Mitigating bias in deep nets with knowledge bases

Mitigating Bias in Deep Nets with Knowledge Bases : the Case of Natural Language Understanding for Robots

Martino Mensio*, Emanuele Bastianelli⋆, Ilaria Tiddi† and Giuseppe Rizzo‡
*Knowledge Media Institute, The Open University, UK, martino.mensio@open.ac.uk
⋆The Interaction Lab, Heriot-Watt University, UK, emanuele.bastianelli@hw.ac.uk
†Department of Computer Science, Vrije Universiteit Amsterdam, NL, i.tiddi@vu.nl
‡LINKS Foundation, Italy, giuseppe.rizzo@linksfoundation.com

Abstract

In this paper, we tackle the problem of lack of understandability of deep learning systems by integrating heterogeneous knowledge sources, and in the specific we present how we used FrameNet to guarantee the correct learning for an LSTM-based semantic parser in the task of Spoken Language Understanding for robots. The problem of the explainability of Artificial Intelligence (AI) systems, i.e. their ability to explain decisions to both experts and end users, has attracted growing attention in the latest years, affecting their credibility and trustworthiness. Trusting these systems is fundamental in the context of AI-based robotic companions interacting in natural language, as the users' acceptance of the robot also relies on the ability to explain the reasons behind its actions. Following similar approaches, we first use the values of the neural attention layers employed in the semantic parser as a clue to analyze and interpret the model's behavior and reveal the intrinsic bias induced by the training data. We then show how the integration of knowledge from external resources such as FrameNet can help minimizing, or mitigating, such bias, and consequently guarantee the model to provide the correct interpretations. Our preliminary, but promising results suggest that (i) attention layers can improve the model understandability; (ii) the integration of different knowledge bases can help overcoming the limitations of machine learning models; and (iii) an approach combining the strengths of both knowledge engineering and machine learning can foster the development of more transparent, understandable intelligent systems.

Introduction

With the dramatic success of new machine learning techniques relying on deep architectures, the number of Artificial Intelligence (AI)-based systems has rapidly increased. Events such as the Cambridge Analytica scandal and the disruptions of the 2016 US elections have brought researchers and practitioners to question the explainability of these systems, i.e. their ability to explain decisions to both experts and end-users, resulting in a number of initiatives to improve their understandability and trustworthiness (cfr. DARPA's eXplainable AI program[1]; the "right to explanation" requested by the European General Data Protection Regulation; and the "Ethics guidelines for Trustworthy AI" published by the European Union in December 2018[2]). In the context of robotic companions interacting in natural language using AI techniques, where our research is placed, trust and transparency are fundamental aspects, as the users' acceptance of the robot assistants will be also based on their ability to explain the reasons behind their actions, if and when required.

Let us take the example of a robot understanding spoken commands given by a human, e.g. "*take the book from the table*", and where a corresponding robot action such as `take(book, table)` has to be instantiated correctly. Such instantiation is generally triggered by a trained model, where noise, over-fitting, and mislabeling could indeed bring to an undesired output, e.g. the robot placing the book on the table. In the view of symbiotic autonomous robots (Rosenthal, Biswas, and Veloso 2010) that rely on humans to overcome their limitations and correct their actions, a transparent model could help identifying and explicit the reason(s) behind the wrong behavior of the robot.

Our motivation is the semantic processing of robotic commands (also called *semantic parsing*) from spoken language utterances, i.e. the process of mapping natural language sentences to formal meaning representations. The formal meaning representation theory we rely upon is Frame Semantics (Fillmore 1985), describing actions and events expressed in language through conceptual structures called *semantic frames*. This theory also states that a frame is evoked in a sentence through the occurrence of specific *lexical units*, i.e. words (such as verbs and nouns) that linguistically express the underlying situation. To identify such frames, we

EB and IT developed the theoretical framework and directed the project; EB and GR designed the experiments; MM derived the models, performed experiments and analysed the results; IT and EB wrote the manuscript in consultation with GR and MM.

[1] https://www.darpa.mil/program/explainable-artificial-intelligence

[2] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

built a semantic parser based on a multi-layer Long-Short Term Memory (LSTM) neural network with attention (Mensio et al. 2018), and trained it over the Human-Robot Interaction Corpus (HuRIC) (Bastianelli et al. 2014). LSTMs as many similar deep nets-based models have an opaque nature, i.e. they do not give clear clues on the way they behave, which may complicate the understanding of undesired behaviors as, in our case, an incorrect robot behavior. Moreover, understanding the inner workings of such models tend to be harder when trained on small, domain-specific datasets (such as HuRIC), as they often lack of effective representativeness of the problem domain. The questions we wish to answer in this work are therefore:

- how can we better understand our LSTM-based model?
- how we can we identify undesired behaviors in the model?
- is there a way to mitigate such undesired behaviors?

