A step beyond local observations with a dialog aware bidirectional GRU network for Spoken Language Understanding

Vedran Vukotić, Christian Raymond, Guillaume Gravier

A step beyond local observations with a dialog aware bidirectional GRU network for Spoken Language Understanding

Vedran Vukotić[1,2]*, Christian Raymond*[1,2], Guillaume Gravier[2,3]

[1]INSA de Rennes, Rennes, France
[2]INRIA/IRISA, Rennes, France
[3]CNRS, France

{vedran.vukotic, christian.raymond, guillaume.gravier}@irisa.fr

Abstract

Architectures of Recurrent Neural Networks (RNN) recently become a very popular choice for Spoken Language Understanding (SLU) problems; however, they represent a big family of different architectures that can furthermore be combined to form more complex neural networks. In this work, we compare different recurrent networks, such as simple Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) networks, Gated Memory Units (GRU) and their bidirectional versions, on the popular ATIS dataset and on MEDIA, a more complex French dataset. Additionally, we propose a novel method where information about the presence of relevant word classes in the dialog history is combined with a bidirectional GRU, and we show that combining relevant word classes from the dialog history improves the performance over recurrent networks that work by solely analyzing the current sentence.

Index Terms: spoken language understanding, deep learning, recurrent neural networks, RNN, long short-term memory, LSTM, gated recurrent units, GRU, bidirectional LSTM, bidirectional GRU, dialog, dialog history

1. Introduction

This paper focuses on Spoken Language Understanding (SLU) or more specifically, on the slot tagging task. In classical SLU systems, the key task is to label words with lexical semantics.

Many solutions have been proposed, ranging from HMM to CRF [1, 2]. However, in recent years, different types of recurrent neural networks have shown promising results and are increasing in popularity [3, 4, 5, 6].

The work presented in this paper is divided into two parts. In the first part, we evaluate different recurrent neural network architectures, such as simple recurrent neural networks (RNN), long short-term memory networks (LSTM) [7, 8] and novel gated recurrent units (GRU) [9]. Experiments are conducted on two datasets: the well know ATIS dataset and MEDIA, a more challenging French SLU dataset where recent RNN approaches competitive on ATIS are still below CRF [10]. Recent promising recurrent neural architectures are analyzed both in their monodirectional and bidirectional versions. In the second part, we explore the possibility of combining dialog awareness with bidirectional [11] GRU networks and we show that further improvements are possible by including parts of the dialog history. This paper is structured as follows: first we present the two datasets, namely ATIS and MEDIA in section 2. In section 3 we present the explored architectures starting by monodirectional models, progressing to bidirectional models and concluding with our proposed dialog aware model. We present the experiments and their respective results in Section 4 and we show which models perform better while also proving the importance of modeling relevant information from the whole dialog in Section 5.

2. Datasets

In this work, two datasets are used: ATIS and MEDIA. ATIS is a publicly available corpus used since the early nineties for SLU evaluation. MEDIA is a more recent, French dataset that has been made available through ELRA since 2008.

2.1. ATIS

The Air Travel Information System (ATIS) task [12] is dedicated to provide flight information. The semantic representation used is frame based. The SLU goal is to find the good frame and fill the corresponding slots.

Three values are used for each word, the word itself, a class to which the word might belong and the target label. There are 37 word classes in ATIS and they represent clusters like *country_name*, *airport_name* etc. Every word utilized within the dataset that belongs to a cluster is replaced by the name of the cluster. The target label is then predicted by using the set of words and/or word classes, where available. The word classes are also used to model the appearance of relevant classes when modeling the dialog history.

The training set consists of 4978 utterances selected from the Class A (context independent) training data in the ATIS-2 and ATIS-3 corpora while the ATIS test set contains both the ATIS-3 NOV93 and DEC94 datasets.

2.2. MEDIA

The research project MEDIA [13] evaluates different SLU models of spoken dialogue systems dedicated to provide tourist information. A 1250 French dialogue corpus has been recorded by ELDA following a Wizard of Oz protocol: 250 speakers have each followed 5 hotel reservation scenarios. This corpus has been manually transcribed, then conceptually annotated according to a semantic representation defined within the project. We used three values for each word: the word itself, a class to which the word might belong and the target label. The classes of words are clusters to which multiple words belong. E.g. all city names used within the corpus belong to a *city_name* class. Most words do not belong to any specific class and are used as such. The target labels are again predicted from the words and/or word classes and the word classes are used to model the appearance of relevant classes in the dialog history.

