directional GRU tagger ($biGRU$). The advantage of this architecture is its flexibility as it can be applied to several semantic parsing schemes such as PropBank (He et al., 2017) and FrameNet (Yang and Mitchell, 2017).

More precisely, the model consists of a 4 layer bi-directional Gated Recurrent Unit (GRU) with highway connections (Srivastava et al., 2015). This model does not rely solely on word embeddings as input. Instead, it has a richer set of features including syntactic, morphological and surface features. (see (Marzinotto et al., 2018b) for more details).

Except for words where we use pre-trained embeddings, we use randomly initialized embedding layers for categorical features.

3.2 Sequence encoding/decoding

For all experiments we use a BIO label encoding. To ensure that output sequences respect the BIO constrains we implement an A* decoding strategy similar to the one proposed by (He et al., 2017). We further apply a coherence filter to the output of the tagging process. This filter ensures that the predicted semantic structure is acceptable. Given a sentence and a word w that is a Lexical Unit (LU) trigger, we select the frame F as being the most probable frame among the ones that can have w as a trigger. Once F is determined, we then mask all FEs that do not belong to F and perform constrained A* decoding. Finally, we improve this strategy by introducing a parameter $\delta \in (-1; 1)$ that is added to the output probability of the *null* label $P(y_t = O)$ at each time-step. By default, with $\delta = 0$ the most probable non-null hypothesis is selected if its probability is higher than $P(y_t = O)$. Varying $\delta > 0$ (resp. $\delta < 0$) is equivalent to being more strict (resp. less strict) on the highest non-null hypothesis. By doing so we can study the precision/recall (P/R) trade-off of our models. This δ parameter is tuned on a validation set and we either provide the P/R curve or report scores for the $Fmax$ setting.

3.3 Adversarial Domain Classifier

In order to design an efficient adversarial task, several criteria have to be met. The task has to be related to the biases it is supposed to alleviate. And furthermore, the adversarial task should not be correlated to the main task (i.e semantic parsing here), otherwise it may harm the model's performances. Determining where these biases lay is not easy, although this is critical for the success of our method. We propose to use a domain classification adversarial task.

Given two data-sets `(X1, Y1)` and `(X2, Y2)` from different domains. The expected gains from introducing an adversarial domain classifier are proportional to how different `X1` and `X2` are (the more dissimilar the better) and proportional to how similar the label distributions `Y1` and `Y2` are (higher similarity is better). Otherwise, if `X1` and `X2` are very similar, there is no need to transfer learning from one domain to another. Under this condition, The adversarial domain classifier will not be able to recognize domains and give proper information on how to build better representations. If `Y1` and `Y2` are extremely dissimilar, to the point where `Y` cannot be predicted without explicit domain information, using adversarial learning may be harmful. In our case, prior probabilities for both frame distribution and word senses are correlated to the thematic domains. However, adversarial learning can still be useful because most of the LUs are polysemous even within a domain. Therefore, the model needs to learn word sense disambiguation through a more complex process than simply using the domain information.

Our adversarial domain classifier is an extension of (Ganin and Lempitsky, 2014) to recurrent neural networks. We start from our $biGRU$ semantic parser and on the last hidden layer, we stack another neural network that implements a domain classifier (called adversarial task). The domain classifier is connected to the $biGRU$ using a gradient reversal layer. Training consists in finding a saddle point where the semantic parser is good and the domain classifier is bad. This way, the model is optimized to be domain independent.

The architecture is shown in Figure 1. The adversarial task can be implemented using a CNN, a RNN or a FNN. In this paper we use CNN as they yield the best results on preliminary experiments.

This architecture is trained following the guidelines of (Ganin and Lempitsky, 2014). During training, the adversarial classifier is trained to predict the target class (*i.e.* to minimize the loss L_{adv}) while the main task model is trained to make the adversarial task fail (*i.e.* to minimize the loss $L_{frame} - L_{adv}$). In practice, in order to ensure stability when the parameters of the shared network are updated, the adversarial task gradient magnitude is attenuated by a factor λ as shown in (1). Here ∇L represents the gradients w.r.t the weights θ for either the frame classifier loss or the adversarial loss, θ are the model's parameters being updated, and μ is the learning rate. This λ factor increases on every epoch following (2), where p is the progress, starting at 0 and increasing linearly up to 1 at the last epoch.

