*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Adapting a FrameNet Semantic Parser for Spoken Language Understanding using Adversarial Learning

*Gabriel Marzinotto*[1,2], *Géraldine Damnati*[1], *Frédéric Béchet*[2]

[1]Orange Labs/ Lannion France
[2]Aix Marseille Université, Univ. Toulon, CNRS, LIS, Marseille, France

gabriel.marzinotto@orange.com, geraldine.damnati@orange.com, frederic.bechet@lis-lab.fr

## Abstract

This paper presents a new semantic frame parsing model, based on Berkeley FrameNet, adapted to process spoken documents in order to perform information extraction from broadcast contents. Building upon previous work that had shown the effectiveness of adversarial learning for domain generalization in the context of semantic parsing of encyclopedic written documents, we propose to extend this approach to elocutionary style generalization. The underlying question throughout this study is whether adversarial learning can be used to combine data from different sources and train models on a higher level of abstraction in order to increase their robustness to lexical and stylistic variations as well as automatic speech recognition errors. The proposed strategy is evaluated on a French corpus of encyclopedic written documents and a smaller corpus of radio podcast transcriptions, both annotated with a FrameNet paradigm. We show that adversarial learning increases all models generalization capabilities both on manual and automatic speech transcription as well as on encyclopedic data.

**Index Terms**: semantic parsing, FrameNet, Adversarial learning, speech recognition, spoken language, understanding

## 1. Introduction

Semantic parsing is essential for Spoken Language Understanding (SLU). Currently, the most common strategy to extract semantic information consists on a pipeline processing composed of an automatic speech recognition (ASR) system and then a semantic parser on the ASR outputs. However, most semantic parsers are built to process written language and are not robust to speech processing and even less to ASR errors.

Even when the speech corpora to train semantic parsers is available, systems trained on perfect transcriptions tend to degrade processing ASR. A common strategy to compensate the system consists on simulating ASR errors in the textual training corpus [1]. Additionally, ASR systems are generally tuned measuring word error rate (WER) on a validation corpus, but this metric is not optimal for the subsequent SLU task (semantic parsing, NER, etc.). To compensate this, some specialized metrics to tune ASR have been proposed [2]. However, the number of SLU task applied on the ASR output may be large and considering dedicated metrics for each one of them is not feasible.

Biases associated to writing, speech and ASR are a major problem in SLU task. Models learn these biases as useful information and experience a significant performance drop whenever they are applied on data from a different source. 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 [3]. In the NLP community, this method has been mostly used for cross-lingual transfer learning [4] and more recently in monolingual setups in order to alleviate domain bias in semantic parsers [5].

In this paper we implement a FrameNet [6] semantic parser, originally designed and trained to process encyclopedic texts, for semantic analysis of spoken documents (radio podcasts) in an Information Extraction perspective. We address the issue of generalization capacities of the Semantic Frame Parser when processing speech. We show that adversarial learning can be used to combine different sources of data to build robust representations and improve the generalization capacities of semantic parsers on speech and ASR data. We propose an adversarial framework based on a domain classification task that we use as a regularization technique on a state-of-the-art semantic parser.

## 2. Related Work

Structured semantic representations for speech processing have been mainly explored in the domain of conversational speech processing. Experiments around the adaptability of Framenet semantic parsing to the context of dialogues are reported on the Communicator2000 corpus [7] and in the LUNA Italian dialogue corpus [8], showing their viability for labeling conversational speech. French conversational speech have also been explored on the DECODA corpus with adaptation of parsers to highly spontaneous speech with a specific adaptation process towards disfluencies and ASR errors [9] and the introduction of multi-task learning to jointly handle syntactic and semantic analysis [10]. Other semantic models have also been experimented for SLU, such as Abstract Meaning Representation (AMR) in [11] where the authors show that syntactic and semantic structured representations can help guiding attention based models neural networks. In a broader perspective, few works have been dedicated to semantic parsing of spoken contents for Information Extraction.