To answer the first two questions, we rely on the idea that linguistic theories could be used in the context of our semantic parser to obtain more understanding of the model, i.e. they could be exploited to provide to explain the model's behavior. Recent trends in deep learning have shown that visual explanations for the models' behavior could be obtained through the analysis of the values of the attention layers in a number of tasks (Machine Translation (Bahdanau, Cho, and Bengio 2014), Sentiment Analysis (Lin et al. 2017), Image Captioning (Xu et al. 2015)) for their ability of correlating inputs and outputs. Inspired by these works, our hypothesis is that we can use attentions to achieve some degree of explainability for the LSTM-based parser, and that Frame Semantics can be the key to drive the interpretation process. We therefore use attentions to capture the interpretation of spoken commands and, more specifically, use the values that the attention layer assign to each word of a given sentence to detect which word is the lexical unit evoking (i.e. causing) the identified frame. We show how this not only gives us a hint on the model behavior, but that attentions help unveiling the intrinsic bias induced by our training data. Here, we exploit the linguistic knowledge encoded in an external resource such as FrameNet (Baker, Fillmore, and Lowe 1998) in a data augmentation strategy, with the goal of mitigating the corpus bias, improve the explanations that the model provides and, consequently, the overall model results.

Although preliminary, our promising results suggest that attention layers combined with Frame Semantics do provide a clue to a more explainable model, and that the integration of external knowledge bases can help overcoming the inner limitations of machine learning models. More importantly, our method suggests that the combination of knowledge engineering and machine learning techniques can be beneficial for the development of more transparent, understandable intelligent systems.

Motivation and Background

In this section, we present the theoretical and technical background of our work. We first discuss Frame Semantics, which we use as linguistic theory of reference, and then describe the technical details of our neural network-based semantic parser.

Fundamentals of Frame Semantics

Frame Semantics is a theory that formalizes how a sentence is related to semantic frames. Each frame is a conceptual structure representing an action or, more in general, an event or situation (e.g. the action of *Taking*). Frames are further specified by a set of frame elements (e.g. the THEME, representing the object taken while performing the action *Taking*), which enhance the meaning of the frame with additional information. According to Frame Semantics, frames are evoked in sentences by specific words, called *lexical units* (LU). Lexical units are responsible to convey the meaning of the frames, representing hooks between the textual surface and the theory itself. In the example of Figure 1, the frame *Taking* is evoked in the sentence "*take the book to the table*" by the LU *take*, while *the book* and *to the table* represent the THEME and the GOAL frame elements respectively:

$$[take]_{\textbf{\textit{Taking}}}$$
$$[the\ book]_{\text{THEME}}$$
$$[from\ the\ table]_{\text{ORIGIN}}$$

Figure 1: Example of semantic frame annotation for the sentence "*take the book from the table*".

The process of annotating Frame Semantics over natural language involves three different tasks. First, all the frames evoked in a sentence are identified looking at the potential LUs contained it. This task is generally called Frame Prediction or Frame Induction. Here, we refer to it as Action Detection (AD), as we are dealing with the action expressed by the person uttering the command to the robot. The second task is called Argument Identification (AI, sometimes also called Boundary Detection) and is responsible to find the spans of text corresponding to possible frame elements. The last task is called Argument Classification (AC) and consists in assigning a label to the spans identified during the AI. Note that the AI and AD tasks are often referred together as the process of Semantic Role Labelling.

If we take the example of "*take the book from the table*", the frame *Taking* would be predicted in the AD step by identifying the LU *take*. In the AI step, *the book* and *from the table* would be identified as 2 frame element spans, and respectively classified as THEME and ORIGIN frame elements in the following AC step.

A multi-layer LSTM-based parser

In our previous work (Mensio et al. 2018), we presented a semantic parser for robotic commands, called 3LSTM-ATT, based on a multi-layer LSTM network exploiting attention mechanisms. The 3LSTM-ATT topology is shown in Figure 2. The network was adapted from (Liu and Lane 2016) so that each layer could carry one of the three semantic parsing tasks presented above. We briefly describe the network in the following, and refer the reader to the original paper for more details.

The input to the network is a tokenized sentence, where each token is embedded using the GloVe word embeddings (Pennington, Socher, and Manning 2014), pre-trained

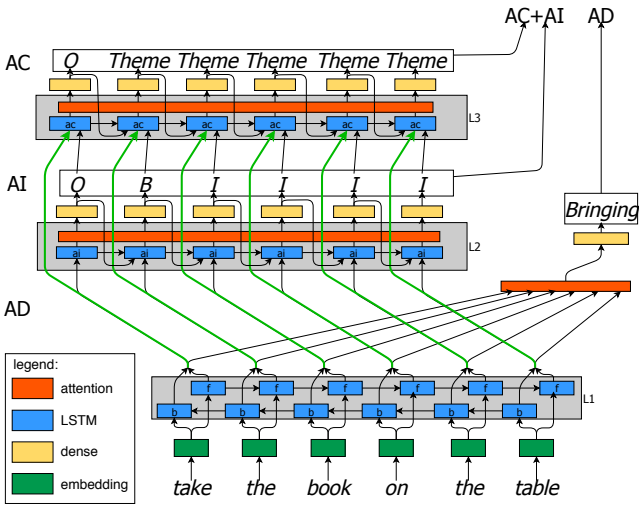

Figure 2: The neural network for the semantic parser. The connections in green represent highway connections between the first and the third layer.

