Hierarchical Multi-Task Natural Language Understanding for Cross-domain Conversational AI: HERMIT NLU

Andrea Vanzo
Interaction Lab
Heriot-Watt University
a.vanzo@hw.ac.uk

Emanuele Bastianelli
Interaction Lab
Heriot-Watt University
e.bastianelli@hw.ac.uk

Oliver Lemon
Interaction Lab
Heriot-Watt University
o.lemon@hw.ac.uk

Abstract

We present a new neural architecture for wide-coverage Natural Language Understanding in Spoken Dialogue Systems. We develop a hierarchical multi-task architecture, which delivers a multi-layer representation of sentence meaning (i.e., Dialogue Acts and Frame-like structures). The architecture is a hierarchy of self-attention mechanisms and BiLSTM encoders followed by CRF tagging layers. We describe a variety of experiments, showing that our approach obtains promising results on a dataset annotated with Dialogue Acts and Frame Semantics. Moreover, we demonstrate its applicability to a different, publicly available NLU dataset annotated with domain-specific intents and corresponding semantic roles, providing overall performance higher than state-of-the-art tools such as RASA, Dialogflow, LUIS, and Watson. For example, we show an average 4.45% improvement in entity tagging F-score over Rasa, Dialogflow and LUIS.

1 Introduction

Research in Conversational AI (also known as Spoken Dialogue Systems) has applications ranging from home devices to robotics, and has a growing presence in industry. A key problem in real-world Dialogue Systems is Natural Language Understanding (NLU) – the process of extracting structured representations of meaning from user utterances. In fact, the effective extraction of semantics is an essential feature, being the entry point of any Natural Language interaction system. Apart from challenges given by the inherent complexity and ambiguity of human language, other challenges arise whenever the NLU has to operate over multiple domains. In fact, interaction patterns, domain, and language vary depending on the device the user is interacting with. For example, chit-chatting and instruction-giving for executing an action are different processes in terms of language, domain, syntax and interaction schemes involved. And what if the user combines two interaction domains: "*play some music, but first what's the weather tomorrow*"?

In this work, we present HERMIT, a HiERarchical MultI-Task Natural Language Understanding architecture[1], designed for effective semantic parsing of domain-independent user utterances, extracting meaning representations in terms of high-level intents and frame-like semantic structures. With respect to previous approaches to NLU for SDS, HERMIT stands out for being a cross-domain, multi-task architecture, capable of recognising multiple intents/frames in an utterance. HERMIT also shows better performance with respect to current state-of-the-art commercial systems. Such a novel combination of requirements is discussed below.

Cross-domain NLU A cross-domain dialogue agent must be able to handle heterogeneous types of conversation, such as chit-chatting, giving directions, entertaining, and triggering domain/task actions. A domain-independent and rich meaning representation is thus required to properly capture the intent of the user. Meaning is modelled here through three layers of knowledge: dialogue acts, frames, and frame arguments. Frames and arguments can be in turn mapped to domain-dependent intents and slots, or to Frame Semantics' (Fillmore, 1976) structures (i.e. semantic frames and frame elements, respectively), which allow handling of heterogeneous domains and language.

Multi-task NLU Deriving such a multi-layered meaning representation can be approached through a multi-task learning approach. Multi-task learning has found success in several NLP

[1] https://gitlab.com/hwu-ilab/hermit-nlu

Proceedings of the SIGDial 2019 Conference, pages 254–263
Stockholm, Sweden, 11-13 September 2019.

problems (Hashimoto et al., 2017; Strubell et al., 2018), especially with the recent rise of Deep Learning. Thanks to the possibility of building complex networks, handling more tasks at once has been proven to be a successful solution, provided that some degree of dependence holds between the tasks. Moreover, multi-task learning allows the use of different datasets to train subparts of the network (Sanh et al., 2018). Following the same trend, HERMIT is a hierarchical multi-task neural architecture which is able to deal with the three tasks of tagging dialogue acts, frame-like structures, and their arguments in parallel. The network, based on self-attention mechanisms, seq2seq bi-directional Long-Short Term Memory (BiLSTM) encoders, and CRF tagging layers, is hierarchical in the sense that information output from earlier layers flows through the network, feeding following layers to solve downstream dependent tasks.