Concerning the crucial issue of model robustness, several strategies have been studied in order to improve generalization in supervised learning. A popular approach that emerged in image processing [12] consists in training models on a double objective composed of a task-specific classifier and an adversarial domain classifier. The latter is called adversarial because it is connected to the task-specific classifier through 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 that this guarantees the resulting model to be domain independent [13]. In Natural Language Processing tasks, this approach has been used to build cross-lingual models, doing transfer learning from English to low resource languages for POS tagging [4] and sentiment analysis [14], by using language classifiers with an adversarial objective to train task-specific but language agnostic representations. This technique is not only useful in cross-lingual transfer prob-

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2732

lems, as it has been used to improve generalization in a document classification[15], Q&A systems [16], duplicate question detection [17] and semantic parsing [5] in a monolingual setup.

In Frame Semantic Parsing, data is scarce and evaluation campaigns rarely study the generalization capacities on out-of-domain test data. Recently, the YAGS corpus was published along with the first in depth study of the domain adaptation problem in Semantic Frame Parsing[18]. They show that the main bottleneck in domain adaptation is at the Frame Identification step and propose a more robust classifier for this task, using predicate and context embeddings to perform Frame Identification. This approach is suitable for cascade systems such as SEMAFOR [19], [20]. In this paper we study the generalization issue within the framework of a sequence tagging semantic frame parser that performs frame selection and argument classification in one step. And we will show that adversarial domain adaptation paradigms can be transposed into speech adaptation.

# 3. Semantic parsing model with an adversarial training scheme

## 3.1. Semantic parsing model: **biGRU**

We implement our semantic frame parser using a sequence tagger that performs frame selection and argument classification in one step. Our model is a 4 layer bi-directional GRU tagger ($biGRU$). The advantage of this architecture is its flexibility as it can be applied on both SRL [21] and Frame Parsing [22, 23]. This model relies on a rich set of features including pretrained word embedding, syntactic, morphological and surface features. More details on the architecture can be found in [24].

## 3.2. Sequence encoding/decoding

We use a BIO label encoding in all our experiments. On inference, we apply the coherence filter [5] that selects the most probable Frame for the LU and filters all incompatible FE. To ensure that output sequences respect the BIO constrains we implement an A$^*$ decoding strategy as the one proposed by [21].

Finally, we introduce a hyper-parameter $\delta \in (-1; 1)$ that is added to the output probability of the *null* label $P(y_t = O)$ at each time-step. For each word, 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. This technique allows to tune the performance of our models and study their precision/recall (P/R) trade-off. The optimal value for $\delta$ is selected on a validation set. In this paper, we either provide the P/R curve or report scores for the $Fmax$ setting.

## 3.3. Adversarial Domain Classifier

Adversarial domain training was initially proposed in [3] and adapted for semantic parsing in [5]. We start from our *biGRU* semantic parser and on the last hidden layer, we stack a CNN with a decision layer to implement 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 bad. This optimizes the model to be domain independent. The architecture diagram is shown in Figure 1.

More precisely, the adversarial classifier is trained to predict domains (*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 during training, we follow the guidelines from [3]. The adversarial task gradient magnitude is attenuated by a factor $\lambda$ as shown in equation (1). Here $\nabla L$ represents the gradients w.r.t the weights $\theta$ for either the frame classifier loss or the adversarial loss, $\theta$ are the model's parameters being updated, and $\mu$ is the learning rate. This $\lambda$ factor increases on every epoch following equation (2), where $p$ is the progress, starting at 0 and increasing linearly up to 1 at the last epoch.