The MEDIA semantic dictionary contains 83 concept labels, 19 word classes (specifiers) and 4 types of modal information. In this study we focus only on concept extraction so we don't use modal information. The MEDIA corpus is split into 3 parts. The first part (720 dialogues, 12K messages) is used for training the models, the second (79 dialogues, 1.3K message) when cross-validation is performed, and the third part (200 dialogues, 3K message) is used for testing.

3. Methodology

In this section, we first present the different architectures of recurrent neural networks that were analyzed in this work. We start by analyzing simple recurrent neural networks and long short-term memory (LSTM) networks that have been successfully used for spoken language understanding [3, 4]. Next, we analyze novel gated recurrent units (GRU) [9], which have recently been also used for spoken language understanding [6] and we explore bidirectionalism in both LSTM and GRU networks. In the last part, we introduce a novel architecture that combines a bidirectional GRU with a separate network that introduces awareness to relevant word concepts from dialog history.

3.1. Recurrent Neural Networks

Recurrent neural networks define a family of neural networks that contain a loop, thus making them recurrent and allowing information to persist in them. Simple recurrent neural networks contain solely a loop while other, more complex recurrent neural networks, are composed with one or more gates which allow them to model information to retain and forget.

3.1.1. Simple Recurrent Neural Networks

Simple RNNs are defined as follows:

$$\boldsymbol{h}_t = act_1(\boldsymbol{W}_h \boldsymbol{h}_{t-1} + \boldsymbol{W}_x \boldsymbol{x}_t)$$
$$\boldsymbol{o}_t = act_2(\boldsymbol{W}_o \boldsymbol{h}_t)$$

where $\boldsymbol{x}_t$ is the input at time t, $\boldsymbol{h}_t$ the hidden state at time t, $\boldsymbol{W}_\square$ are the weight matrices, act are activation functions and $\boldsymbol{o}_t$ the output of the recurrent network at time t. Variations of this basic architecture, namely Jordan and Elman architectures (that contain additional *context units*) have also been successfully used for spoken language understanding [3, 10]. In practice, RNNs have difficulties modeling long-term dependencies [14]. Gated recurrent networks such as LSTM and GRU networks improve upon this problem.

3.1.2. Long Short-Term Memory Networks

Long short-term memory networks [7, 8] introduce a series of gates (input gate, forget gate and output gate) that help model the information retained by the recurrent network. LSTMs are defined as follows:

$$\boldsymbol{f}_t = act_1(\boldsymbol{W}_f[\boldsymbol{h}_{t-1} \| \boldsymbol{x}_t] + \boldsymbol{b}_f)$$
$$\boldsymbol{i}_t = act_1(\boldsymbol{W}_i[\boldsymbol{h}_{t-1} \| \boldsymbol{x}_t] + \boldsymbol{b}_i)$$
$$\widehat{\boldsymbol{C}}_t = act_2(\boldsymbol{W}_c[\boldsymbol{h}_{t-1} \| \boldsymbol{x}_t] + \boldsymbol{b}_c)$$

where f_t and i_t represent the *forget* and *input* gates respectively, the $\|$ symbol denotes concatenation and $\widehat{\boldsymbol{C}}$ represents the new candidate value for the LSTM cell state. The cell state is then updated as follows:

$$\boldsymbol{C}_t = \boldsymbol{f}_t C_{t-1} + \boldsymbol{i}_t \widehat{\boldsymbol{C}}_t$$

Finally, the current output $\boldsymbol{o}_t$ and hidden state $\boldsymbol{h}_t$ are updated:

$$\boldsymbol{o}_t = act_1(\boldsymbol{W}_o[\boldsymbol{h}_{t-1} \| \boldsymbol{x}_t] + \boldsymbol{b}_o)$$
$$\boldsymbol{h}_t = \boldsymbol{o}_t act_2(\boldsymbol{C}_t)$$

LSTM networks have also been successfully used in spoken language understanding, either by themselves [4] or as encoder-decoder (sequence to sequence) architectures [5] that are more commonly used in machine translation tasks [15, 16].

3.1.3. Gated Recurrent Units

Gated recurrent units [9] are a more novel variation of LSTM networks. They combine the forget and input gates into one update gate and merge the hidden state and cell state into one state. More formally:

$$\boldsymbol{z}_t = act_1(\boldsymbol{W}_z[\boldsymbol{h}_{t-1} \| \boldsymbol{x}_t])$$
$$\boldsymbol{r}_t = act_1(\boldsymbol{W}_r[\boldsymbol{h}_{t-1} \| \boldsymbol{x}_t])$$
$$\widehat{\boldsymbol{h}}_{\boldsymbol{t}} = act_2(\boldsymbol{W}[\boldsymbol{h}_{t-1} \| \boldsymbol{x}_t])$$
$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) + \boldsymbol{z}_t \widehat{\boldsymbol{h}}_{\boldsymbol{t}}$$

where $\boldsymbol{r}_t$ is a reset gate and $\boldsymbol{z}_t$ is an update gate. GRUs have been shown to perform better than regular LSTMs while also being faster due to a simpler architecture [17].