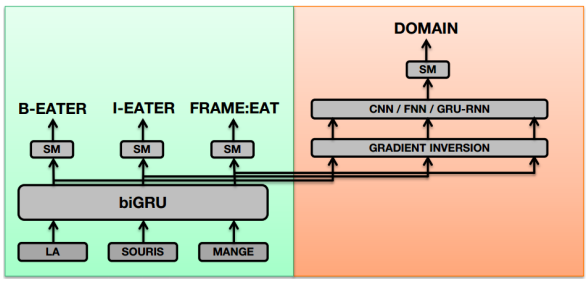

Figure 1: Adversarial Domain Classifier model

$$\theta \leftarrow \theta - \mu * (\nabla L_{frame} - \lambda \nabla L_{adv}) \quad (1)$$

$$\lambda = \frac{2}{1 + \exp(-10 \cdot p)} - 1 \quad (2)$$

3.4 Unsupervised domain inference

The originality of our approach lies in the design of an adversarial task in an unsupervised way. Our purpose is to design a general method that could easily and efficiently apply in any realistic conditions, independently of the fact that the training sentences can be linked to an *a priori* explicit domain or topic. To this end, an unsupervised clustering algorithm is used to partition the training corpus into *clusters* that are supposed to reflect topics, lexical or stylistic variation within the corpus, depending on the metric used for the clustering. In a first attempt for our experiments, we use K-means, in the `sklearn` implementation, to cluster training sentences. For clustering purpose, sentences are represented by the average of their word embedding vectors. K-means with a euclidean distance is then used to group these representations into K clusters. We use a standard *Kmeans++* initialization. The clustering process is repeated 10 times and the one that produces the

Document Source	# Sentence	# Frame	# FE
D1 Wikipedia WW1	30994	14227	32708
D2 Wikipedia ARCH	27023	9943	19892
D3 Vikidia ANC	5841	85034	3246

Table 1: Description of the CALOR-Frame corpus

minimal intra-cluster inertia is kept. Eventually, the resulting clusters are used as classes that the CNN will be trained to recognize for each corresponding training sentence. The underlying assumption is that the clustering process will capture domain-related regularities and biases that will be harnessed by the adversarial task in order to increase the model's generalization capacities.

4 Evaluation setting

To create an experimental setting that shows the effect of domain on the semantic parsing task, we run experiments on the CALOR-Frame corpus (Marzinotto et al., 2018a), which is a compilation of French encyclopedic documents, with manual FrameNet annotations (Baker et al., 1998). The CALOR-Frame corpus has been designed in the perspective of Information Extraction tasks (automatic filling of Knowledge Bases, document question answering, semantically driven reading comprehension, etc.). Due to the *partial parsing* policy, the CALOR corpus presents a much higher amount of occurrences per Frame (504) than the FrameNet 1.5 corpus (33).

We selected three subsets from the corpus, each one from a different source and/or thematic domain: Wikipedia World War 1 portal (D1), Wikipedia Archeology portal (D2) and Vikidia[1] Ancient history portal (D3). These sources allow to study the impact of both style changes (associated to differences on syntactic structures) and thematic changes (associated to lexical differences).

For the study, we focus on a set of 53 Frames that occur in all the three sub-corpora. Each partition has a different prior on the Frames and lexical units (LUs) distributions. Figure 2 shows the normalized Frame distributions for the three subsets, illustrating the thematic domain dependence. Frames such as *Attack* and *Leadership* are frequent in D1 while *Age* and *Scrutiny* are characteristic Frames for D2. The same analysis can be done using LUs, yielding similar conclusions. We also observe that D2 and D3 are more similar but the difference between these sub-corpora lie more in the syntactic structure of sentences and in the polysemic use of some LU, as will be discussed in the experiments.

[1] Vikidia is an encyclopedia for children vikidia.org

5 Results

In these experiments we evaluate the impact of our adversarial training scheme on the semantic frame parser presented in section 3.1. The parser is trained on 80% of the $D1$ and $D2$ corpora and evaluated on the remaining 20%, considered here as *in-domain* data, as well as the whole $D3$ corpus for the *out-of-domain* data. For the domain inference, we use the method presented in section 3.4. Several experiments were made, varying the number of clusters K ($K = 2$, $K = 5$ and $K = 10$). The performance obtained were very similar, the influence of K being negligible in our experiments, therefore we report here only those done with $K = 5$.

We report results using *precision/recall/f-measure* metrics at the *Target*, *Frame* and *Argument identification* levels. The errors are cumulative: to obtain a correct argument identification, we need to have both its frame and target correct. Moreover we use a hard-span metric for argument identification, meaning that both the label and the span have to be correct for an hypothesis to be considered as correct.

