ABSTRACT REPRESENTATION FOR MULTI-INTENT SPOKEN LANGUAGE UNDERSTANDING

Rim Abrougui, Géraldine Damnati, Johannes Heinecke, Frederic Bechet

ABSTRACT REPRESENTATION FOR MULTI-INTENT SPOKEN LANGUAGE UNDERSTANDING

Rim Abrougui[1,2]*, Géraldine Damnati*[1]*, Johannes Heinecke*[1]*, Frédéric Béchet*[2]

[1]Orange Innovation, Lannion, France
[2]Aix-Marseille University, CNRS, Marseille, France

ABSTRACT

Current sequence tagging models based on Deep Neural Network models with pretrained language models achieve almost perfect results on many SLU benchmarks with a flat semantic annotation at the token level such as ATIS or SNIPS. When dealing with more complex human-machine interactions (multi-domain, multi-intent, dialog context), relational semantic structures are needed in order to encode the links between slots and intents within an utterance and through dialog history. We propose in this study a new way to project annotation in an abstract structure with more compositional expressive power and a model to directly generate this abstract structure. We evaluate it on the MultiWoz dataset in a contextual SLU experimental setup. We show that this projection can be used to extend the existing flat annotations towards graph-based structures.

Index Terms— Natural Language Understanding, Spoken Language Understanding, sequence tagging, sequence-to-sequence models

1. INTRODUCTION

Spoken Language Understanding (SLU) is a task that has been studied mostly in the context of human-machine spoken dialog systems on benchmark corpora implementing a *domain/intent/slot/value* model where each query contains a list of slot/value pairs corresponding to a single intent from a single domain. Such a *flat* semantic annotation scheme can be used to annotate dialogue corpora by labelling each query with a *domain/intent* label, then projecting the *slot/value* pairs at the word level with structure labels such as *B,I,O* (*Begin, Inside, Outside*) as prefix to the slot label to predict for each word. Current sequence-tagging models based on *Deep Neural Networks* with pretrained language models achieve almost perfect results on many of these *flat* SLU benchmarks [1, 2, 3, 4]. When dealing with more complex human-machine interactions (multi-domain, multi-intent, multi-turn), relational semantic structures are needed in order to encode the meaning of a given turn, leading to a change in both the way we annotate the semantic structure of an utterance and in the choice of the prediction models used to generate this annotation.

In this study we propose a novel encoding, called ARMILU (*Abstract Representation for Multi-Intent spoken and natural Language Understanding*), for representing SLU semantic structures that can be used for graph-based or flat semantic structures and which is particularly adapted to be used with generation models. We will compare standard token-level classification models with B,I,O encoding to generation models using the ARMILU encoding on the MultiWOZ dataset.

2. RELATED WORK

Some recent work try to overcome the flat intent-slot representation and allow compositional interpretation. [5] propose an embedded representation in the TOP corpus (followed by its multilingual version MTOP [6]) where intents can be nested through a tree structure and implement a shift-reduce algorithm to parse utterances. [7] extend this approach with decoupled representation (SB-TOP) in order to handle discontinuity and long-distance dependencies. They propose a parser based on Pointer-Generator seq2seq models. However TOP and SB-TOP corpora remain limited to simple commands (single utterances on navigation, events and navigation to events for TOP and short sessions on calling, weather, music and reminder for SB-TOP). Additionally [8] have introduce DMR which is intended to be more general across domains and propose a new ontology with domain-agnostic and domain-specific parts. Although DMR representation shares common ideas with our proposition, it is a new ontology applied to a new corpus. In this study, we propose to project already existing ontologies in a formalism that can extend their expressive power.

3. ABSTRACT REPRESENTATION FOR MULTI-INTENT SLU

As discussed in the introduction, one goal of this study is to go further than the simple mono-intent, flat slot/value semantic annotation scheme. We are interested in 3 features that have to be handled to build human-machine dialog systems and

which are often missing in benchmark corpora: dialog context, multi-intent, categorical slots (slots not associated to a given word span). We propose in this study the ARMILU encoding for representing SLU annotations into a graph-based formalism that can handle these 3 features. ARMILU, for *Abstract Representation for Multi-Intent spoken Language Understanding*, was inspired by semantic representations such as AMR (*Abstract Meaning Representation* [9]) encoded with the *Penman* format [10], a linear representation using parentheses and variables to express relations and concepts.