Multi-dialogue act and -intent NLU Another degree of complexity in NLU is represented by the granularity of knowledge that can be extracted from an utterance. Utterance semantics is often rich and expressive: approximating meaning to a single user intent is often not enough to convey the required information. As opposed to the traditional single-dialogue act and single-intent view in previous work (Guo et al., 2014; Liu and Lane, 2016; Hakkani-Tur et al., 2016), HERMIT operates on a meaning representation that is multi-dialogue act and multi-intent. In fact, it is possible to model an utterance's meaning through multiple dialogue acts and intents at the same time. For example, the user would be able both to request tomorrow's weather and listen to his/her favourite music with just a single utterance.

A further requirement is that for practical application the system should be **competitive with state-of-the-art**: we evaluate HERMIT's effectiveness by running several empirical investigations. We perform a robust test on a publicly available NLU-Benchmark (NLU-BM) (Liu et al., 2019) containing 25K cross-domain utterances with a conversational agent. The results obtained show a performance higher than well-known off-the-shelf tools (i.e., Rasa, DialogueFlow, LUIS, and Watson). The contribution of the different network components is then highlighted through an ablation study. We also test HERMIT on the smaller Robotics-Oriented MUltitask Language Understanding (ROMULUS) corpus, annotated with Dialogue Acts and Frame Semantics. HERMIT produces promising results for the application in a real scenario.

2 Related Work

Much research on Natural (or Spoken, depending on the input) Language Understanding has been carried out in the area of Spoken Dialogue Systems (Chen et al., 2017), where the advent of statistical learning has led to the application of many data-driven approaches (Lemon and Pietquin, 2012). In recent years, the rise of deep learning models has further improved the state-of-the-art. Recurrent Neural Networks (RNNs) have proven to be particularly successful, especially uni- and bi-directional LSTMs and Gated Recurrent Units (GRUs). The use of such deep architectures has also fostered the development of joint classification models of intents and slots. Bi-directional GRUs are applied in (Zhang and Wang, 2016), where the hidden state of each time step is used for slot tagging in a seq2seq fashion, while the final state of the GRU is used for intent classification. The application of attention mechanisms in a BiLSTM architecture is investigated in (Liu and Lane, 2016), while the work of (Chen et al., 2016) explores the use of memory networks (Sukhbaatar et al., 2015) to exploit encoding of historical user utterances to improve the slot-filling task. Seq2seq with self-attention is applied in (Li et al., 2018), where the classified intent is also used to guide a special gated unit that contributes to the slot classification of each token.

One of the first attempts to jointly detect domains in addition to intent-slot tagging is the work of (Guo et al., 2014). An utterance syntax is encoded through a Recursive NN, and it is used to predict the joined domain-intent classes. Syntactic features extracted from the same network are used in the per-word slot classifier. The work of (Hakkani-Tur et al., 2016) applies the same idea of (Zhang and Wang, 2016), this time using a context-augmented BiLSTM, and performing domain-intent classification as a single joint task. As in (Chen et al., 2016), the history of user utterances is also considered in (Bapna et al., 2017), in combination with a dialogue context encoder. A two-layer hierarchical structure made of a combination of BiLSTM and BiGRU is used

for joint classification of domains and intents, together with slot tagging. (Rastogi et al., 2018) apply multi-task learning to the dialogue domain. Dialogue state tracking, dialogue act and intent classification, and slot tagging are jointly learned. Dialogue states and user utterances are encoded to provide hidden representations, which jointly affect all the other tasks.

Many previous systems are trained and compared over the ATIS (Airline Travel Information Systems) dataset (Price, 1990), which covers only the flight-booking domain. Some of them also use bigger, not publicly available datasets, which appear to be similar to the NLU-BM in terms of number of intents and slots, but they cover no more than three or four domains. Our work stands out for its more challenging NLU setting, since we are dealing with a higher number of domains/scenarios (18), intents (64) and slots (54) in the NLU-BM dataset, and dialogue acts (11), frames (58) and frame elements (84) in the ROMULUS dataset. Moreover, we propose a multi-task hierarchical architecture, where each layer is trained to solve one of the three tasks. Each of these is tackled with a seq2seq classification using a CRF output layer, as in (Sanh et al., 2018).

The NLU problem has been studied also on the Interactive Robotics front, mostly to support basic dialogue systems, with few dialogue states and tailored for specific tasks, such as semantic mapping (Kruijff et al., 2007), navigation (Kollar et al., 2010; Bothe et al., 2018), or grounded language learning (Chai et al., 2016). However, the designed approaches, either based on formal languages or data-driven, have never been shown to scale to real world scenarios. The work of (Hatori et al., 2018) makes a step forward in this direction. Their model still deals with the single 'pick and place' domain, covering no more than two intents, but it is trained on several thousands of examples, making it able to manage more unstructured language. An attempt to manage a higher number of intents, as well as more variable language, is represented by the work of (Bastianelli et al., 2016) where the sole Frame Semantics is applied to represent user intents, with no Dialogue Acts.

