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

Vedran Vukotić, Christian Raymond, Guillaume Gravier

IRISA/INRIA Rennes & INSA Rennes, France

INTERSPEECH 2016
September 12th 2016, San Francisco, CA

Introduction

Spoken Language Understanding

- (previously) state-of-the-art were Conditional Random Fields [Vukotic et al., 2015]
- recently **Recurrent Neural Networks** became promising and popular [Yao et al., 2013, Yao et al., 2014, Kurata et al., 2016, Zhilin Yang, 2016]

Introduction

Spoken Language Understanding

- (previously) state-of-the-art were Conditional Random Fields [Vukotic et al., 2015]
- recently **Recurrent Neural Networks** became promising and popular [Yao et al., 2013, Yao et al., 2014, Kurata et al., 2016, Zhilin Yang, 2016]

questions

- which RNN architecture is best suited for SLU?
- are there architectural extensions that can improve performance?
- will any dataset help answer the previous two questions?

Introduction

test different RNNs

- simple RNNs (standard; Elman and Jordan architectures tested previously)
- Long Short-Term Memory (LSTM) networks
- Gated Recurrent Unit (GRU) networks

architectural extensions

- single direction modelling *vs.* bidirectional modelling
- adding dialog awareness

ATIS & Media presentation

ATIS: obtain air travel information such as flight schedules, fares, and ground transportation from a relational database

x=	*list*	twa	*flights from*	washington	*to*	philadelphia
y=	<null>	<airline>	<null>	<depart.city>	<null>	<arrive.city>

MEDIA: reservation of hotel rooms with tourist information.

x=	*euh*	une	chambre pour deux personnes	au novotel
y=	<null>	<number>	<room-type>	<hotel-mark>

ATIS & Media datasets

Air Travel Information System

- training corpus: 4978 utterances
- testing corpus: 893 utterances
- 572 words, 64 labels
- words supporting concept 49%
  - segmentation: easy: almost one word to concept correspondence
  - classification: easy: main ambiguity $\rightarrow$ departure *vs* arrival info

Media

- training corpus: 12922 utterances
- testing corpus: 4772 utterances
- 2460 words, 75 labels
- words supporting concept 72%
  - segmentation: hard
  - classification: hard: hierarchical attributes, complex dependencies

Simple RNNs

- simplest form of recurrent neural networks
- hidden state dependent on previous hidden state
- output dependent on hidden state

	ATIS		MEDIA	
Method	F1 (%)	impr.	F1(%)	impr.
Classic RNN	94.63	-	78.46	-

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

Long Short-Term Memory (LSTM) networks

- designed to efficiently model long-term dependencies
- introduces a series of gates (input gate, forget gate and output gate)

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

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

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

Long Short-Term Memory (LSTM) networks

- designed to efficiently model long-term dependencies
- introduces a series of gates (input gate, forget gate and output gate)

	ATIS		MEDIA	
Method	F1 (%)	impr.	F1(%)	impr.
Classic RNN	94.63	-	78.46	-
LSTM	95.12	✓	81.54	✓

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

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

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

- modeling long-term dependencies helps
- LSTMs outperform RNNs on both ATIS and MEDIA ✓

Gated Recurrent Unit (GRU) networks

- a recent simplification / improvement over LSTMs [Cho et al., 2014]
- forget and input gates are merged into **one** update gate
- hidden state and cell state combined

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

Gated Recurrent Unit (GRU) networks

- a recent simplification / improvement over LSTMs [Cho et al., 2014]
- forget and input gates are merged into **one** update gate
- hidden state and cell state combined

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

	ATIS		MEDIA	
Method	F1 (%)	impr.	F1(%)	impr.
Classic RNN	94.63	-	78.46	-
LSTM	95.12	✓	81.54	✓
GRU	95.43	✓	83.15	✓

- GRUs outperform LSTMs (and are also faster!) ✓

Bidirectional LSTMs / GRUs

- modeling left to right or right to left?
- why not both?

- two possibilities:
  - integrate double connections within the architecture(s)
  - merge two architectures working in opposing directions

Bidirectional LSTMs / GRUs

- modeling left to right or right to left?
- why not both?