3.1.4. Bidirectionalism

Recurrent neural networks typically model information solely in one direction. In some cases, it's been shown that reversing the sequence can improve the performance of a recurrent network in machine translation applications [18]. It's thus best to model sequences in both directions (both $\ldots x_{i-1}, x_i, x_{i+1} \ldots$ and $\ldots x_{i+1}, x_i, x_{i-1} \ldots$). Modeling information in both directions can be done by implementing a bidirectional structure directly within the architecture of a recurrent neural network [11], or two recurrent neural networks working with opposing directions can be combined to achieve the same goal [6]. The latter method is more common with complex recurrent neural networks and is also used in our work.

3.2. Bidirectional GRU and Dialog Awareness

Equal sequences of words may be labeled differently depending on context. This relevant context can be part of the current sentence or part of previous sentences that are part of the current dialog history. For example, let's illustrate the case of the words *{l', hôtel}* in two phrases from the MEDIA dataset. The phrase *{je, souhaite, réserver, à, **l'**, **hôtel**, ibis} (I'd like to reserve at the Ibis hotel)* is labeled as *{B-command-tache, I-command-tache, I-command-tache, **B-hotel-marque**, **I-hotel-marque**, I-hotel-marque}* while the phrase *{savoir, s', il, y, a, un, restaurant, dans, **l'**, **hôtel**} (<missing begining> know if there is a restaurant inside the hotel)* is labeled as *{O, O, O, O, O, B-hotel-services, I-hotel-services, O, **B-lienRef-coRef**, **B-objetBD**}*. Each of the target labels that are not null *(O)* starts either with *B (begin)* or *I (inside)*, indicating whether it's the first word of a section labeled as such or one of the subsequent words inside the same section. In this example, we focus on three labels: *hotel-marque* (hotel name/mark), *lienRef-CoRef* (referential/coreferential link) and *objetBD* (object). The difference between the first and the second sentence can be inferred by how far the dialog has progressed and what has already been mentioned so far. In this particular example, if the user is asking for specifics about a hotel, it means that there has probably been a previous mention of a hotel and that *"l'hôtel"* should be interpreted differently - as a reference to a previously mentioned object and not as a new hotel. Sometimes solely the current sentence suffice while in most cases, the knowledge of what has been previously mentioned improves the understanding of the current sentence.

Network Architecture	Parameters	Accuracy	Precision	Recall	F1
ATIS					
RNN	emb=200, win=11, dim=200	97.71 (0.06)	94.02 (0.10)	95.26 (0.20)	94.63 (0.14)
LSTM	emb=200, win=11, dim=200	97.89 (0.04)	94.47 (0.18)	95.80 (0.19)	95.12 (0.17)
Bidirectional LSTM	emb=200, win=11, dim=200	97.91 (0.05)	94.61 (0.13)	95.86 (0.13)	95.23 (0.11)
GRU	emb=200, win=11, dim=200	97.95 (0.05)	94.72 (0.11)	96.14 (0.04)	95.43 (0.06)
Bidirectional GRU	emb=200, win=11, dim=200	98.00 (0.06)	94.86 (0.15)	96.21 (0.19)	95.53 (0.17)
Bi-GRU with d. awareness	emb=200, win=11, dim=200, dns=37	97.97 (0.08)	94.85 (0.26)	96.06 (0.22)	95.54 (0.16)
MEDIA					
RNN	emb=200, win=11, dim=200	86.08 (0.25)	76.13 (0.67)	80.95 (0.23)	78.46 (0.45)
LSTM	emb=200, win=11, dim=200	87.80 (0.73)	80.49 (1.55)	82.61 (1.20)	81.54 (1.33)
Bidirectional LSTM	emb=200, win=11, dim=200	88.45 (0.05)	82.54 (0.85)	83.61 (0.22)	83.07 (0.37)
GRU	emb=200, win=11, dim=200	88.39 (0.16)	82.73 (0.56)	83.63 (0.38)	83.18 (0.47)
Bidirectional GRU	emb=200, win=11, dim=200	88.81 (0.09)	82.93 (0.42)	84.34 (0.33)	83.63 (0.16)
Bi-GRU with d. awareness	emb=200, win=11, dim=200, dns=19	88.76 (0.25)	83.55 (0.41)	84.23 (0.20)	**83.89** (0.27)

Table 1: Performances of the various recurrent architectures on ATIS and MEDIA. Averaged (over multiple runs) accuracy, precision, recall, F1 measure (%) and their respective standard deviations (in parenthesis).