over the Common Crawl resource[3]. The sequence is firstly encoded with a bidirectional LSTM (L1). For the AD task, a single contextual representation for the whole sequence c is computed through an attention layer (Bahdanau, Cho, and Bengio 2014), which is in turn passed through a fully connected layer with a final softmax activation to obtain per-frame probabilities. The sequence out of L1 is further encoded with a LSTM (L2) with self-attention (Cheng, Dong, and Lapata 2016). Single hidden representations of the tokens are classified through a dense layer with softmax into `IOB` labels, which denote whether a word is the `B`eginning, the `I`nside or it is `O`utside of a frame element span. The LSTM at L2 is modified so that, at each time step, the output of the dense layer at $t-1$ is provided as additional input to the LSTM cell at time t. The third and final encoding layer (L3) takes as input the output of L2 and the output of L1 through highway connections. The same type of encoder used in L2 is applied in L3, with the difference that the dense layer outputs frame element labels instead of `IOB` ones.

The simple attention mechanism (Bahdanau, Cho, and Bengio 2014) used for the AD task is a layer that gives an insight of the contribution that a certain input gives in the production of a given output. The final contextual representation of a sentence c is evaluated as the weighted sum:

$$c = \sum_i a_i h_i,$$

where h_i represents the encoding of the i-th token and the attention value (or score) a_i is evaluated through a simple feedforward network $f_{att}(h_i)$. Roughly speaking, this attention layer evaluates a value a_i for each encoded input token. Since the AD classification layer operates over the contextual representation c, each value a_i indicates how much each word in a sentence contributes to the final classification of a frame. For this reason, it can intrinsically provide an explanation for the model behavior, as it summarizes a much broader set of values that can be more difficult to interpret (e.g. looking at all the values of the self-learned weights). In fact, it enables to underline a restricted subset of features, because not all the inputs have the same importance. The self-attentions used in the two other layers (L2 and L3) instead encode the relationship among all the input objects, e.g. of much each token contributes to the representation of all the other tokens for a given task. We point the reader to the original paper for more details about the self-attention layers.

[3]http://commoncrawl.org/

Hypotheses and Challenges

Taking back our research questions, at this point we ask:

- how can we better understand the LSTM-based model we built?
- how can we identify an undesired behavior in such model?
- is there a way to mitigate any undesired behaviors?

Our first question can be answered looking at the attention layer values to get hints on the model's behavior. As previously discussed, attentions give the chance to explore the intermediate classification steps, enabling the interpretability of how the system processes a given input – an aspect that we can exploit for a better understanding of our process. As a first attempt, this work aims at answering the previous questions by taking into account the sole ability of the system to detect the correct frame. For this reason, we will focus on the analysis of the attention values for the sole AD task. We leave the analysis of the other two tasks for forthcoming work.

We thus answer our second question by aligning attentions and the linguistic theory. On the one hand, we have the Frame Semantics theory that states that frames are evoked in natural language by specific words called lexical units. On the other, we have the attention values computed by the network to balance the input words in the final contextual representation used to classify the frame. Our assumption therefore is that, by annotating data with Frame Semantics, the algorithm learning from such data should encode implicitly the theory itself, through an attempt of learning it (or a good approximation of it). If the network is learning correctly from the data, we should therefore observe an alignment between the values produced by the attention of the AD layer and what is stated by Frame Semantics, e.g. we should notice relevant values attributed to words that could possibly be lexical units for the classified frame. Should the network not follow the underlying Frame Semantic theory, this could mean not only that the model is only following patterns statistically evident in the data (and not related to the theory), but also that an incorrect explanation for its behavior would be provided if requested. Our challenge is first to verify whether the words receiving the highest attention values are the correct lexical units of a classified frame (e.g. given the sentence "*take the book from the table*", the word *take* should be given a high attention value).

Finally, we need a mitigation strategy to overcome the cases where the attention turns out to be focused on the incorrect lexical element and consequently ensure that the correct explanation for a decision can be provided. Given that

HuRIC's annotations are based on Frame Semantics, we propose to augment the dataset using additional examples from the FrameNet corpus (Baker, Fillmore, and Lowe 1998). Although FrameNet cover a different domain w.r.t HuRIC, i.e. written vs. spoken language, we believe that, by using a data augmentation strategy, the algorithm can be driven to rely on patterns consistent with the theory, and thus to achieve better generalization.

Approach

In this section, we show the design of the overall approach, namely (1) how we align the model to the Frame Semantics theory; (2) how we use these alignments to identify misbehavior by the model; and (3) the data augmentation strategy we use to mitigate the bias in the model.

Aligning Attentions and Linguistic Theory

As previously explained, the attention values produced by our 3LSTM-ATT parser during the AD stage can be used to guess which words in a sentence are more relevant to the classified frame. We can use these values to attempt an alignment between words and the linguistic theory, namely which words are lexical units or, other relevant words such as prepositions.