The goal of ARMILU is not to propose a new generic ontology representing semantics in a dialog context as proposed by [8], but to directly project the origin annotations without any modifications to the ontology. ARMILU only defines a graph structure among 3 kinds of elements: *intents*, *slots* and *types*. In our model, slots are relations between intents and values of a given type. Among possible types one can find *name*, *dates*, or any kind of named entities; values can be associated to spans in the utterance or to categorical values such as *yes* or *no*. The ontology of intents, slots and types is not part of ARMILU and is taken directly from existing corpora.

The main advantage of ARMILU is to be able to represent existing flat SLU annotations of standard benchmark such as ATIS or SNIPS as well as more complex ones such as MTOP [6] or MultiWOZ [11]. No extra human annotations is needed to translate existing SLU annotations in ARMILU.

3.1. Definition

Let U be an utterance composed of two intents (`Intent_1` and `Intent_2`). The first intent is associated to a span-based slot `slot_a` of type `Type_x`, whose value is explicitly extracted from a span of two tokens in the utterance (`token_1` and `token_2`). The second intent is associated to a categorical slot `slot_b` whose value comes from a closed set of predefined labels `Label_y` without an explicit correspondence with the tokens.

Its ARMILU representation would be as represented in figure 1. `:op#` symbols are used to mark the enumeration for the different intents or the different tokens. Slots are relations in the graph between an intent and a value or between an intent and a category. Mono-intent utterances can be represented similarly with the intent being at the initial state of the graph. The variables allow to encode potential coreferences within a turn such as in the central part of figure 1 where two slots refer to the same entity, or between turns to encode dialog context. An example of such a graph structure representation could be instanciated as in the right part of figure 1.

3.2. Generating ARMILU representation

The recent success of AMR prediction models has been obtained thanks to *seq2seq* models that directly learn to generate the Penman linear representation of these graphs with powerful pretrained generative models. One of the advantage of adopting the Penman projection is that we can rely on such libraries that have already been designed to efficiently produce consistent semantic graphs. The Spring tools [12] and its X-AMR variant [13] are based on the pre-trained mBART model [14]. The AMRlib library (https://github.com/bjascob/amrlib) is based on T5 [15] and we have modified it to be able to use the multilingual version mT5 [16]. Both are agnostic to the relations nomenclature and the variables naming, and include a Penman syntax validation module.

4. EXPERIMENTS ON THE MULTIWOZ CORPUS

To illustrate the potential of our ARMILU semantic representation we choose to work with the MultiWOZ corpus [11] which is a large-scale multi-domain English dataset frequently used in Dialog State Tracking (DST), dialog policy, and dialog generation tasks. It was collected using a Wizard of Oz approach by *crowd-sourcing* for 8 domains (*Train*, *Taxi*, *Hotel*, *Restaurant*, *Attraction*, *Hospital*, *Bus* and *Police*). Although this is a text corpus, we believe it is relevant for SLU studies as the style of the dialogs is clearly oral, moreover a speech version of the corpus has been developed for the last DSTC shared task and should be available soon and could be used order to predict directly semantic annotation from speech with end-to-end SLU methods [17, 18].

We chose to work on version *MultiWOZ2.3* which turns out to be the most stable one at the moment. Although the SLU and DST tasks are obviously related, the goal of DST is not to produce a structured semantic representation but rather a flat list of concept/values representing the current state of the validated facts through the dialog. *MultiWOZ* however contains semantic annotations at the utterance level made of semantic frames[1] composed of an intent (the concatenation of the domain and the dialog act) with a set of arguments in the form of slot-value pairs. The annotation can be linked to word span information in the turns for the slots with open values (such as names or dates) or to the whole turn for normalized values (also called *categorical slots*) such as *yes*, *no*, *expensive*, ... In the first example of table 1, the frame (corresponding to the Intent in our general approach) consists of the domain `Hotel` and the act `Inform` with the slots `Parking` and `Price`. The value `yes` has no exact match, while the span is provided for the second value. Two different intents from two domains are present in the second example. These examples highlight the fact that MultiWoz annotation is intrinsically contextual. The association between "the postcode for *that*" and the `Attraction` domain cannot be solved by taking into account the single current utterance.

As we can see in these examples, the semantic annotation scheme of *MultiWOZ* is flat, each frame is simply represented by an intent followed by a set of slot/value elements. However due to the fact that the same word span can appear in several

[1] Semantic frames are denoted as "`dialog_act`" in MultiWOZ *json* annotation files but we prefer to adopt the conventional SLU terminology.