3 Jointly parsing dialogue acts and frame-like structures

The identification of Dialogue Acts (henceforth DAs) is required to drive the dialogue manager to the next dialogue state. General frame structures (FRs) provide a reference framework to capture user intents, in terms of required or desired actions that a conversational agent has to perform. Depending on the level of abstraction required by an application, these can be interpreted as more domain-dependent paradigms like *intent*, or to shallower representations, such as *semantic frames*, as conceived in FrameNet (Baker et al., 1998). From this perspective, semantic frames represent a versatile abstraction that can be mapped over an agent's capabilities, allowing also the system to be easily extended with new functionalities without requiring the definition of new ad-hoc structures. Similarly, frame arguments (ARs) act as *slots* in a traditional intent-slots scheme, or to *frame elements* for semantic frames.

In our work, the whole process of extracting a complete semantic interpretation as required by the system is tackled with a multi-task learning approach across DAs, FRs, and ARs. Each of these tasks is modelled as a seq2seq problem, where a task-specific label is assigned to each token of the sentence according to the IOB2 notation (Sang and Veenstra, 1999), with "B-" marking the Beginning of the chunk, "I-" the tokens Inside the chunk while "O-" is assigned to any token that does not belong to any chunk. Task labels are drawn from the set of classes defined for DAs, FRs, and ARs. Figure 1 shows an example of the tagging layers over the sentence *Where can I find Starbucks?*, where Frame Semantics has been selected as underlying reference theory.

3.1 Architecture description

The central motivation behind the proposed architecture is that there is a dependence among the three tasks of identifying DAs, FRs, and ARs. The relationship between tagging frame and arguments appears more evident, as also developed in theories like Frame Semantics – although it is defined independently by each theory. However, some degree of dependence also holds between the DAs and FRs. For example, the FrameNet semantic frame *Desiring*, expressing a desire of the user for an event to occur, is more likely to be used in the context of an INFORM DA, which indicates the state of notifying the agent with an information, other than in an INSTRUCTION. This is clearly visible in interactions like "*I'd like a cup of hot chocolate*" or "*I'd like to find a shoe shop*", where

	Where	*can*	*I*	*find*	*Starbucks*	*?*
DAs	B-REQ_INFO	I-REQ_INFO	I-REQ_INFO	I-REQ_INFO	I-REQ_INFO	O
FRs	B-*Locating*	I-*Locating*	I-*Locating*	I-*Locating*	I-*Locating*	O
ARs	O	O	B-COGNIZER	B-LEXICAL_UNIT	B-ENTITY	O

Figure 1: Dialogue Acts (DAs), Frames (FRs – here semantic frames) and Arguments (ARs – here frame elements) IOB2 tagging for the sentence *Where can I find Starbucks?*

the user is actually notifying the agent about a desire of hers/his.

In order to reflect such inter-task dependence, the classification process is tackled here through a hierarchical multi-task learning approach. We designed a multi-layer neural network, whose architecture is shown in Figure 2, where each layer is trained to solve one of the three tasks, namely labelling dialogue acts (DA layer), semantic frames (FR layer), and frame elements (AR layer). The layers are arranged in a hierarchical structure that allows the information produced by earlier layers to be fed to downstream tasks.

The network is mainly composed of three BiLSTM (Schuster and Paliwal, 1997) encoding layers. A sequence of input words is initially converted into an embedded representation through an ELMo embeddings layer (Peters et al., 2018), and is fed to the DA layer. The embedded representation is also passed over through shortcut connections (Hashimoto et al., 2017), and concatenated with both the outputs of the DA and FR layers. Self-attention layers (Zheng et al., 2018) are placed after the DA and FR BiLSTM encoders. Where w_t is the input word at time step t of the sentence $\mathbf{w} = (w_1, ..., w_T)$, the architecture can be formalised by:

$$
\begin{aligned}
e_t = ELMo(w_t),\ s_t^{DA} = BiLSTM(e_t) \\
a_t^{DA} = SelfAtt(s_t^{DA},\ \mathbf{s}^{DA}), \\
s_t^{FR} = BiLSTM(e_t \oplus a_t^{DA}), \\
a_t^{FR} = SelfAtt(s_t^{FR},\ \mathbf{s}^{FR}), \\
s_t^{AR} = BiLSTM(e_t \oplus a_t^{FR})
\end{aligned}
$$

where $\oplus$ represents the vector concatenation operator, e_t is the embedding of the word at time t, and $\mathbf{s}^L = (s_1^L, ..., s_T^L)$ is the embedded sequence output of each L layer, with $L = \{DA, FR, AR\}$. Given an input sentence, the final sequence of labels $\mathbf{y}^L$ for each task is computed through a CRF tagging layer, which operates on the output of the DA and FR self-attention, and of the AR BiLSTM embedding, so that:

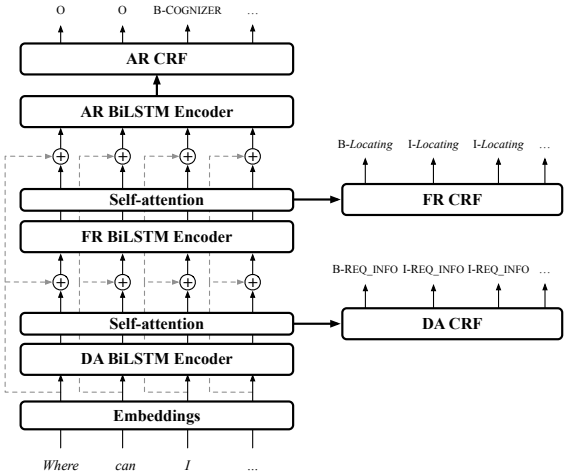

Figure 2: HERMIT Network topology

$$
\begin{aligned}
\mathbf{y}^{DA} = CRF^{DA}(\mathbf{a}^{DA}),\ \mathbf{y}^{FR} = CRF^{FR}(\mathbf{a}^{FR}) \\
\mathbf{y}^{AR} = CRF^{AR}(\mathbf{s}^{AR}),
\end{aligned}
$$

where $\mathbf{a}^{DA}$, $\mathbf{a}^{FR}$ are attended embedded sequences. Due to shortcut connections, layers in the upper levels of the architecture can rely both on direct word embeddings as well as the hidden representation a_t^L computed by a previous layer. Operationally, the latter carries task specific information which, combined with the input embeddings, helps in stabilising the classification of each CRF layer, as shown by our experiments. The network is trained by minimising the sum of the individual negative log-likelihoods of the three CRF layers, while at test time the most likely sequence is obtained through the Viterbi decoding over the output scores of the CRF layer.

4 Experimental Evaluation

In order to assess the effectiveness of the proposed architecture and compare against existing off-the-shelf tools, we run several empirical evaluations.

4.1 Datasets

We tested the system on two datasets, different in size and complexity of the addressed language.

NLU-Benchmark dataset The first (publicly available) dataset, NLU-Benchmark (NLU-BM), contains 25,716 utterances annotated with targeted *Scenario*, *Action*, and involved *Entities*. For example, "*schedule a call with Lisa on Monday morning*" is labelled to contain a `calendar` scenario, where the `set_event` action is instantiated through the entities [`event_name`: *a call with Lisa*] and [`date`: *Monday morning*]. The Intent is then obtained by concatenating scenario and action labels (e.g., `calendar_set_event`). This dataset consists of multiple home assistant task domains (e.g., scheduling, playing music), chit-chat, and commands to a robot (Liu et al., 2019).[2]

	NLU-BM	NLU-BM (reduced)
Sentences	25715	11020
Sentences length	7.06	6.84
Scenario labels set	18	18
Action labels set	54	51
Intent labels set	68	64
Entity labels set	56	54
Number of intent	25715	11020
Number of entities	20597	9130
Intents/sentence	1	1
Entities/sentence	0.8	0.83

Table 1: Statistics of the NLU-Benchmark dataset (Liu et al., 2019).

ROMULUS dataset The second dataset, ROMULUS, is composed of 1,431 sentences, for each of which dialogue acts, semantic frames, and corresponding frame elements are provided. This dataset is being developed for modelling user utterances to open-domain conversational systems for robotic platforms that are expected to handle different interaction situations/patterns – e.g., chit-chat, command interpretation. The corpus is composed of different subsections, addressing heterogeneous linguistic phenomena, ranging from imperative instructions (e.g., "*enter the bedroom slowly, turn left and turn the lights off* ") to complex requests for information (e.g., "*good morning I want to buy a new mobile phone is there any shop nearby?*") or open-domain chit-chat (e.g., "*nope thanks let's talk about cinema*"). A considerable number of utterances in the dataset is collected through Human-Human Interaction studies in robotic domain (≈70%), though a small portion has been synthetically generated for balancing the frame distribution.