We utilize a set of 37 word classes for ATIS and 19 word classes for MEDIA. To illustrate, a few of word classes utilized for ATIS are: *{aircraft_code, airline_code, airline_name, airport_code, airport_name, city_name, class_type, cost_relative, country_name, day_name, ... }*. The presence of a word that belongs to one of those classes within the current dialog history (the sentences of the current dialog history, from the first sentence, until the current sentence) is encoded in a binary vector of length 37 or 19 respectively. In the MEDIA dataset, dialogues contain from 1 to 56 user sentences. ATIS does not provide a dialog so only the word classes from the current sentence are modeled.

Binary vectors containing information about the presence of word concepts are fed to a neural network in parallel with the context windows of the currently analyzed sentence. Words from the context window are first passed though an embedding layer after which two GRUs working in opposing directions follow. Their outputs are concatenated and dropout is applied. Dialog awareness vectors are instead passed through a dense, fully-connected layer of the same length as the input vectors. The two parts are then merged by a fully-connected dense layer of the size equal to the number of output labels and they are passed through a final activation function. Our proposed architecture is illustrated in Figure 1.

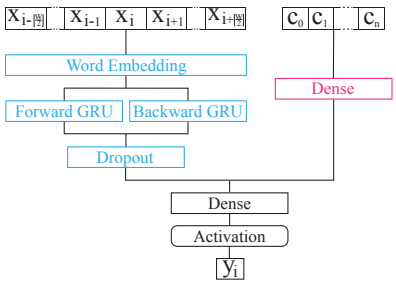

Figure 1: Our proposed architecture: a bidirectional GRU combined with a dialog aware fully-connected dense layer

4. Experiments

All the described architectures were implemented with the *Keras* framework [19]. Each experiment was run multiple times, the results were averaged and their standard deviations were computed, as shown in Table 1. Before performing the experiments, different hyperparameters (including number of layers, their sizes, dropout values and embedding sizes) were tested and the minimal characteristics were chosen after their performances ceased to improve.

4.1. ATIS

The parameters that worked best consist of an embedding size of 200 (the embeddings are learned jointly, while training the whole network), a context window of 11 (5 words before and 5 words after the current word) and an output size of the recurrent network (either RNN, LSTM or GRU) of 200. Simple RNNs stopped improving with a smaller context window. However, a window of 11 did not make the results worse, so we kept the same window size over all experiments to have a more sensible comparison. It's important to emphasize that we evaluated solely the standard model of simple RNNs, as described in Section 3. Jordan and Elman architectures are special cases of the previously described architecture that include additional context units and perform slightly better [3, 10] (F1=83.25% on MEDIA). The last fully-connected dense layer is always of a size equal to the number of output classes, after which a sigmoid activation layer follows. It was determined experimentally that a sigmoid activation layer performs better than layers with other common activation functions. We found that dropout of 50% worked best.

For bidirectional networks (namely bidirectional LSTM and bidirectional GRU networks), the recurrent networks were duplicated, made to work in the opposing direction and combined with the original forward networks by concatenating their outputs. Except for the additional backward recurrent neural network, all the other hyperparameters remain the same.

Dialog aware bidirectional GRU networks were formed by adding a fully-connected dense layer that reacts to the vector describing which word classes were mentioned in the current dialog history (which is just the current sentence in the case of ATIS). The best results were obtained with only one fully-connected dense layer of size 37 (same size as the input - number of possible word classes) connecting the input to the merging layer.

Although it might seem apparent from Table 1 that some architectures have a better F1 score than others, not many improvements come with a sensible confidence interval. It's clear that *ATIS does not present a challenging dataset*; that most

methods obtain very similar performances and that it's difficult to derive statistically more significant conclusions based on this dataset. However, it's clear that LSTM networks perform better than simple RNNs ($\alpha = 0.005$) in terms of F1 measure and GRU networks outperform LSTM networks ($\alpha = 0.02$). With $\alpha = 0.2$, we can also state that bidirectional networks (bidirectional LSTMs and bidirectional GRUs) show improved performance over their monodirectional versions, although this statement is not very strong. Utilizing word classes from the dialog does not improve the results since, on ATIS, the word classes are modeled solely within the current phrase.