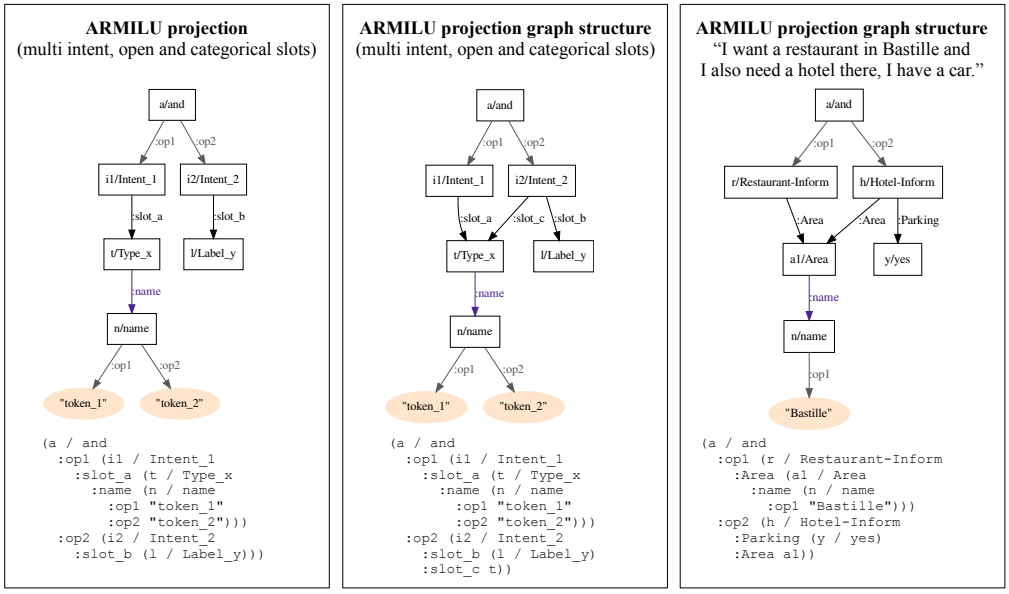

Fig. 1. ARMILU representation for multi-intent and categorical slots

frames, that multiple intents can occur in one turn, that there are categorical slots not associated to a given word span and finally that dialog context is needed to remove frame ambiguities, we believe that *MultiWOZ* is a good candidate for evaluating our ARMILU semantic representation. Moreover, the current flat frame semantic annotation can be projected automatically into the ARMILU format, as presented in the right part of figure 1, therefore no additional human annotation effort is needed. Note that in MultiWoz there is no explicit notion of *Type* for values, hence for the ARMILU projection, we duplicate the *Slot* label as the *Type* label in the nodes. One of the contributions of this work is to provide to the SLU research community a version of *MultiWOZ* in the ARMILU format [2] that can be used easily for SLU tasks, this was not the case with the original annotations, dedicated exclusively to the DST task. Table 2 presents the characteristics of the *MultiWOZ* dataset we have annotated and that we use in our experiments.

utterances	semantic annotation
◇ I need to make sure it's cheap and I have a car	`"Hotel-Inform": [["Parking", "yes"], ["Price", "cheap", 7, 7]]`
◇ I'll need the address for one that does have wifi please	`"Hotel-Inform": [["Internet", "yes"]], "Hotel-Request": [["Addr", "?"]]`

Table 1. Examples of utterance annotations in MultiWOZ2.3

[2]available at: https://gitlab.lis-lab.fr/armilu/multiwoz2.3_armilu

characteristics	MultiWOZ
#user turns (train/test)	56775 / 7372
#words (train/test)	765080 / 126178
#slots (open/categorical)	27 / 22
#Intents	32

Table 2. MultiWOZ dataset description

5. RESULTS

We compare two models in these experiments: a sequence tagging model using a flat annotation of the MultiWOZ dataset and a graph generation model producing directly our ARMILU representation from text with a seq2seq paradigm.

Sequence tagging models. In addition to the ARMILU projection, we also have produced a flat B,I,O annotation of *MultiWOZ* in order to compare standard SLU sequence tagging approaches with the ARMILU graph generation method. Following previous work on the Convlab [19] platform, we use a projection in order to deal with multi-intent annotations suitable for use with a BERT pretrained language model representation making use of the `[CLS]` token for handling intents. We systematically predict at the level of the `[CLS]` token the set of intents of the utterance. The slots that have an associated span are identified at the token level while the categorical slots are directly predicted at the `[CLS]` level.