[2]Available at https://github.com/xliuhw/NLU-Evaluation-Data.

	ROMULUS dataset
Sentences	1431
Sentences length	7.24
Dialogue act labels set	11
Frame labels set	58
Frame element labels set	84
Number of dialogue acts	1906
Number of frames	2013
Number of frame elements	5059
Dialogue act/sentence	1.33
Frames/sentence	1.41
Frame elements/sentence	3.54

Table 2: Statistics of the ROMULUS dataset.

Note that while the NLU-BM is designed to have at most one intent per utterance, sentences are here tagged following the IOB2 sequence labelling scheme (see example of Figure 1), so that multiple dialogue acts, frames, and frame elements can be defined at the same time for the same utterance. For example, three dialogue acts are identified within the sentence [*good morning*]$_{\text{OPENING}}$ [*I want to buy a new mobile phone*]$_{\text{INFORM}}$ [is there any shop nearby?]$_{\text{REQ\_INFO}}$. As a result, though smaller, the ROMULUS dataset provides a richer representation of the sentence's semantics, making the tasks more complex and challenging. These observations are highlighted by the statistics in Table 2, that show an average number of dialogue acts, frames and frame elements always greater than 1 (i.e., 1.33, 1.41 and 3.54, respectively).

4.2 Experimental setup

All the models are implemented with Keras (Chollet et al., 2015) and Tensorflow (Abadi et al., 2015) as backend, and run on a Titan Xp. Experiments are performed in a 10-fold setting, using one fold for tuning and one for testing. However, since HERMIT is designed to operate on dialogue acts, semantic frames and frame elements, the best hyperparameters are obtained over the ROMULUS dataset via a grid search using early stopping, and are applied also to the NLU-BM models.[3] This guarantees fairness towards other systems, that do not perform any fine-tuning on the training data. We make use of pre-trained 1024-dim ELMo embeddings (Peters et al., 2018) as word vector representations without re-training the weights.

[3]Notice that in the NLU-BM experiments only the number of epochs is tuned, using 10% of the training data.

4.3 Experiments on the NLU-Benchmark

This section shows the results obtained on the NLU-Benchmark (NLU-BM) dataset provided by (Liu et al., 2019), by comparing **HERMIT** to off-the-shelf NLU services, namely: **Rasa**[4], **Dialogflow**[5], **LUIS**[6] and **Watson**[7]. In order to apply HERMIT to NLU-BM annotations, these have been aligned so that Scenarios are treated as DAs, Actions as FRs and Entities as ARs.

To make our model comparable against other approaches, we reproduced the same folds as in (Liu et al., 2019), where a resized version of the original dataset is used. Table 1 shows some statistics of the NLU-BM and its reduced version. Moreover, micro-averaged Precision, Recall and F1 are computed following the original paper to assure consistency. TP, FP and FN of intent labels are obtained as in any other multi-class task. An entity is instead counted as TP if there is an overlap between the predicted and the gold span, and their labels match.

Experimental results are reported in Table 3. The statistical significance is evaluated through the Wilcoxon signed-rank test. When looking at the intent F1, HERMIT performs significantly better than Rasa $[Z = -2.701, p = .007]$ and LUIS $[Z = -2.807, p = .005]$. On the contrary, the improvements w.r.t. Dialogflow $[Z = -1.173, p = .241]$ do not seem to be significant. This is probably due to the high variance obtained by Dialogflow across the 10 folds. Watson is by a significant margin the most accurate system in recognising intents $[Z = -2.191, p = .028]$, especially due to its Precision score.

The hierarchical multi-task architecture of HERMIT seems to contribute strongly to entity tagging accuracy. In fact, in this task it performs significantly better than Rasa $[Z = -2.803, p = .005]$, Dialogflow $[Z = -2.803, p = .005]$, LUIS $[Z = -2.803, p = .005]$ and Watson $[Z = -2.805, p = .005]$, with improvements from 7.08 to 35.92 of F1.[9]

Following (Liu et al., 2019), we then evaluated a metric that combines intent and entities, computed by simply summing up the two confusion matrices (Table 4). Results highlight the contribution of the entity tagging task, where HERMIT outperforms the other approaches. Paired-samples t-tests were conducted to compare the HERMIT combined F1 against the other systems. The statistical analysis shows a significant improvement over Rasa $[Z = -2.803, p = .005]$, Dialogflow $[Z = -2.803, p = .005]$, LUIS $[Z = -2.803, p = .005]$ and Watson $[Z = -2.803, p = .005]$.