Results are given in figure 3 where the precision/recall curve on argument identification is given with and without adversarial training for the in-domain corpus $D1$ and the out-of domain corpus $D3$. Table 2 presents F-measure (F-max) for the 3 identification tasks (target, frame, argument) for $D1$, $D2$ and $D3$.

Figure 3 clearly illustrates the drop in performance between in-domain and out-of-domain corpora. The difference is significant, it accumulates at each step resulting on a 9 points drop in F-max for the argument identification task as shown in table 2. When applying our adversarial method during model training, we clearly increase the generalization capabilities of our model on out-of-domain data ($D3$), as the $biGRU + AC$ curve outperforms the $biGRU$ curve at every operating point in figure 3. This is confirmed on the F-max values for each level in table 2.

Interestingly our adversarial training method not only improves parser performance on out-of-domain data, but also on in-domain data, as shown for $D1$ in figure 3. This improvement is mainly

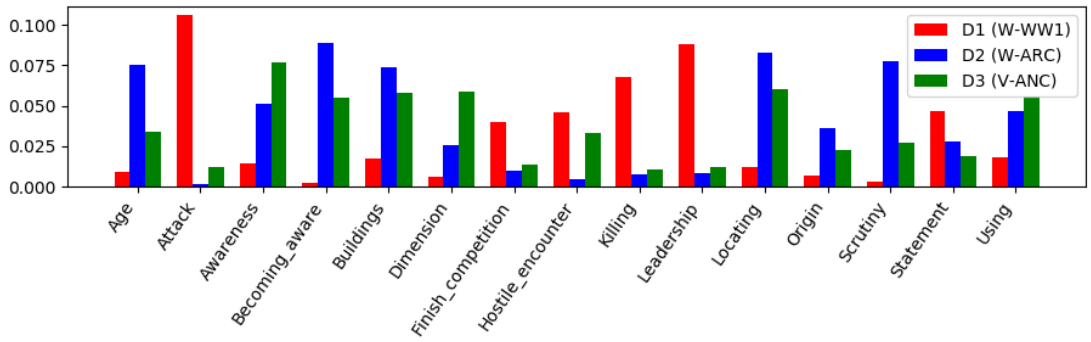

Figure 2: Most frequent frames and their normalized distribution for each partition (from left to right: `W-WW1` (`D1`), `W-ARC` (`D2`) and `V-ANC` (`D3`))

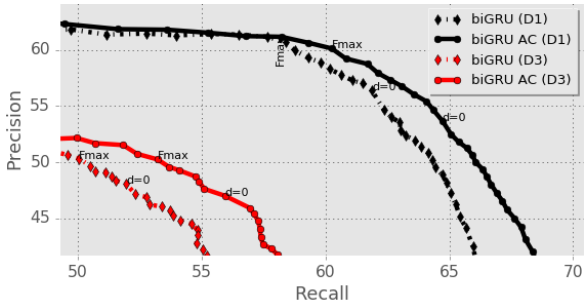

Figure 3: Precision/Recall trade-off with ($biGRU$+AC) and without ($biGRU$) adversarial training on $D1$ (in-domain) and $D3$ (out-of-domain)

due to a gain in recall, and is confirmed on the F-max measures for $D1$ and $D2$ in table 2.

We believe that, since domains are correlated with the lexicon, the adversarial objective pushes the model to rely less on lexical features and to give more importance to features such as syntax and part-of-speech. This is inline with the observation that most of the improvement comes from a higher recall. This can also explain the performance gains on in-domain data. The high dimensionality of word embeddings may have lead to some over-fitting on our initial $biGRU$ model. On the other hand, adversarial learning can act as a regularization technique that makes the model rely as much as possible on domain independent features.

In order to have a better understanding of the behavior of our method we performed two additional contrastive experiments where we used firstly *gold domain labels* instead of inferred ones, and secondly a single domain corpus for training rather than a multi-domain one.

Gold domain labels. We consider here the *true* domain labels for $D1$ and $D2$ as the classes for our adversarial classifier. Therefore only two domains are considered in the adversarial training process. Results are presented in the second part of table 2, in line $biGRU$+AC-$gold$. As we can see results are very similar to those obtained with automatic clustering ($biGRU$+AC). With an average difference of only 0.3pts (Fmax) for the argument identification task across the different domains. This confirms that our unsupervised clustering approach is as efficient as a supervised domain classification for adversarial learning. Moreover, our approach has the advantage that it can be applied in situations where no domain information is available.