- two possibilities:
  - integrate double connections within the architecture(s)
  - merge two architectures working in opposing directions

	ATIS		MEDIA	
Method	F1 (%)	impr.	F1(%)	impr.
Classic RNN	94.63	-	78.46	-
LSTM	95.12	✓	81.54	✓
Bi-LSTM	95.23	~	83.07	✓
GRU	95.43	✓	83.15	✓
Bi-GRU	95.53	~	83.63	✓

- poor significance on ATIS ($\alpha = 0.1$)
- MEDIA: bidirectional modeling is always a better choice ✓

Adding dialog awareness

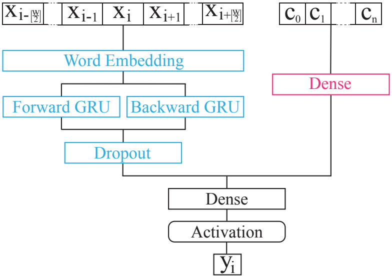
- modeling the presence of specific word classes within the dialog history (including the current sentence, until the current word)
  - e.g. {aircraft_code, airline_code, airline_name, airport_code, airport_name, city_name, class_type, cost_relative, country_name, day_name, ...}
  - binary features

Adding dialog awareness

- modeling the presence of specific word classes within the dialog history (including the current sentence, until the current word)
  - e.g. {aircraft_code, airline_code, airline_name, airport_code, airport_name, city_name, class_type, cost_relative, country_name, day_name, ...}
  - binary features
- history length:
  - MEDIA: 1 to 56 sentences per dialog
  - ATIS: limited to one sentence

Dialog awareness - implementation

- modeling the presence of specific word classes within the dialog history (until the current word)
  - word classes from a database
  - binary features: 37 for ATIS, 19 for MEDIA
  - fully-connected dense layer
- merging with a Bidirectional GRU to obtain a final decision



Dialog awareness - influence

- improvement on MEDIA ✓
- no significant improvement on ATIS
  - for ATIS the "dialog" is limited to the current sentence
  - lack of challenging segmentation in ATIS

	ATIS		MEDIA	
Method	F1 (%)	impr.	F1(%)	impr.
Classic RNN	94.63	-	78.46	-
LSTM	95.12	✓	81.54	✓
Bi-LSTM	95.23	~	83.07	✓
GRU	95.43	✓	83.15	✓
Bi-GRU	95.53	~	83.63	✓
Bi-GRU+diag aw.	95.54	✗	83.89	✓

Conclusion

1. Gated Recurrent Networks are best suited for SLU
   - RNN < LSTM < GRU

Conclusion

1. Gated Recurrent Networks are best suited for SLU
   - RNN < LSTM < GRU
2. modeling is best done in both directions
   - LSTM < **Bi-LSTM** < GRU < **Bi-GRU**

Conclusion

1. Gated Recurrent Networks are best suited for SLU
   - RNN < LSTM < GRU
2. modeling is best done in both directions
   - LSTM < **Bi-LSTM** < GRU < **Bi-GRU**
3. modeling key parts of the dialog helps!
   - when there is a "real" dialog
   - **future work:** smarter dialog awareness (e.g. attention model)

Conclusion

1. Gated Recurrent Networks are best suited for SLU
   - RNN < LSTM < GRU
2. modeling is best done in both directions
   - LSTM < **Bi-LSTM** < GRU < **Bi-GRU**
3. modeling key parts of the dialog helps!
   - when there is a "real" dialog
   - **future work:** smarter dialog awareness (e.g. attention model)
4. ATIS is not challenging enough
   - hard to obtain reasonable significance
   - MEDIA is a solid dataset that helps differentiating different approaches

Thank you!

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

Kurata, G., Xiang, B., Zhou, B., and Yu, M. (2016).
Leveraging Sentence-level Information with Encoder LSTM for Natural Language Understanding.
arXiv preprint arXiv:1601.01530.

Vukotic, V., Raymond, C., and Gravier, G. (2015).
Is it time to switch to word embedding and recurrent neural networks for spoken language understanding?
In *InterSpeech*, Dresde, Germany.

Yao, K., Peng, B., Zhang, Y., Yu, D., Zweig, G., and Shi, Y. (2014).

Spoken language understanding using long short-term memory neural networks.
In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 189–194. IEEE.

Yao, K., Zweig, G., Hwang, M.-Y., Shi, Y., and Yu, D. (2013).
Recurrent neural networks for language understanding.
In *InterSpeech*, pages 2524–2528.

Zhilin Yang, Ruslan Salakhutdinov, W. C. (2016).
Multi-Task Cross-Lingual Sequence Tagging from Scratch.
In *arXiv*.