4.3.1 Ablation study

In order to assess the contributions of the HERMIT's components, we performed an ablation study. The results are obtained on the NLU-BM, following the same setup as in Section 4.3.

Results are shown in Table 5. The first row refers to the complete architecture, while –SA shows the results of HERMIT without the self-attention mechanism. Then, from this latter we further remove shortcut connections (– SA/CN) and CRF taggers (– SA/CRF). The last row (– SA/CN/CRF) shows the results of a simple architecture, without self-attention, shortcuts, and CRF. Though not significant, the contribution of the several architectural components can be observed. The contribution of self-attention is distributed across all the tasks, with a small inclination towards the upstream ones. This means that while the entity tagging task is mostly lexicon independent, it is easier to identify pivoting keywords for predicting the intent, e.g. the verb *"schedule"* triggering the `calendar_set_event` intent. The impact of shortcut connections is more evident on entity tagging. In fact, the effect provided by shortcut connections is that the information flowing throughout the hierarchical architecture allows higher layers to encode richer representations (i.e., original word embeddings + latent semantics from the previous task). Conversely, the presence of the CRF tagger affects mainly the lower levels of the hierarchical architecture. This is not probably due to their position in the hierarchy, but to the way the tasks have been designed. In fact, while the span of an entity is expected to cover few tokens, in intent recognition (i.e., a combination of Scenario and Action recognition) the span always covers all the tokens of an utterance. CRF therefore preserves consistency of IOB2 sequences structure. However, HERMIT seems to be the most stable ar-

[4] https://rasa.com/

[5] https://dialogflow.com/

[6] https://www.luis.ai/

[7] https://www.ibm.com/watson

[9] Results for Watson are shown for the non-contextual training. Due to Watson limitations, i.e. 2000 training examples for contextual training, we could not run the whole test in such configuration. For fairness, we report results made on 8 random samplings of 2000/1000 train/test examples a each (F1): Intent=72.64 ± 7.46, Slots=77.01 ± 10.65, Combined=74.85 ± 7.54

	Intent			Entity		
	P	R	F1	P	R	F1
Rasa	86.31±1.07	86.31±1.07	86.31±1.07	85.93±1.05	69.40±1.66	76.78±1.27
Dialogflow	86.97±2.02	85.87±2.33	86.42±2.18	78.21±3.35	70.85±4.70	74.30±3.74
LUIS	85.53±1.14	85.51±1.15	85.52±1.15	83.69±1.31	72.46±2.05	77.66±1.45
Watson[8]	**88.41**±0.68	**88.08**±0.74	**88.24**±0.70	35.39±0.93	78.70±2.01	48.82±1.14
HERMIT	87.41±0.63	87.70±0.64	87.55±0.63	**87.65**±0.98	**82.04**±2.12	**84.74**±1.18

Table 3: Comparison of HERMIT with the results obtained in (Liu et al., 2019) for Intents and Entity Types.

	Combined		
	P	R	F1
Rasa	86.16±0.90	78.66±1.28	82.24±1.08
Dialogflow	83.19±2.43	79.07±3.10	81.07±2.64
LUIS	84.76±0.67	79.61±1.25	82.1±0.90
Watson	54.02±0.75	83.83±1.02	65.7±0.75
HERMIT	**87.52**±0.61	**85.03**±1.11	**86.25**±0.66

Table 4: Comparison of HERMIT with the results in (Liu et al., 2019) by combining Intent and Entity.

	Intent	Entity	Combined
HERMIT	**87.55**±0.63	84.74±1.18	**86.25**±0.66
– SA	87.03±0.74	84.35±1.15	85.81±0.81
– SA/CN	87.09±0.78	82.43±1.42	84.97±0.72
– SA/CRF	83.57±0.75	**84.77**±1.06	84.09±0.79
– SA/CN/CRF	83.78±1.10	82.22±1.41	83.10±1.06

Table 5: Ablation study of HERMIT on the NLU-BM.

chitecture, both in terms of standard deviation and task performance, with a good balance between intent and entity recognition.

4.4 Experiments on the ROMULUS dataset

In this section we report the experiments performed on the ROMULUS dataset (Table 6). Together with the evaluation metrics used in (Liu et al., 2019), we report the span F1, computed using the CoNLL-2000 shared task evaluation script, and the Exact Match (EM) accuracy of the entire sequence of labels. It is worth noticing that the EM Combined score is computed as the conjunction of the three individual predictions – e.g., a match is when all the three sequences are correct.