The parser has been trained over the previously mentioned HuRIC dataset, which contains transcriptions of user commands tagged with Frame Semantics. The annotated frames generally correspond to actions like taking objects or moving to a specific position. The dataset contains 585 frame occurrences over 526 sentences on 16 different frame types for an average of ∼36 sentences per frame. The results, obtained over this dataset through a 5-fold cross validation stratified on the frame types, are reported in Table 1. Compared to results of (Bastianelli et al. 2016) (BAS16 henceforth)[4], our parser obtains better results for both the AD and AI tasks.

Table 1: Parser performances in terms of F-Measure for the AD, AI and AC, compared to BAS16. Only gold values considered as input of each task.

Corpus	AD	AI	AC
BAS16	94.67%	90.74%	**94.93%**
3LSTM-ATT	**96.33%**	**94.35%**	91.77%

(Mis-)alignment of Attention Values

Differently from BAS16, we can take advantage of the attention layer in the AD step to understand our system's behavior when classifying a frame for a given sentence. As explained, our assumption is that the word receiving the highest value from the AD attention layer may be the LU for the classified frame.

In order to prove such hypothesis, we need to quantitatively measure the alignment between the attention values and the "gold" LU for a given frame. Let $S = (w_1, ..., w_m)$ be a sentence as a sequence of m words w. Gold LUs are available in HuRIC, so let $\hat{w}_i$ be the gold LU for the i-th sentence S_i[5]. Let us consider the attention layer as a (simplified) function $f_{att}(w)$ that attributes an attention value to a word w (for clarity, w is a shortcut for the hidden representation h). The $\mathcal{A}_{LU}$ (LU-alignment) measure can be then calculated as follows:

$$\mathcal{A}_{LU} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\arg\max_{w \in S_i} f_{att}(w) = \hat{w}_i) \quad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Although lexical units carry most of the meaning for a frame, there are still many ambiguous cases, where a verb alone may evoke different frames. Consider for example the verb *take*, which may evoke the frame *Bringing*, e.g. in the sentence *take the book to the table*, or *Taking*, e.g. *take the book from the table*. The meaning in this case is not carried only by the LU alone, but also by the co-occurrence with other specific words or syntactic structures. The preposition *to* in the first example clearly introduces an argument representing the destination of a motion (i.e. the GOAL frame element), helping in choosing the frame *Bringing* over *Taking* for the word *take*. It is thus legit to think that, in these cases, part of the attention values should also focus on such discriminant words.

We thus designed a second measure, that we call $\mathcal{A}_D$ (discriminant alignment), with the aim of taking into account additional discriminant words in addition to the LU. To this end, we annotated the discriminant words for each sentence in the dataset. For each sentence $S = (w_1, ..., w_m)$, we created a vector of gold discriminant word indexes $v_g = (gd_1, ..., gd_m)$ where each $gd_j \in \{0, 1\}$ is set to 1 if its position corresponds to a discriminant word in S. Given the attentions values obtained from the AD layer, we created a vector of classified discriminant word indexes $v_c = (cd_1, ..., cd_m)$ where each $cd_j = \mathbb{I}(f_{att}(w_j) \geq 0.01)$[6]. Finally, we calculated Precision and Recall over these vectors the following way:

$$P = \frac{\sum_{j=1}^{m} \mathbb{I}(cd_j = gd_j = 1)}{\sum_{j=1}^{m} cd_j} \quad (2)$$

$$R = \frac{\sum_{j=1}^{m} \mathbb{I}(cd_j = gd_j = 1)}{\sum_{j=1}^{m} gd_j} \quad (3)$$

through which we obtained the F-Measure. The $\mathcal{A}_D$ was finally calculated as macro-average over the F-Measure of all the sentences S_i in the dataset.

The HuRIC→HuRIC row of Table 2 shows the scores for $\mathcal{A}_{LU}$ and $\mathcal{A}_D$ obtained when training and testing over HuRIC. As we can see from the 11.17% on $\mathcal{A}_{LU}$ and 20.53% on $\mathcal{A}_D$ values, the model reaches good results on the AD task (96.33% of F-Measure), but is quite misaligned

[4]Please note that the BAS16 makes use also of perceptual features, while our parser relies only on linguistic inputs.

[5]Sentence splitting was applied in order to have 1 frame per sentence, for the rare HuRIC cases containing more than one frame per sentence.

[6]This threshold was set to filter attention noises. The study how to properly set this threshold is left for future work.

from the linguistic theory. Indeed, the error analysis we carried on the attention values reported that the model is following latent patterns, which are completely unrelated to the theory, rather than generalizing the linguistic theory as expected. In other words, the model concentrates its attention on recurrent words that are not discriminative with respect to the respective frame; yet, it was able to produce the correct classification.

Model	Frame gold	Frame pred	take	the	red	shoes
			LU	-	-	-
HuRIC→HuRIC	*Taking*	*Taking*	0.001	0.627	0.371	0

Model	Frame gold	Frame pred	inspect	the	red	shoes
			LU	-	-	-
HuRIC→HuRIC	*Inspecting*	*Taking*	0.065	0.214	0.716	0

Figure 3: Attention analysis for two different input sentences. The attention falls mostly on words that are not LUs, e.g. *the*, *red*.