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

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

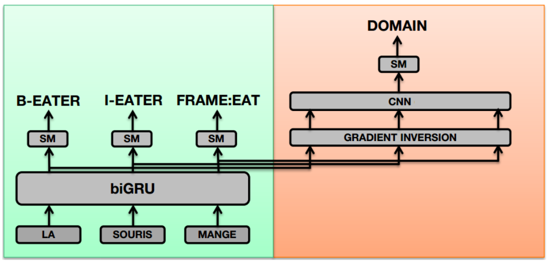

Figure 1: *Adversarial Domain Classifier model*

Table 1: *Statistics of both the* `CALOR` *corpus of encyclopedic documents and the* `PODCAST` *corpus of transcriptions*

| **Corpus** | **# Sentence** | **# Frame** | **# FE** |
|---|---|---|---|
| CALOR | 67381 | 26725 | 57688 |
| PODCAST | 4233 | 2298 | 5474 |

# 4. Evaluation setting

Our experimental setting allows to study the differences between encyclopedic texts and speech transcriptions on the semantic parsing task. To do this, we run experiments on two corpora. First, the `CALOR` corpus [25], which is a compilation of French encyclopedic documents (from Wikipedia, Vikidia and ClioTexte) with manual FrameNet [26] annotations. Second, the `PODCAST` corpus is a compilation of radio podcasts from *Les P'tits Bateaux* a show from France Inter broadcast. Our corpus gathers 210 sequences of a children asking a general knowledge question through a phone call followed by an answer in the form of a conversation between an expert and a journalist, each sequence is 3 to 4 minutes long. The corpus has been manually transcribed and annotated in FrameNet semantics. Statistics about the corpora are presented in Table 1.

Even though both corpora are related to general knowledge, they are very different in terms of style. `CALOR` is composed of well written encyclopedic text dealing with three subjects (WW1, Archaeology and Ancient History). On the other hand, `PODCAST` contains transcriptions of a radio show addressed to children using a simpler discourse but dealing with a broader set of general knowledge topics. These corpora have been designed in the perspective of targeted Information Extraction tasks. Due to this, we used a *partial parsing* policy where only 53 Frames have been annotated on the whole corpus. This allows to rapidly annotate large corpora and yields a much higher amount of occurrences per Frame (i.e. 504 in `CALOR` vs. 33 in `FrameNet`).

Each corpus has a different prior on the Frames and LU distributions. Figure 2 shows the normalized Frame distributions for both sets, illustrating the domain dependence. Frames such

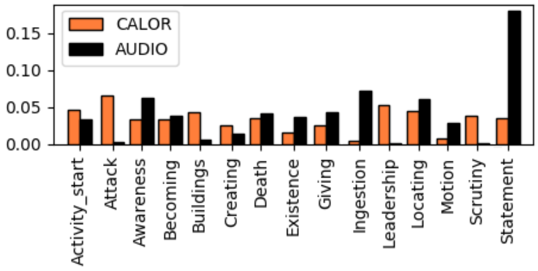

Figure 2: *Most frequent frames and their normalized distribution for each partition (*`CALOR` *and* `PODCAST`*)*

as *Attack* and *Leadership* are frequent in `CALOR` while *Statement* and *Ingestion* are common Frames in `PODCAST`. When we analyze the Lexical Units (LUs) distribution and their associated Frames, we observe an imbalance between corpora. In `PODCAST` we find that 38% of LUs do not trigger any Frame. This is a very large number compared to the 13% in the `CALOR` corpus. The reason for this is that in `PODCAST` many LUs appear as part of idioms or oral expressions that do not convey semantic Frames. Examples of these expressions are: *c'est-à-dire (that is), disons (let's say), comment dire (how to say).* Moreover, speakers in `PODCAST` tend to address children in a simplified language, using very common words which are often polysemous, this makes LUs in `PODCAST` more ambiguous. Examples for this are: *ça donne* (*"results in"* instead of *"it gives"*), *arriver à* (*"being able to"* as opposed to *"arriving"*), *on se demande* (*"we ask ourselves"* and not *"to request"*). As a consequence, processing "simplified" language for children is not "simpler" from a semantic parsing perspective.

To generate an ASR corpus, we process our `PODCAST` samples using the Cobalt Speech Recognition system developed at Orange Labs. It is a Kaldi-based decoder using a time-delay neural network based acoustic model [27] trained on more than 2000 h of clean and noisy speech, with a 1.7 million word lexicon, and a 5-gram language model trained on 3 billion words. We further align the ASR outputs with the manual transcriptions in order to project the FrameNet annotations into the ASR corpus. The evaluation of our ASR system on the `PODCAST` corpus yields a WER of 14.2%, with a large variation between children speech in telephone recording conditions (41% WER) and journalist and expert studio conversation (13% WER).