Single-thematic corpus: This contrastive experiment consists in using as training a single-domain corpus. We want to check if the gains obtained on both in and out-of-domain corpora in table 2 hold when the training corpus does not explicitly contain several domains. Here, the models are trained only on the training set from `D1`. We evaluate them on `D1` (in-domain) and `D2,D3` (out-of-domain). The adversarial task is obtained by running our domain inference algorithm only on `D1` training set. Here again, we have chosen to partition training data into 5 clusters. Alternative experiments not reported in this paper using only `D2` as in-domain training data have also been performed and yielded similar conclusions. F-max values reported in table 3 are lower than those of table 2. This is expected as the training corpus considered here is much smaller (only $D1$), however performances follow the same trend: some gains are obtained for all 3 levels both for in and out-of-domain corpora. This is a very interesting result as it shows that there is no need for an explicitly multi-thematic training corpus in order to run the adversarial training and to obtain some

	Target Identification			Frame Identification			Argument Identification		
	in-domain		out-of-domain	in-domain		out-of-domain	in-domain		out-of-domain
	D1	D2	D3	D1	D2	D3	D1	D2	D3
biGRU	97.1	97.5	94.2	93.4	95.4	91.0	59.2	56.3	50.2
biGRU+AC	97.3	98.7	94.7	94.2	95.9	91.9	60.0	57.0	51.7
biGRU+AC-gold	97.7	98.2	94.9	94.9	95.9	92.0	60.1	56.7	51.3

Table 2: F-measure (Fmax) on target, frame and argument identification with ($biGRU$+AC) and without ($biGRU$) adversarial training. Clustering with exact domain labels is given in line $biGRU$+AC-$gold$

gains in terms of model generalization.

5.1 Error Analysis

5.1.1 Target and Frame Identification

When looking carefully at the generalization capabilities of the initial model, we observed that the frames with the highest performance drops on D3 are those associated to LUs that are polysemous in a general context, but are unambiguous given a thematic domain. For example, *installer (to install)* triggers the frame *Installing* in both D1 and D2, but triggers *Colonization* in D3. Sometimes the confusion comes from changes in the writing style. For example *arriver (to arrive)* means *Arriving* in both D1 and D2, but in D3 it is used as a modal verb (*arriver à* meaning *to be able to*) triggering no frame.Under these circumstances, a model trained on a single domain underestimates the complexity of the frame identification task, mapping the LU to the frame without further analysis of its sense.

When we apply $biGRU + AC$, the gains observed on Frame Identification are not constant across LUs. To analyze the impact of the adversarial approach, we compare for each LU the distribution across clusters of the sentences containing the given LU. This is done separately for D1 and D3 (for D3, sentences are projected into the clusters by choosing the less distant cluster centroid). In table 4, we present the LUs for which the cluster distribution on D1 and D3 are the most dissimilar. These are also the LUs that are the most positively affected by the adversarial strategy.

This means that whenever a LU has similar distribution of context words across the different domains this context information is already exploited by the system to perform frame disambiguation. On the other hand, when the context words of a LU depend on the domain, the model can take advantage of adversarial learning to build a higher level representation that abstracts as much as possible from the lexical variations of the words surrounding the LU.

5.1.2 Argument identification

In this section, we focus on the Frame Argument (or FE for Frame Element) Identification level, and propose contrastive experiments following the complexity factors analysis proposed by (Marzinotto et al., 2018b). In this study, the authors have shown that FE identification is performing better for verbal LU triggers than for nominal ones, and for triggers that are at the root of the syntactic dependency tree. We want to see here how these complexity factors are affected by the adversarial learning strategy. Additionally, we want to see if the system behaves equivalently over core and non-core Frame Elements. Actually, in the usual evaluation Framework of the Framenet 1.5 shared task, non-core FEs are assigned a 0.5 weight for the F-measure computation, reducing the impact of errors on non-core FEs. In this paper, all FEs are rated equally, but here we separate them to observe their behaviour. As we can see in table 5, adversarial training consistently improves the FE identification results, in all conditions. Moreover, bigger gains are observed for the *difficult* conditions.

6 Generalization to PropBank Parsing

We further show that this adversarial learning technique can be used on other semantic frameworks such as Propbank. In PropBank Semantic Role Labeling, CoNLL-2005 uses Wall Street Journal (WSJ) for training and two test corpora. The in-domain (ID) test set is derived from WSJ and the out-of-domain (OOD) test set contains 'general fiction' from the Brown corpus. In published works, there is always an important gap in performances between ID and OOD. Several studies have tried to develop models with better generalization capacities (Yang et al., 2015), (FitzGerald et al., 2015). In recent works, PropBank SRL systems have evolved and span classifier approaches have been replaced by current state of the art sequence tagging models that use recurrent neural networks (He et al., 2017) and neural atten-

	Target Identification			Frame Identification			Argument Identification		
	in-domain	out-of-domain		in-domain	out-of-domain		in-domain	out-of-domain	
	D1	D2	D3	D1	D2	D3	D1	D2	D3
biGRU	97.6	95.5	93.3	93.8	93.4	90.9	58.2	46.1	43.6
biGRU+AC	97.6	95.6	94.3	95.3	94.5	91.2	60.0	47.1	45.2

Table 3: Frame semantic parsing performances (Fmax). Models trained on D1. Adversarial learning with inferred domains $biGRU + AC$.