4.2. MEDIA

For MEDIA, the best hyperparameters are equal to those used on the ATIS dataset: embeddings (learned jointly with the whole network) of size 200, context windows of size 11 (5 words before and 5 words after the current word), recurrent units with an output of size 200, 50% dropout, a fully dense layer of the size equal to the number of target classes and a sigmoidal activation function (other activation functions showed less good results also on MEDIA). The only difference is present in the case of dialog aware bidirectional GRU networks where, in the case of MEDIA, the fully-connected dense layer connecting the input representing the word concepts from the dialog is still equal to the number of possible word classes but is now of size 19.

MEDIA is clearly a more challenging dataset and conclusions that are statistically more significant can drawn. We show that, in terms of F1 measure, LSTM networks outperform simple RNNs ($\alpha = 0.01$) and that GRU networks perform better than LSTM networks ($\alpha = 0.05$). It's also clear that bidirectional gated recurrent networks work better than gated recurrent networks that work solely in one direction: bidirectional LSTMs outperform LSTMs ($\alpha = 0.1$) and bidirectional GRUs outperform GRUs ($\alpha = 0.1$). Finally, we show that combining word concepts from the dialog with bidirectional GRU networks further improves the results over bidirectional GRU networks that utilize solely the current sentence ($\alpha = 0.1$).

5. Conclusions

We analyzed different recurrent neural architectures on two datasets: ATIS and MEDIA. We show that gated recurrent neural networks, known for better long dependency modeling, clearly outperform simple RNNs. Within gated recurrent neural networks, we show that novel GRU networks outperform LSTM networks. Gated recurrent networks model information in one direction. Modeling information in both direction by combining two networks with opposing direction improves performance, as demonstrated with both bidirectional LSTM networks and bidirectional GRU networks. Finally, we show that adding information about the presence of specific word classes within the current dialog history further improves the performance of the previously best-performing bidirectional GRU networks. Unfortunately, simple CRF methods still slithgtly outperform RNN methods [10]. We believe this is due to better target dependency modeling that CRF offers. However, RNNs represent a competitive framework that might offer easier extensions such as attention models implemented over the dialog history.

We believe that there is not much improvement left to be done with architectures that utilize solely the current sentence. As shown, improvement can be achieved by integrating more distant dependencies that are part of the dialog but are not necessarily part of the current sentence. In our opinion, future work should address means of incorporating knowledge from the entire dialog, either by engineering relevant features or by deploying appropriate attention models. As a consequence, experiments performed on datasets more complex than ATIS are also required.

6. References

[1] C. Raymond and G. Riccardi, "Generative and Discriminative Algorithms for Spoken Language Understanding," in *InterSpeech*, Antwerp, Belgium, August 2007, pp. 1605–1608.

[2] Y. He and S. Young, "Semantic Processing using the Hidden Vector State Model," *Computer Speech and Language*, vol. 19, pp. 85–106, 2005.

[3] K. Yao, G. Zweig, M.-Y. Hwang, Y. Shi, and D. Yu, "Recurrent neural networks for language understanding," in *InterSpeech*, 2013, pp. 2524–2528.

[4] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi, "Spoken language understanding using long short-term memory neural networks," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 189–194.

[5] G. Kurata, B. Xiang, B. Zhou, and M. Yu, "Leveraging Sentence-level Information with Encoder LSTM for Natural Language Understanding," *arXiv preprint arXiv:1601.01530*, 2016.

[6] W. C. Zhilin Yang, Ruslan Salakhutdinov, "Multi-Task Cross-Lingual Sequence Tagging from Scratch," in *arXiv*, 2016.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[9] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[10] V. Vukotic, C. Raymond, and G. Gravier, "Is it time to switch to word embedding and recurrent neural networks for spoken language understanding?" in *InterSpeech*, Dresde, Germany, September 2015.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *Signal Processing, IEEE Transactions on*, vol. 45, no. 11, pp. 2673–2681, 1997.

[12] D. A. Dahl, M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and E. Shriberg, "Expanding the scope of the ATIS task: the ATIS-3 corpus," in *HLT*, 1994, pp. 43–48.

[13] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic Annotation of the French Media Dialog Corpus," in *InterSpeech*, Lisbon, September 2005.

[14] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *Neural Networks, IEEE Transactions on*, vol. 5, no. 2, pp. 157–166, 1994.

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

[16] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Advances in Neural Information Processing Systems*, 2015, pp. 2755–2763.

[17] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[19] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.