## 5. Results

### 5.1. Without Adaptation

In these experiments we first evaluate the performances of our *biGRU* semantic Frame parser for different training configurations in order to better understand the main difficulties of the transfer learning for SLU. For each corpus, we split data into 60% for training 20% for validation and 20% for test. We report results using *F-measure* metrics at the *Frame* and *Argument Identification* (respectively FI and AI) levels. Errors are cumulative: in order to obtain a correct argument identification, we need to identify the correct Frame. Since we test our model on the ASR outputs, we evaluate using a soft-span metric for AI, meaning that for an argument hypothesis to be correct the label has to match the reference but the span may not be exactly the same (an overlap with the matching reference is required). This metric does not oversimplify the argument detection task since the median length of an argument in `PODCAST` corpus is 2 words (slightly lower than for the original `CALOR` corpus where the median length is 3 words).

Table 2: *F-measure (Fmax) on Frame and Argument Identification using (biGRU) and different training datasets.*

| | Frame Ident | | Argument Ident | |
|---|---|---|---|---|
| | PODCAST | | PODCAST | |
| | `GOLD` | `ASR` | `GOLD` | `ASR` |
| CALOR | 79.6 | 79.1 | 52.3 | 51.4 |
| PODCAST$_{GOLD}$ | 86.3 | 85.0 | 55.7 | 51.0 |
| PODCAST$_{ASR}$ | 86.3 | 86.3 | 56.0 | 53.6 |
| CALOR+PODCAST$_{GOLD}$ | 84.8 | 84.5 | 61.3 | 56.7 |
| CALOR+PODCAST$_{ASR}$ | 86.0 | 85.6 | 59.9 | 57.8 |
| PODCAST$_{BOTH}$ | 87.3 | 86.9 | 58.4 | 55.9 |
| CALOR+PODCAST$_{BOTH}$ | 87.4 | 85.0 | 65.4 | 59.9 |

Results for this experience are reported in table 2. We observe that simply learning a model on the `CALOR` textual corpus and applying it on the speech corpus yields very low performances. In `PODCAST`, for the Frame Identification phase, the parser fails to analyse idioms and oral expressions where the LU should not trigger Frames. Training a parser on the small `PODCAST-GOLD` or `PODCAST-ASR` corpus fixes this problem increasing the FI score from 79.6% to 86.3% on manual transcripts and from 79.1% to 86.3% on automatic transcripts. However, the `PODCAST` corpus is too small to train a model for the Argument Identification task. For this task the only models that give acceptable performances are those that were trained using data from both `CALOR` and `PODCAST`.

Even though our ASR system has a WER or 15% when recognizing LUs, most of these errors were confusions with other inflected forms of the LU, which do not affect FI. For this reason, performances in this task are almost the same for `ASR` and `GOLD` transcripts. We cannot say the same thing from the Argument Identification task. In our best training configuration (`CALOR+PODCAST(BOTH)`), performances on `ASR` are 5.5 points bellow the performances on `GOLD`. Even though WER inside the Arguments is lower (only 12%), these errors deeply affect the semantic parser. There are two reasons for this: the first one is that insertions and deletions appear mostly on short words, these words are often pronouns, prepositions and articles. When a pronoun is missing a whole Argument can be lost, similarly, articles and preposition are strong indicators of the presence of a specific argument. The second reason is that semantic roles are strongly correlated to syntax and the ASR errors easily introduce syntax errors, for example confusing *"a"(to have)* with *"à"(to)* or *"est"(to be)* with *"et"(and)* completely change the structure and the meaning of a sentence.