Results in terms of EM reflect the complexity of the different tasks, motivating their position within the hierarchy. Specifically, dialogue act identification is the easiest task (89.31%) with respect to frame (82.60%) and frame element (79.73%), due to the shallow semantics it aims to catch. However, when looking at the span F1, its score (89.42%) is lower than the frame element identification task (92.26%). What happens is that even though the label set is smaller, dialogue act spans are supposed to be longer than frame element ones, sometimes covering the whole sentence. Frame elements, instead, are often one or two tokens long, that contribute in increasing span based metrics. Frame identification is the most complex task for several reasons. First, lots of frame spans are interlaced or even nested; this contributes to increasing the network entropy. Second, while the dialogue act label is highly related to syntactic structures, frame identification is often subject to the inherent ambiguity of language (e.g., *get* can evoke both *Commerce_buy* and *Arriving*). We also report the metrics in (Liu et al., 2019) for consistency. For dialogue act and frame tasks, scores provide just the extent to which the network is able to detect those labels. In fact, the metrics do not consider any span information, essential to solve and evaluate our tasks. However, the frame element scores are comparable to the benchmark, since the task is very similar.

Overall, getting back to the combined EM accuracy, HERMIT seems to be promising, with the network being able to reproduce all the three gold sequences for almost 70% of the cases. The importance of this result provides an idea of the architecture behaviour over the entire pipeline.

4.5 Discussion

The experimental evaluation reported in this section provides different insights. The proposed architecture addresses the problem of NLU in wide-coverage conversational systems, modelling semantics through multiple Dialogue Acts and Frame-like structures in an end-to-end fashion. In addition, its hierarchical structure, which reflects the complexity of the single tasks, allows providing rich representations across the whole network. In this respect, we can affirm that the architecture successfully tackles the multi-task problem, with results that are promising in terms of usability and applicability of the system in real scenarios.

	P	R	F1	span F1	EM
Dialogue act	96.49±0.98	95.95±1.41	96.21±1.13	89.42±3.74	89.31±3.28
Frame	95.26±0.95	94.02±1.20	94.64±1.09	84.40±2.99	82.60±2.68
Frame element	95.62±0.61	93.98±0.76	94.79±0.56	92.26±1.22	79.73±2.03
Combined	93.90±0.89	92.95±0.86	93.42±0.83	–	**69.53**±2.50

Table 6: HERMIT performance over the ROMULUS dataset. P,R and F1 are evaluated following (Liu et al., 2019) metrics

However, a thorough evaluation in the wild must be carried out, to assess to what extent the system is able to handle complex spoken language phenomena, such as repetitions, disfluencies, etc. To this end, a real scenario evaluation may open new research directions, by addressing new tasks to be included in the multi-task architecture. This is supported by the scalable nature of the proposed approach. Moreover, following (Sanh et al., 2018), corpora providing different annotations can be exploited within the same multi-task network.

We also empirically showed how the same architectural design could be applied to a dataset addressing similar problems. In fact, a comparison with off-the-shelf tools shows the benefits provided by the hierarchical structure, with better overall performance better than any current solution. An ablation study has been performed, assessing the contribution provided by the different components of the network. The results show how the shortcut connections help in the more fine-grained tasks, successfully encoding richer representations. CRFs help when longer spans are being predicted, more present in the upstream tasks.

Finally, the seq2seq design allowed obtaining a multi-label approach, enabling the identification of multiple spans in the same utterance that might evoke different dialogue acts/frames. This represents a novelty for NLU in conversational systems, as such a problem has always been tackled as a single-intent detection. However, the seq2seq approach carries also some limitations, especially on the Frame Semantics side. In fact, label sequences are linear structures, not suitable for representing nested predicates, a tough and common problem in Natural Language. For example, in the sentence "*I want to buy a new mobile phone*", the [*to buy a new mobile phone*] span represents both the DESIRED_EVENT frame element of the *Desiring* frame and a *Commerce_buy* frame at the same time. At the moment of writing, we are working on modeling nested predicates through the application of bilinear models.

5 Future Work

We have started integrating a corpus of 5M sentences of real users chit-chatting with our conversational agent, though at the time of writing they represent only 16% of the current dataset.

As already pointed out in Section 4.5, there are some limitations in the current approach that need to be addressed. First, we have to assess the network's capability in handling typical phenomena of spontaneous spoken language input, such as repetitions and disfluencies (Shalyminov et al., 2018). This may open new research directions, by including new tasks to identify/remove any kind of noise from the spoken input. Second, the seq2seq scheme does not deal with nested predicates, a common aspect of Natural Language. To the best of our knowledge, there is no architecture that implements an end-to-end network for FrameNet based semantic parsing. Following previous work (Strubell et al., 2018), one of our future goals is to tackle such problems through hierarchical multi-task architectures that rely on bilinear models.