For the inference model we chose a multilabel sequence tagging model by *finetuning* a pretrained language model on the task. The models are learned with a batch size equal to 100 and a maximum of 50 epochs. We fixed a threshold of 0.55 (optimised on the dev corpus) for classification probabilities for each label. Two pretrained language models are compared

	Negation		Implicit values		Value "?"		Multi-label	
nb of samples	**+ (470)**	**– (6902)**	**+ (574)**	**– (6798)**	**+ (1499)**	**– (5837)**	**+ (3417)**	**– (3667)**
BIO mBERT	62.3	88.2	55.7	91.4	76.3	89.1	80.0	92.1
	[61.2-63.4]	*[87.4-88.9]*	*[54.6-56.8]*	*[90.7-92.0]*	*[75.3-77.3]*	*[88.4-89.8]*	*[79.1-80.9]*	*[91.5-92.7]*
BIO mT5	56.0	86.6	48.4	90.6	71.8	88.0	77.9	90.7
	[54.7-57.1]	*[85.8-87.4]*	*[47.3-49.5]*	*[89.9-91.2]*	*[70.8-72.8]*	*[87.2-88.7]*	*[76.9-78.8]*	*[90.0-91.3]*
ARMILU mT5	**68.3**	**89.0**	**59.4**	**91.9**	**78.9**	**89.9**	**81.7**	**93.2**
	[67.2-69.4]	*[88.3-89.7]*	*[58.3-60.5]*	*[91.3-92.5]*	*[78.0-79.8]*	*[89.2-90.6]*	*[80.8-82.6]*	*[92.6-93.8]*

Table 3. Global accuracy with confidence intervals at the utterance level on different partitions of the MultiWOZ test set

in our experiments: mBERT (**BIO mBERT**) and the encoder-only part of mT5 (**BIO mT5**).

Graph generation models. For directly generating the graph ARMILU representation, we used the AMRlib library with the mT5 model (mT5-base version), over 20 epochs, with a batch size of 3 and a maximum sequence length of 100. This model is called **ARMILU mT5** in our experiments.

As mentioned in section 4, the MultiWOZ corpus annotation is intrinsically contextual. In order to improve the current utterance interpretation, we augment the input of the models with the two previous turns (with anchors to denlimit [USER] and [SYSTEM] turns) both for BIO and ARMILU graph generation model. SLU performance is usually evaluated through classification accuracy at the intent level and F1 score at the slot/value level. As the semantic model accepts several intents per utterance in MultiWOZ, we consider the list of *Intent(Slot, Value)* elements following the same rules as the *seqeval* [3] library. Finally, a global accuracy evaluation at the turn level is also estimated where the whole semantic structure of a given turn has to be completely correct.

model	**int.**	**(S,V)**	**int.(S,V)**	**turn**
***BIO BERTNLU* [20]**	*89.0*	*90.7*	*89.7*	*78.3*
***BIO BERTNLU* (ours)**	95.5	94.6	**94.1**	86.6
BIO mBERT	96.1	94.6	93.9	86.3
BIO mt5 enc.	95.2	94.0	93.1	84.7
ARMILU mt5	**96.2**	**94.7**	**94.1**	**87.5**

Table 4. Comparison of BIO and ARMILU models on MultiWOZ: F1 at intent level (**int.**), slot+value (**(S,V)**), intent+slot+value, and global accuracy (**turn**)

Table 4 compares the results obtained on our BIO classification models and our ARMILU generation model. As we can see, the **ARMILU mt5** model brings a small gain in performance over the mBERT and mt5-encoder classification models. Few results have been published for NLU on MultiWoz2.3. When delivering the corpus the authors in [20] provided performances of the BERTNLU model [19] but the evaluation protocole is not exactly comparable. We re-ran experiments with the model available in [21] and obtained the results of the second line with the same evaluation scripts.

[3] https://github.com/chakki-works/seqeval.git

Beyond global evaluation we are interested in assessing the behaviour of models when facing more complex utterances. Hence, we have partitioned the MultiWOZ test corpus into several sets according to several criteria related to potential difficulty levels in the automatic processing.