### 5.2. With Adaptation

In this experiment we compare our initial *biGRU* model with a model trained using adversarial domain adaptation. We ran experiments using some of the different configurations of the training corpus presented in Table 2. However, since the adaptation technique behaves similarly in all these configurations, we present the results for the most common setup, which consists in training a model on all data sources (`CALOR` and `PODCAST(BOTH)`). Under this setup we trained our adversarial model (*biGRU+adv*) with a "domain" classifier that distinguishes between two sources, determining if samples come either from `CALOR` or `PODCAST(BOTH)`. In earlier experiments we tested different domain classification tasks varying the classes, and using unsupervised inferred domains, but the simple 2 domain task yields the best results on `PODCAST-ASR`.

Results are given in figure 3 where the precision/recall

Table 3: *F-measure (Fmax) on FI and AI with (biGRU+adv) and without (biGRU) adversarial training.*

| | Frame Ident. | | Argument Ident. | |
|---|---|---|---|---|
| | PODCAST | | PODCAST | |
| | GOLD | ASR | GOLD | ASR |
| *biGRU* | 87.4 | 85.0 | 65.4 | 59.9 |
| *biGRU+adv* | 89.3 | 88.3 | 65.9 | 62.4 |

curve on argument identification is obtained by varying a threshold over final argument detection in two conditions: with (*biGRU+adv*) and without (*biGRU*) adversarial training for the PODCAST-GOLD and PODCAST-ASR tests sets. For the sake of comparison, the results are also provided over the initial CALOR test set, showing that these results are not harmed by the adaptation process. Table 3 presents F-measure (F-max) for both Frame and Argument Identification tasks on each test corpus. When applying our adversarial method, we clearly increase the generalization capabilities of our model on both test sets, as the *biGRU+adv* curve outperforms the *biGRU* curve at every operating point in figure 3. This is confirmed on the F-max values in table 3. The corpus with the highest improvements is the ASR corpus, with +2.5 points. This shows that our approach can help building higher level representation that are more independent from data source (written or spoken language) especially when dealing with transcription errors.

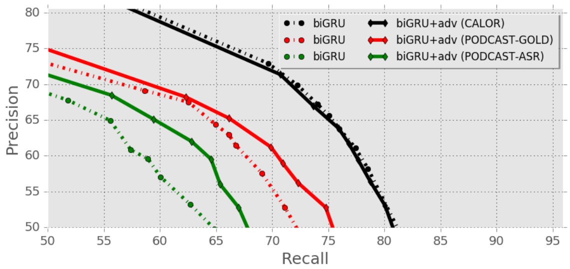

Figure 3: *Precision/Recall trade-off with (biGRU+adv) and without (biGRU) adversarial training for the different test sets*

### 5.3. Error Analysis

We observed that most of the errors in **Frame Identification** are associated to LUs that are polysemous in a general context, but are unambiguous given a thematic domain. For example, *dire (to say)* triggers the frame *Statement* in most of the CALOR corpus, but in the PODCAST corpus most of the time it does not trigger any Frame as it is used as part of an oral expression. Under these circumstances, the model underestimates the FI task and assigns a Frame to a LU without doing a proper analysis of its meaning using appropriated features. In Table 4 we show the FI score for the LUs suffering important changes in their meaning distribution across corpora. These changes are mostly due to LUs used in oral expression, rather than to a thematic change. Adversarial training is beneficial for these LUs, specially for the ASR. This result supports the idea that adaptation can be useful even if the spoken and written thematic domains are similar.

We focus now on the **Argument Identification** (AI) level. We try to identify precisely in which situations the adversarial learning strategy improves AI. To do so, we focus on the PODCAST-ASR test data and we evaluate our strategy on the scope of some complexity factors [24]. As we can see in table 5, adversarial training largely improves the identification of non-core Frame Elements (FE) such as *Place* or Time, while having a moderate impact on core FEs (specific arguments with typical agent and patient semantic roles). This is not surprising since non-core FEs are often shared across several Frames despite having non trivial lexical variations for each Frame/domain. Consider the non-core FE *Place* whose content can vary from names of countries and cities to common nouns (*"in the park"*) and adverbs (*"there"*). Under these circumstances adversarial learning can successfully build a higher level representation of the FE. As for the other complexity factors, bigger gains are observed for the *difficult* conditions (i.e. nominal triggers and long sentences). On the other hand, adversarial learning slightly harmed performances on the short sentences.