LU	*biGRU*		*biGRU* + *AC*	
	D1	D3	D1	D3
arriver	88.2	63.4	91.8	70.0
écrire	96.5	73.3	97.8	88.4
expression	49.8	66.6	56.5	92.4

Table 4: Frame Identification score for LUs with the highest variation in cluster distribution

D3	*biGRU*	*biGRU* + *AC*
overall	50.2	51.7 (+3%)
core FE	56.5	57.0 (+0.9%)
non-core FE	48.9	50.4 (+3.1%)
verbal trigger	53.4	54.9 (+2.8%)
nominal trigger	34.6	39.0 (+12.7%)
root trigger	59.5	61.3 (+3.0%)
non-root trigger	45.4	47.2 (+4.0%)

Table 5: Frame Element Identification results according to complexity factors on D3 (Fmax)

tion (Tan et al., 2017; Strubell et al., 2018). However, these parsers still suffer performances drops of up to 12 points in F-measure on OOD with respect to ID. For this experiment we have applied the same adversarial approach over a state-of-the-art Propbank parser (He et al., 2017) using a single straight classifier model. As there is no explicit domain partitions in the training corpus, we apply our inferred domain adversarial task approach, running the clustering algorithm with 5 clusters. We were not able to reproduce the same results as the one published in the paper, we hence provide the results obtained running the downloaded system in our lab. Similarly to the previous FrameNet parsing model, we vary a threshold on the output probabilities of Semantic Roles in order to optimize the F-measure and we provide F-max values, computed using the official evaluation script. We observe in table 6 that the adversarial approach outperforms the original system on both the ID and OOD tests sets.

7 Conclusion

We have presented a study on improving the robustness of a frame semantic parser using adversarial learning. Results obtained on a multi-domain publicly available benchmark, called *CALOR-Frame*, showed that domain adversarial training can effectively be used to improve the generalization capacities of the tagging models, even without prior information about domains in the training corpus. We showed that our technique can be applied to other semantic models, by implementing it into a state-of-the-art PropBank parser and showing some consistent gains. This positive result suggests that our approach could apply successfully to more NLP tasks.

	ID WSJ	OOD BROWN
(He et al., 2017)	82.4	71.7
(He et al., 2017)+ *AC*	83.0	72.3

Table 6: SRL performance (Fmax) on CoNLL-2005, based on (He et al., 2017) model

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *ArXiv e-prints*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational linguistics*, 40(1):9–56.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *EMNLP*.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Semantic role

labeling with neural network factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 960–970.

Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*.

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

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(59):1–35.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain framenet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 471–482.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *ACL (1)*, pages 1448–1458.

Richard Johansson and Pierre Nugues. 2008. The effect of syntactic representation on semantic role labeling. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 393–400. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *CoRR*, abs/1704.05742.

Gabriel Marzinotto, Jeremy Auguste, Frédéric Béchet, Géraldine Damnati, and Alexis Nasr. 2018a. Semantic Frame Parsing for Information Extraction : the CALOR corpus. In *LREC 2018*, Miyazaki, Japan.

Gabriel Marzinotto, Frédéric Béchet, Géraldine Damnati, and Alexis Nasr. 2018b. Sources of Complexity in Semantic Frame Parsing for Information Extraction. In *International FrameNet Workshop 2018*, Miyazaki, Japan.

Rebecca J Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012. The masc word sense sentence corpus. In *Proceedings of LREC*.

Darsh J. Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial domain adaptation for duplicate question detection. In *EMNLP*.

Anders Søgaard, Barbara Plank, and Hector Martinez Alonso. 2015. Using frame semantics for knowledge extraction from twitter. In *AAAI*, pages 2447–2452.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *NIPS 2015*, pages 2377–2385, Montral, Qubec, Canada.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. *arXiv preprint arXiv:1804.08199*.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2017. Deep semantic role labeling with self-attention. *CoRR*, abs/1712.01586.

Bishan Yang and Tom Mitchell. 2017. A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256. Association for Computational Linguistics.

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.

Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen. 2018. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 682–690. ACM.