An example of such behavior is reported in Figure 3: while the *Taking* frame is indeed correctly identified, the attention values reveal that the model attention falls on the two words *the* and *red*, which do not convey any frame meaning in this context, while the correct LU *take* receives only 0,1% of attention. As an additional proof, a similar sentence with a different frame, e.g. *inspect the red shoes*, is classified with the same frame *Taking* (instead of *Inspecting*), with most of the attention falling again on words *the, red*.

A first consideration that can arise from the above analysis is that linguistic phenomena are not equally represented in HuRIC (i.e. some frames happen in correspondence of more frequent, but not necessarily significant, grammatical patterns), and this lack of representativeness might cause intrinsic bias. This prevents the model to learn the underlying linguistic theory, and to generalize from it.

Mitigating the Data Bias

If we hypothesize that our model does not generalize towards the linguistic theory as it should due to the lack of representativeness of the dataset, a natural solution is to try to increase the number of training examples to see if the alignment measures improve without compromising the performances. Since HuRIC is tagged with Frame Semantics following the same scheme as the FrameNet corpus (Baker, Fillmore, and Lowe 1998), the first solution at hand to attenuate the bias with more examples consists in integrating HuRIC with examples from FrameNet itself. For the purpose of comparison, we selected only the FrameNet examples annotated with frames also contained in HuRIC. This selection resulted in a subset of 6,814 frame examples, for an average of ∼425 examples per frame.

Although sharing the same background linguistic theory, however, the two datasets belong to two different domains, namely written text vs. spoken commands. This indeed may lead to a drop in terms of performances. Let us take the example of FrameNet annotated-sentence for the frame *Taking*:

In the late 1870s, he defaulted on a loan from rancher Archibald Stewart, so [*Stewart*]$_{\text{AGENT}}$ [***took***]$_{\textbf{\textit{Taking}}}$ [*the Las Vegas Ranch*]$_{\text{THEME}}$ [*for his own*]$_{\text{EXPLANATION}}$.

Indeed, the label set of frame elements, and, in general, the variability of the language in FrameNet is, in fact, much higher than HuRIC. On the one hand, this can negatively contribute to the overall performance, as the complexity of the task increases. On the other, the network will access more evidence in terms of theory-related patterns, e.g. seeing more often the association of the frame *Taking* with co-occurring verbs like *take*, than with other unrelated words like *shoes*. Our aim is therefore to reach a good trade-off between the model's performance and its degree of generalization that, in turn, reveals the degree of understandability (explainability) of its behavior.

Experiments and Results

In order to support our hypotheses about the mitigation strategy, we designed two additional experimental settings with the goal of evaluating the changing in the model behavior:

- FN→HuRIC: a model is trained over the full subset of samples coming from FrameNet, and is tested on the whole HuRIC dataset;
- FN+Hu→HuRIC: the evaluation follows a 5-fold cross validation. At each validation turn, the training set consists in FrameNet + 80% of HuRIC, leaving the remaining 20% as test set. The distribution of frames is uniformly stratified.

Table 2 presents the results of the $\mathcal{A}_{LU}$ and $\mathcal{A}_D$ for both configurations. The performances of the semantic parser in terms of F-Measure for the AD, AI and AC tasks are reported as well. Please note that HuRIC→HuRIC results differ from the ones in Table 1, showing performances of the single tasks in isolation (i.e. each task receives gold information from the previous steps). Instead, we consider here the full semantic parsing pipeline.

It appears clear how the parser performances and the alignment measure scores are reversed for the two different settings. The models trained only on FrameNet do not achieve high performances, reaching only approx. 68% for the AD task. When the two datasets are combined, an increase of ∼19% points is achieved for the same task. This is still very low when compared to the 96.33% achieved with HuRIC only. With that said, by looking at the alignment measure scores, we notice that this drop of performance comes at the advantage of the model's explainability. When trained only on FrameNet, in fact, the $\mathcal{A}_{LU}$ and $\mathcal{A}_D$ scores reach 93.92% and 84.86% respectively. This confirms that the AD attention layer is focusing on the relevant words, hence giving us a hint that the model is correctly learning the linguistic theory. The introduction of HuRIC to the training sample helps in raising the parsing performances to convincing levels, while not deteriorating completely the alignment. The $\mathcal{A}_{LU}$ and $\mathcal{A}_D$ still drop by ∼40 and ∼34 points respectively, but considering the performances reached by the

Model	Frame gold	Frame pred	get	the	dishes	from	the	dining	room
			LU	-	-	-	-	-	-
HuRIC→HuRIC	*Taking*	*Taking*	0.001	0.908	0.018	0.041	0.021	0.011	0
FN→HuRIC	*Taking*	*Entering*	0.994	0.002	0	0.004	0	0	0
FN+Hu→HuRIC	*Taking*	*Taking*	0.984	0	0.001	0.012	0.003	0	0

Figure 4: Result of the attention analysis over the three different training conditions for the sentence *get the dishes from the dining room*.

parser, this can be considered an encouraging trade-off. Although the bias has been corrected to a certain extent, the overall results suggest that HuRIC is still introducing some noise, which diverts the system from the full alignment with the underlying theory. Testing the use of different amount of examples from FrameNet and HuRIC may result in an even better balancing of linguistic variance and the domain-specificity.