Table 4: *Frame Identification score for LUs with the highest variation of sense distribution across corpora*

| LU | GOLD | | ASR | |
|---|---|---|---|---|
| | *biGRU* | *+adv* | *biGRU* | *+adv* |
| dire | 80.9 | 81.8 | 77.1 | 85.4 |
| donner | 78.8 | 84.8 | 77.4 | 87.1 |
| demander | 75.8 | 75.8 | 56.0 | 72.0 |

Table 5: *Argument Identification results on the PODCAST-ASR corpus according to complexity factors (Fmax)*

| D3 | *biGRU* | *biGRU+adv* |
|---|---|---|
| overall | 59.9 | 62.4 (+2.5%) |
| core FE | 63.6 | 65.5 (+1.9%) |
| non-core FE | 37.8 | 43.0 (+5.2%) |
| verbal trigger | 61.2 | 63.8 (+2.6%) |
| nominal trigger | 36.5 | 41.0 (+4.5%) |
| short sentences | 65.3 | 64.8 (-0.5%) |
| long sentences | 58.8 | 61.9 (+3.1%) |

Table 6: *Frame Element Identification results (Fmax) on the PODCAST-ASR corpus under different WER*

| WER | #frames | *biGRU* | *biGRU+adv* |
|---|---|---|---|
| overall | 489 | 59.9 | 62.4 (%) |
| $0 \leq WER < 5$ | 50 | 70.1 | 72.5 (+2.4%) |
| $5 \leq WER < 10$ | 167 | 63.6 | 65.4 (+1.8%) |
| $10 \leq WER < 15$ | 126 | 61.1 | 63.8 (+2.7%) |
| $15 \leq WER < 20$ | 81 | 55.0 | 60.7 (+5.7%) |
| $20 \leq WER$ | 65 | 51.8 | 58.0 (+6.2%) |

Finally, intuition says a higher WER translates into lower performance for our Frame parser. To corroborate this, Table 6 shows the performance of our models on different subsets of PODCAST-ASR presenting different WER. We observe that indeed higher WER yields lower performance for the semantic parsing task. Table 6 also shows that our adversarial learning strategy is more beneficial for transcriptions with a high WER.

## 6. Conclusions

We have presented a study on the robustness of Frame semantic parsing under changes in the elocutionary style. We presented an adaptation technique based on adversarial learning. This technique combines data from different sources (speech and text) to train more robust models that perform semantic parsing on a higher level of abstraction. Results showed that domain adversarial training can be effectively used to improve the generalization capacities of our semantic frame parser on spoken documents. This positive result suggests that our approach could apply successfully to more Spoken Language Understanding tasks.

# 7. References

[1] E. Simonnet, S. Ghannay, N. Camelin, and Y. Estève, “Simulating ASR errors for training SLU systems,” in *LREC 2018*, Miyazaki, Japan, May 2018. [Online]. Available: https://hal-univ-lemans.archives-ouvertes.fr/hal-01715923

[2] M. A. B. Jannet, O. Galibert, M. Adda-Decker, and S. Rosset, “Investigating the effect of asr tuning on named entity recognition,” in *Proc. Interspeech 2017*, 2017, pp. 2486–2490. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1482

[3] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” *arXiv preprint arXiv:1409.7495*, 2014.

[4] J.-K. Kim, Y.-B. Kim, R. Sarikaya, and E. Fosler-Lussier, “Cross-lingual transfer learning for pos tagging without cross-lingual resources,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 2832–2838. [Online]. Available: http://aclweb.org/anthology/D17-1302

[5] G. Marzinotto, G. Damnati, F. Béchet, and B. Favre, “Robust semantic parsing with adversarial learning for domain generalization,” in *Proc. of NAACL*, 2019.