The first column highlights utterances that have negation marks, (*not* or *n't*), since negative sentences are more complex to process than affirmative ones. The second one separates utterances with only categorical values (`dontcare`, `yes`, `no`), or no value (`none`) from those which contain values with an associated span. The hypothesis is that these concepts are more difficult to handle than span-based ones. The third set concerns interrogative sentences (value `?`), the fourth partition distinguishes utterances that have multiple labels from single-label ones. Performance in terms of global accuracy is reported in table 3 for our BIO models and our ARMILU model with confidence intervals obtained with the Wilson binomial proportion confidence interval at 95%.

Results show that the ARMILU projection with the graph generation model is significantly better on all the potentially difficult configurations. The improvement is increased for implicit values, or for utterances with negations, confirming that abstract representations can help going beyond classical span based slot/value utterances.

These observations confirm the highest expressive power of abstract representations and show that it is possible to process simple and complex phenomena with a single inference model. We will try in future work to enrich the annotations in order to generate graph semantic interpretations and push the potential of the approach.

6. CONCLUSIONS

We have proposed a new paradigm to project semantic annotations for SLU into an abstract graph representation. Our ARMILU representation exploits the Penman representation also used for Abstract Meaning Representation. The advantage of this structure is to be able to take advantage of powerful seq2seq generative models to produce structured semanting interpretations. This approach yields improved results on the MultiWoz corpus with higher improvements on more complex utterances, and paves the way to handle potentially more complex semantic annotations for dialogue.

7. REFERENCES

[1] Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang, "Multi-domain joint semantic frame parsing using bi-directional rnn-lstm.," in *Proc. Interspeech'16*, 2016.

[2] Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen, "Slot-gated modeling for joint slot filling and intent prediction," in *Proc. NAACL'18: HLT*, 2018.

[3] Qian Chen, Zhu Zhuo, and Wen Wang, "Bert for joint intent classification and slot filling," *arXiv preprint arXiv:1902.10909*, 2019.

[4] Frédéric Béchet and Christian Raymond, "Is ATIS too shallow to go deeper for benchmarking Spoken Language Understanding models?," in *Proc. Interspeech'18*, Hyderabad, India, Sept. 2018.

[5] Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis, "Semantic parsing for task oriented dialog using hierarchical representations," *arXiv preprint arXiv:1810.07942*, 2018.

[6] Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad, "Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark," 2020.

[7] Armen Aghajanyan, Jean Maillard, Akshat Shrivastava, Keith Diedrick, Mike Haeger, Haoran Li, Yashar Mehdad, Ves Stoyanov, Anuj Kumar, Mike Lewis, et al., "Conversational semantic parsing," *arXiv preprint arXiv:2009.13655*, 2020.

[8] Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss, "Dialogue-AMR: Abstract Meaning Representation for dialogue," in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 684–695, European Language Resources Association.

[9] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider, "Abstract Meaning Representation for Sembanking," in *Proc. LAWID'13*, 2013, pp. 178–186.

[10] Robert T. Kasper, "A flexible interface for linking applications to Penman's sentence generator," in *Proc. Speech and Natural Language Workshop SNL'89*, 1989.

[11] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić, "MultiWOZ – a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling," in *Proc. EMNLP'18*, 2018.

[12] Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli, "One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline," in *Proc. AAAI AI'21*, 2021.

[13] Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam, "Multilingual AMR Parsing with Noisy Knowledge Distillation," in *FACL: EMNLP'21*, 2021.

[14] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *CoRR'20*, 2020.

[15] Colin Raffel, Noam Shazeer, Katherine Lee, Sharan Narang, Michael Matena, Yanqui Zhou, Wei Li, and Liu Peter J., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *JMLR'20*, vol. 21, 2020.

[16] Linting Xue, Noa Constant, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in *Proc. NAACL'21*, 2021.

[17] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[18] Siddhant Arora, Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda, Yifan Peng, Yuekai Zhang, Sujay Kumar, Karthik Ganesan, Brian Yan, et al., "Espnet-slu: Advancing spoken language understanding through espnet," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7167–7171.

[19] Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Xiang Li, Yaoqin Zhang, Zheng Zhang, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, et al., "Convlab: Multi-domain end-to-end dialog system platform," *arXiv:1904.08637*, 2019.

[20] Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang, "Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation," *arXiv:2010.05594*, 2020.

[21] Convlab-2 BERTNLU, ," https://github.com/thu-coai/ConvLab-2/tree/master/convlab2/nlu/jointBERT/multiwoz.