Table 2: End-to-end performances and alignment scores of the 3LSTM-ATT parser for the three different training settings. F-Measure is reported for the AD, AI and AC tasks.

	HuRIC→HuRIC	FN→HuRIC	FN+Hu→HuRIC
AD	96.33%	68.06%	87.60%
AI	93.57%	77.14%	81.27%
AC	87.22%	62.70%	72.44%
$\mathcal{A}_{LU}$	11.17%	93.92%	51.31%
$\mathcal{A}_{DS}$	20.53%	84.64%	50.83%

In order to better demonstrate the trade-off between parser performances and theory alignment, we also perform a qualitative analysis on the test examples. In Figure 4, we show the AD tagging and attention values produced by the different training settings (i.e. HuRIC, FN, FN+Hu) for the sentence *get the dishes from the dining room*. When trained on HuRIC only, the network learns again unwanted patterns and, although the frame *Taking* is correctly classified, the attention mostly falls on the article *the*, also spreading with minor values on the rest of the words. By using FrameNet as training set, the attention falls back to the verb *take* that corresponds to the current LU. However, the frame classification fails, predicting *Entering*. The correct frame classification (*Taking*) with attention values matching the correct LU and, to a minor extent, the discriminant preposition *from* is finally obtained when using a combination of the two corpora, as in the last row.

The same behavior can be observed if we consider also other discriminant words in the sentence. Figure 5 shows again frame parsing and attentions values over different sentences. Discriminative words are here reported as well (DISC). In all the four examples it appears clear that when the system is trained only over the HuRIC resource, the attention is unstable, i.e. either it distributes similarly among more or less relevant words (5a), or more strongly attending on non-discriminant words at all (5b). In other cases, the attention indeed does attend on discriminant words, but either the final frame classification is wrong for a lack of value on the LU (5c), or, even if the frame is correct, we lose the dependence of the classification outcome on the LU (5d). When using FrameNet as a training set, the system is able to better attend on LUs (5b–5d). The distance in the application domain seems to still prevent the system to attend also on discriminant words. For the same reason, in other cases the attention still spreads its mass over non-relevant words (5a–5c). This leads to errors in the frame classifications. A more stable behavior can be observed when both FrameNet and HuRIC are used as training set. The attention values, in fact, stabilize mostly over LUs and discriminant words, although with more dense or sparse values. This contributes to a much better frame classifications, giving us an insight of the difference of the results in Table 2.

This confirms the idea that the use of a compatible external resource such as FrameNet can help in reducing the bias of poorly represented corpora that can affect deep network architectures. At the same time, attention values can be analyzed to interpret the outcome of the model classification (frames/actions in our case). More importantly, this method promotes the idea that knowledge engineering, which helps encodes and elicit expert's knowledge (e.g. FrameNet), and machine learning techniques can be combined to develop more transparent and understandable systems.

Related Work

We divided the related work in three parts: (i) approaches to enable more explainable deep learning-based applications, with a particular focus on text classification and attention methods, (ii) approaches to mitigate bias in data and (iii) approaches for semantic parsing in the robotics domain.

Explainability for Deep Learning

Explainability for deep learning methods can be divided in three families. A first family, including perturbation experiments (Zeiler and Fergus 2014), saliency map-based methods (Simonyan, Vedaldi, and Zisserman 2013), LIME (Ribeiro, Singh, and Guestrin 2016) and influence functions (Koh and Liang 2017), relies on methods trying to identify the relevant features treating the model as a black box. An approximated model is built by observing concurrent changes between the input and the output, so that it can provide simple explanations.

A second family of approaches focuses on inspecting the internal representations and input processing. By observing the inner parameters (weights of the neural network, or other latent variables), these methods try to give a meaning to layers and operations in a bottom-up way (Zhang and Zhu 2018). For this reason, their application is difficult to scale for networks with lots of layers and parameters.

A third family consists in the intrinsically explainable

Model	Frame gold	Frame pred	bring	it	to	the	side	of	the	bathtub
			LU	-	DISC	-	-	-	-	-
HuRIC→HuRIC	*Bringing*	*Bringing*	0.0003	0.003	0.2815	0.2205	0.1785	0.1267	0.1796	0.01
FN→HuRIC	*Bringing*	*Placing*	0	0.0001	0.4605	0	0.5394	0	0	0
FN+Hu→HuRIC	*Bringing*	*Bringing*	0.689	0.0107	0.3002	0	0	0	0	0

(a)

Model	Frame gold	Frame pred	look	for	the	wrench	in	the	bathroom
			LU	DISC	-	-	-	-	-
HuRIC→HuRIC	*Searching*	*Perception_active*	0.0032	0.0567	0.0294	0.0789	0.4687	0.3628	0.0001
FN→HuRIC	*Searching*	*Perception_active*	0.9999	0	0	0.0001	0	0	0
FN+Hu→HuRIC	*Searching*	*Searching*	0.1034	0.8966	0	0	0	0	0

(b)