[6] C. J. Fillmore, C. F. Baker, and H. Sato, “Framenet as a ”net”.” in *LREC*, 2004.

[7] S. Stoyanchev, A. Stent, and S. Bangalore, “Evaluation of semantic dependency labeling across domains,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI’16. AAAI Press, 2016, pp. 2814–2820. [Online]. Available: http://dl.acm.org/citation.cfm?id=3016100.3016295

[8] B. Coppola, A. Moschitti, and G. Riccardi, “Shallow semantic parsing for spoken language understanding,” in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, 2009, pp. 85–88.

[9] F. Bechet, A. Nasr, and B. Favre, “Adapting dependency parsing to spontaneous speech for open domain spoken language understanding,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[10] J. Tafforeau, F. Bechet, T. Artières, and B. Favre, “Joint syntactic and semantic analysis with a multitask deep learning framework for spoken language understanding.” in *Interspeech*, 2016, pp. 3260–3264.

[11] Y.-N. Chen, D. Hakanni-Tür, G. Tur, A. Celikyilmaz, J. Guo, and L. Deng, “Syntax or semantics? knowledge-guided joint semantic frame parsing,” in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 348–355.

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-239.html

[13] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International Conference on Machine Learning*, 2015, pp. 1180–1189.

[14] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, “Adversarial deep averaging networks for cross-lingual sentiment classification,” *ArXiv e-prints*, Jun. 2016. [Online]. Available: https://arxiv.org/abs/1606.01614

[15] P. Liu, X. Qiu, and X. Huang, “Adversarial multi-task learning for text classification,” *CoRR*, vol. abs/1704.05742, 2017. [Online]. Available: http://arxiv.org/abs/1704.05742

[16] J. Yu, M. Qiu, J. Jiang, J. Huang, S. Song, W. Chu, and H. Chen, “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*. ACM, 2018, pp. 682–690.

[17] D. J. Shah, T. Lei, A. Moschitti, S. Romeo, and P. Nakov, “Adversarial domain adaptation for duplicate question detection,” in *EMNLP*, 2018.

[18] S. Hartmann, I. Kuznetsov, T. Martin, and I. Gurevych, “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*, vol. 1, 2017, pp. 471–482.

[19] D. Das, D. Chen, A. F. Martins, N. Schneider, and N. A. Smith, “Frame-semantic parsing,” *Computational linguistics*, vol. 40, no. 1, pp. 9–56, 2014.

[20] K. M. Hermann, D. Das, J. Weston, and K. Ganchev, “Semantic frame identification with distributed word representations.” in *ACL (1)*, 2014, pp. 1448–1458.

[21] L. He, K. Lee, M. Lewis, and L. Zettlemoyer, “Deep semantic role labeling: What works and what’s next,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017.

[22] B. Yang and T. Mitchell, “A joint sequential and relational model for frame-semantic parsing,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 1247–1256. [Online]. Available: http://aclweb.org/anthology/D17-1128

[23] A. Celikyilmaz, , J. Gao, , and Y.-Y. Wang, “Multi-domain joint semantic frame parsing using bi-directional rnn-lstm.” ISCA, June 2016. [Online]. Available: https://www.microsoft.com/en-us/research/publication/multijoint/

[24] G. Marzinotto, F. Béchet, G. Damnati, and A. Nasr, “Sources of Complexity in Semantic Frame Parsing for Information Extraction,” in *International FrameNet Workshop 2018*, Miyazaki, Japan, May 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01731385

[25] G. Marzinotto, J. Auguste, F. Béchet, G. Damnati, and A. Nasr, “Semantic Frame Parsing for Information Extraction : the CALOR corpus,” in *LREC 2018*, Miyazaki, Japan, May 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01959187

[26] C. F. Baker, C. J. Fillmore, and J. B. Lowe, “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*. Association for Computational Linguistics, 1998, pp. 86–90.

[27] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Purely sequence-trained neural networks for asr based on lattice-free mmi.” in *Interspeech*, 2016, pp. 2751–2755.