6 Conclusion

In this paper we presented HERMIT NLU, a hierarchical multi-task architecture for semantic parsing sentences for cross-domain spoken dialogue systems. The problem is addressed using a seq2seq model employing BiLSTM encoders and self-attention mechanisms and followed by CRF tagging layers. We evaluated HERMIT on a 25K sentences NLU-Benchmark and outperform state-of-the-art NLU tools such as Rasa, Dialogflow, LUIS and Watson, even without specific fine-tuning of the model.

Acknowledgement

This research was partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688147 (MuMMER project[10]).

[10] http://mummer-project.eu/

References

Martín Abadi et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of ACL and COLING*, Association for Computational Linguistics, pages 86–90.

Ankur Bapna, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck. 2017. Sequential dialogue context modeling for spoken language understanding. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 103–114. Association for Computational Linguistics.

Emanuele Bastianelli, Danilo Croce, Andrea Vanzo, Roberto Basili, and Daniele Nardi. 2016. A discriminative approach to grounded spoken language understanding in interactive robotics. In *Proceedings of the 2016 International Joint Conference on Artificial Intelligence (IJCAI)*, New York, USA.

Chandrakant Bothe, Fernando García, Arturo Cruz-Maya, Amit Kumar Pandey, and Stefan Wermter. 2018. Towards dialogue-based navigation with multivariate adaptation driven by intention and politeness for social robots. In *ICSR*, volume 11357 of *Lecture Notes in Computer Science*, pages 230–240. Springer.

Joyce Y Chai, Rui Fang, Changsong Liu, and Lanbo She. 2016. Collaborative language grounding toward situated human-robot dialogue. *AI Magazine*, 37(4):32–45.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35.

Yun-Nung Chen, Dilek Hakkani-Tür, Gökhan Tür, Jianfeng Gao, and Li Deng. 2016. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *INTERSPEECH*, pages 3245–3249. ISCA.

François Chollet et al. 2015. Keras. `https://keras.io`.

Charles J Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Daniel Guo, Gökhan Tür, Wen-tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop, SLT 2014, South Lake Tahoe, NV, USA, December 7-10, 2014*, pages 554–559.

Dilek Hakkani-Tur, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-Domain Joint Semantic Frame Parsing using Bi-directional RNN-LSTM. In *Proceedings of Interspeech*.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933. Association for Computational Linguistics.

Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. 2018. Interactively picking real-world objects with unconstrained spoken language instructions. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 3774–3781.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2010. Toward understanding natural language directions. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, HRI '10, pages 259–266, Piscataway, NJ, USA. IEEE Press.

Geert-Jan M. Kruijff, H. Zender, P. Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems*, 4(2).

Oliver Lemon and Olivier Pietquin. 2012. *Data-Driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces*. Springer Publishing Company, Incorporated.

Changliang Li, Liang Li, and Ji Qi. 2018. A self-attentive model with gate mechanism for spoken language understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3833. Association for Computational Linguistics.

Bing Liu and Ian Lane. 2016. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 685–689.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking Natural Language Understanding Services for building Conversational Agents. In *Proceedings of the International Workshop on Spoken Dialogue System*, page to appear, Siracusa, Sicily, Italy.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

P. J. Price. 1990. Evaluation of spoken language systems: The atis domain. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '90, pages 91–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abhinav Rastogi, Raghav Gupta, and Dilek Hakkani-Tur. 2018. Multi-task Learning for Joint Language Understanding and Dialogue State Tracking. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 376–384. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 173–179, Stroudsburg, PA, USA. Association for Computational Linguistics.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2018. A hierarchical multi-task approach for learning embeddings from semantic tasks. *arXiv preprint arXiv:1811.06031*.

M. Schuster and K.K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Igor Shalyminov, Arash Eshghi, and Oliver Lemon. 2018. Multi-task learning for domain-general spoken disfluency detection in dialogue systems. In *Proceedings of SemDIAL 2018 (AixDial)*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-Informed Self-Attention for Semantic Role Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038. Association for Computational Linguistics.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc.

Xiaodong Zhang and Houfeng Wang. 2016. A Joint Model of Intent Determination and Slot Filling for Spoken Language Understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16. AAAI Press.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1049–1058, New York, NY, USA. ACM.