Model	Frame gold	Frame pred	robot	please	take	the	mug	to	the	sink
			-	-	LU	-	-	DISC	-	-
HuRIC→HuRIC	*Bringing*	*Taking*	0	0	0.0004	0.0192	0.0343	0.9104	0.0352	0
FN→HuRIC	*Bringing*	*Following*	0	0.4691	0.4703	0	0.0606	0	0	0
FN+Hu→HuRIC	*Bringing*	*Bringing*	0	0	0.1773	0	0	0.8227	0	0

(c)

Model	Frame gold	Frame pred	take	the	jar	to	the	table	of	the	kitchen
			LU	-	-	DISC	-	-	-	-	-
HuRIC→HuRIC	*Bringing*	*Bringing*	0	0.0002	0.0019	0.7515	0.0545	0.1101	0.0192	0.0567	0.0057
FN→HuRIC	*Bringing*	*Taking*	0.9394	0	0.0606	0	0	0	0	0	0
FN+Hu→HuRIC	*Bringing*	*Bringing*	0.089	0	0.0001	0.9109	0	0	0	0	0

(d)

Figure 5: Attention analysis in relation to both the LU and discriminant words (DISC) for the three training settings.

models, which are complex enough to reach good performances, yet providing good hints for interpretation. Attention layers exactly provide a relevance measure between the inputs and outputs, by learning a salience map between the two other network layers, which can be further visualized using heat-maps independently from the domain considered. Visual attention (Mnih et al. 2014) has been used in the automatic Image Captioning task (Xu et al. 2015; You et al. 2016) where, given an input picture a textual caption is generated. The attention values can be observed to highlight the area of the picture which most contributed to generate specific words in the caption. and which can be visualized using heat-maps independently from the domain considered. Self-attentions (Bahdanau, Cho, and Bengio 2014) have also been widely applied in many text processing tasks, such as Sentiment Analysis (Lin et al. 2017) and Question Answering (Hermann et al. 2015). Visual explanations were used in these cases to explain alignments between the words of the input and output sentences.

Bias in Data

In their work, (Zhao et al. 2017) studies the problem of quantifying gender bias in data and models for multi-label object classification and visual semantic role labeling, developing a calibration strategy that introduces frequency-constraints on the training corpus. In the context of recommender systems, (Adomavicius et al. 2014) propose to mitigate the biased customers' ratings after the classification, both with a systematic algorithm and with an interactive user-interface.

Several data augmentation methods for Generative Adversarial Networks that use image intensity normalization, rotation, re-scaling, cropping, flipping, and Gaussian noise injection were presented in the context of medical image analysis (Drozdzal et al. 2018; Hu et al. 2018; Roth et al. 2015).

Little work has been done on how to exploit alignments between knowledge bases for machine learning systems. The Knowledge Representation community has mostly focused on empirically analyzing the effects of data links, i.e. (Tiddi, d'Aquin, and Motta 2014) uses alignments to quantify bias in datasets pairwise, without suggesting mitigation solutions; (Ding et al. 2010) discussed the confusion of provenance and ground truth generated by `owl:sameAs` in the context of bioinformatics datasets; (Beek et al. 2018) gathers and fixed erroneous identity statements offering them in a large-scale dataset.

Knowledge bases integrated with deep nets have so far been used to improve the embedding space at training time or to explain the model's outputs a posteriori (cfr. (Hitzler et al. 2019) for a representative selection). To the best of our knowledge, our work is the first using an external knowledge bases aligned to the training corpus to mitigate the bias in a training dataset in the context of deep nets.

Semantic Parsing for Robotic Applications

A variety of approaches have been proposed in the last two decades to create semantic parsers for commands of virtual and real autonomous agents. With the breakthrough of statistical models, many machine learning techniques have been applied to semantically parse robot instructions, from sequential labelling (Kollar et al. 2010), Statistical Machine Translation (Chen and Mooney 2011), learning-to-rank (Kim and Mooney 2013) and probabilistic graphical models (Tellex et al. 2011). Statistical methods have also been applied to induce grammars to parse human commands into suitable meaning representations as well (Artzi and Zettlemoyer 2013; Thomason et al. 2015). These approaches were implemented mostly in discretized environments, relying on ad-hoc and formulaic representation formalisms, and often dealing with constrained vocabularies. Our work, on the contrary, builds upon the idea of relying on linguistically sound theories of meaning representation, e.g. Frame Semantics, to bridge between linguistic knowledge and robot internal representations. We build upon (Bastianelli et al. 2016) to design a parser to identify semantic frames expressed in robot commands but rely on the bidirectional LSTM network.

Conclusions

In this paper, we have presented an approach relying on the integration of heterogeneous knowledge sources to mitigate the biased results of a deep learning-based semantic parser for Spoken Language Understanding for robots, and improve the model's understandability. We discussed how current models do not necessarily learn the underlying linguistic theory, but rather focus on unwanted, unexpected patterns, because of an intrinsic bias induced by the size and domain-specificity of the training dataset. We showed how the values of the attention layers of the network can be used as a clue to analyze and interpret the model's behavior, as the classification of frames in our case. Finally, we have provide evidence that external resources such as FrameNet can help to reduce the bias in the training data, also guaranteeing the correct interpretations (or explanations) for the model's behavior. While being a preliminary attempt to measure a more complex phenomenon, our work suggests that the strengths of both knowledge engineering and machine learning can be combined to foster the development of more transparent, understandable intelligent systems.

The future work will be focused in a first instance on designing more thorough evaluation schemes to obtain better quantitative understandings of the model's behavior. Secondly, we will focus on identifying the correct balance between the domain-specific samples and the external ones, also testing new pairs of datasets if possible. An analysis carried by gradually combining the samples and showing how the performances and the explainability measures behave across several datasets and domain is indeed crucial. Extending the use of more knowledge bases through their links (e.g. WordNet, ConceptNet) is another route we wish to follow. Finally, we will explore the idea of interactive, symbiotic explanations, where the model can be corrected through spoken dialogue with the user.

References

Adomavicius, G.; Bockstedt, J.; Curley, S.; and Zhang, J. 2014. De-biasing user preference ratings in recommender systems. In *RecSys 2014 Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2014)*, 2–9.

Artzi, Y., and Zettlemoyer, L. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics* 1(1):49–62.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The Berkeley FrameNet project. In *Proceedings of ACL and COLING*, Association for Computational Linguistics, 86–90.

Bastianelli, E.; Castellucci, G.; Croce, D.; Iocchi, L.; Basili, R.; and Nardi, D. 2014. Huric: a human robot interaction corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Bastianelli, E.; Croce, D.; Vanzo, A.; Basili, R.; and Nardi, D. 2016. A discriminative approach to grounded spoken language understanding in interactive robotics. In *Proceedings of the 2016 International Joint Conference on Artificial Intelligence (IJCAI)*.

Beek, W.; Raad, J.; Wielemaker, J.; and Van Harmelen, F. 2018. sameas. cc: The closure of 500m owl: sameas statements. In *European Semantic Web Conference*, 65–80. Springer.

Chen, D. L., and Mooney, R. J. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on AI*, 859–865.

Cheng, J.; Dong, L.; and Lapata, M. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 551–561. Austin, Texas: Association for Computational Linguistics.

Ding, L.; Shinavier, J.; Finin, T.; McGuinness, D. L.; et al. 2010. owl: sameas and linked data: An empirical study. In *Proceedings of the Second Web Science Conference*.

Drozdzal, M.; Chartrand, G.; Vorontsov, E.; Shakeri, M.; Di Jorio, L.; Tang, A.; Romero, A.; Bengio, Y.; Pal, C.; and Kadoury, S. 2018. Learning normalized inputs for iterative estimation in medical image segmentation. *Medical image analysis* 44:1–13.

Fillmore, C. J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* 6(2):222–254.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, 1693–1701.

Hitzler, P.; Bianchi, F.; Ebrahimi, M.; and Sarker, M. K. 2019. Neural-symbolic integration and the semantic web. *Semantic Web* (Preprint):1–9.

Hu, X.; Chung, A. G.; Fieguth, P.; Khalvati, F.; Haider, M. A.; and Wong, A. 2018. Prostategan: Mitigating data bias via prostate diffusion imaging synthesis with generative adversarial networks. *arXiv preprint arXiv:1811.05817*.

Kim, J., and Mooney, R. J. 2013. Adapting discriminative reranking to grounded language learning. In *ACL (1)*, 218–227. The Association for Computer Linguistics.

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*.

Kollar, T.; Tellex, S.; Roy, D.; and Roy, N. 2010. Toward understanding natural language directions. In *Proceedings of the 5th ACM/IEEE*, HRI '10, 259–266.

Lin, Z.; Feng, M.; Santos, C. N. d.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Liu, B., and Lane, I. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In *INTERSPEECH*, 685–689. ISCA.

Mensio, M.; Bastianelli, E.; Tiddi, I.; and Rizzo, G. 2018. A multi-layer lstm-based approach for robot command interaction modeling. *Workshop on Language and Robotics, IROS 2018*.

Mnih, V.; Heess, N.; Graves, A.; et al. 2014. Recurrent models of visual attention. In *Advances in neural information processing systems*, 2204–2212.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. ACM.

Rosenthal, S.; Biswas, J.; and Veloso, M. 2010. An effective personal mobile robot agent through symbiotic human-robot interaction. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, 915–922. International Foundation for Autonomous Agents and Multiagent Systems.

Roth, H. R.; Lu, L.; Liu, J.; Yao, J.; Seff, A.; Cherry, K.; Kim, L.; and Summers, R. M. 2015. Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE transactions on medical imaging* 35(5):1170–1181.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M.; Banerjee, A.; Teller, S.; and Roy, N. 2011. Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine* 32(4):64–76.

Thomason, J.; Zhang, S.; Mooney, R.; and Stone, P. 2015. Learning to interpret natural language commands through human-robot dialog. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI)*, IJCAI'15, 1923–1929. AAAI Press.

Tiddi, I.; d'Aquin, M.; and Motta, E. 2014. Quantifying the bias in data links. In *International Conference on Knowledge Engineering and Knowledge Management*, 531–546. Springer.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4651–4659.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Zhang, Q.-s., and Zhu, S.-C. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19(1):27–39.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.