北京大學
PEKING UNIVERSITY

Marrying Up Regular Expressions with Neural Networks: A Case Study for Spoken Language Understanding

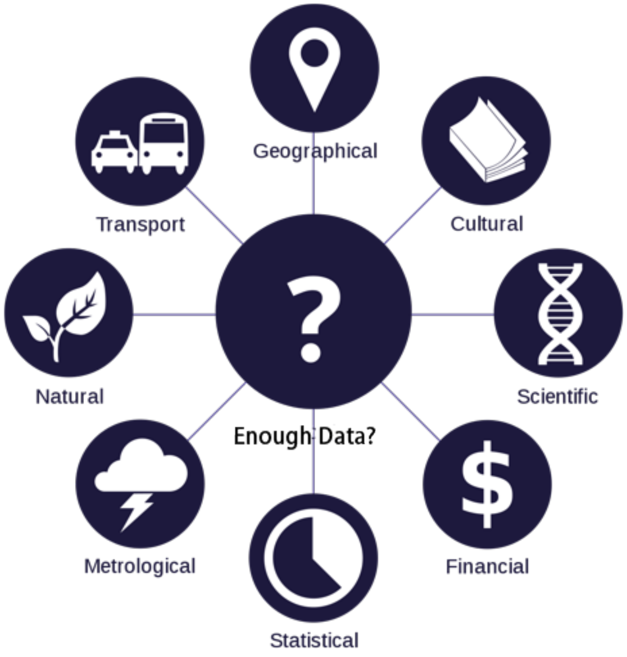
Bingfeng Luo, Yansong Feng, Zheng Wang, Songfang Huang, Rui Yan and Dongyan Zhao

2018/07/18

Data is Limited

- **Most of the popular models in NLP are data-driven**

- **We often need to operate in a specific scenario → Limited data**



Data is Limited

- Take **spoken language understanding** as an example
  - Understanding user query
  - Need to be implemented for many domains

Intent Detection	flights from Boston to Tokyo →	intent: ***flight***
Slot Filling	flights from Boston to Tokyo →	fromloc.city: ***Boston*** toloc.city: ***Tokyo***

Data is Limited

- Take spoken language understanding as an example
  - Need to be implemented for many domains → Limited data
  - E.g., intelligent customer service robot
- What can we do with limited data?

Intent Detection	flights from Boston to Tokyo	→	intent: *flight*
Slot Filling	flights from Boston to Tokyo	→	fromloc.city: *Boston* toloc.city: *Tokyo*

Regular Expression Rules

- When data is limited → Use **rule-based system**
- **Regular expression** is the most commonly used rule in NLP
  - Many regular expression rules in company

Intent Detection

/^flights? from/

flights from Boston to Tokyo → **intent: *flight***

Slot Filling

/from (_CITY) to (_CITY)/

flights from Boston to Tokyo → **fromloc.city: *Boston***
toloc.city: *Tokyo*

_CITY=Boston | Tokyo | Beijign | ...

Regular Expression Rules

- However, regular expressions are hard to generalize
- Neural networks are potentially good at generalization
- Can we combine the advantages of two worlds?

Regular Expressions

/^flights? from/

Pro: controllable, do not need data

Con: need to specify every variation

Neural Network

[0.23, 0.11, -0.32, ...]

Pro: semantic matching

Con: need a lot of data

Which Part of Regular Expression to Use?

- **Regular expression (RE) output is useful**
  - **As feature**
  - **Fusion in output**

Intent Detection

/^flights? from/

flights from Boston to Tokyo → **intent: *flight***

Slot Filling

/from (_CITY) to (_CITY)/

flights from Boston to Tokyo →

fromloc.city: *Boston*
toloc.city: *Tokyo*

Which Part of Regular Expression to Use?

- **Regular expression (RE) output is useful**
- **RE contains clue words**
    - **NN should attend to these clue words for prediction**
    - **Guide attention module**

Intent Detection

/^flights? from/

flights from Boston to Tokyo → **intent: *flight***

Slot Filling

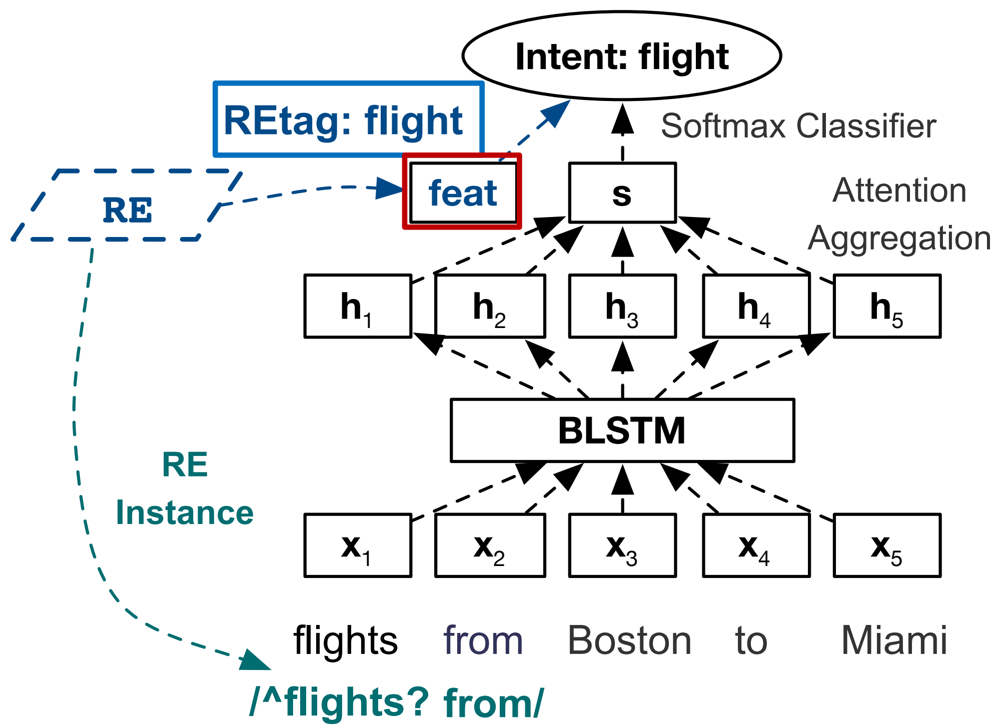
/from (_CITY) to (_CITY)/

flights from Boston to Tokyo → **fromloc.city: *Boston***
toloc.city: *Tokyo*

Method 1: RE Output - As Features

- Embed the REtag, append to input

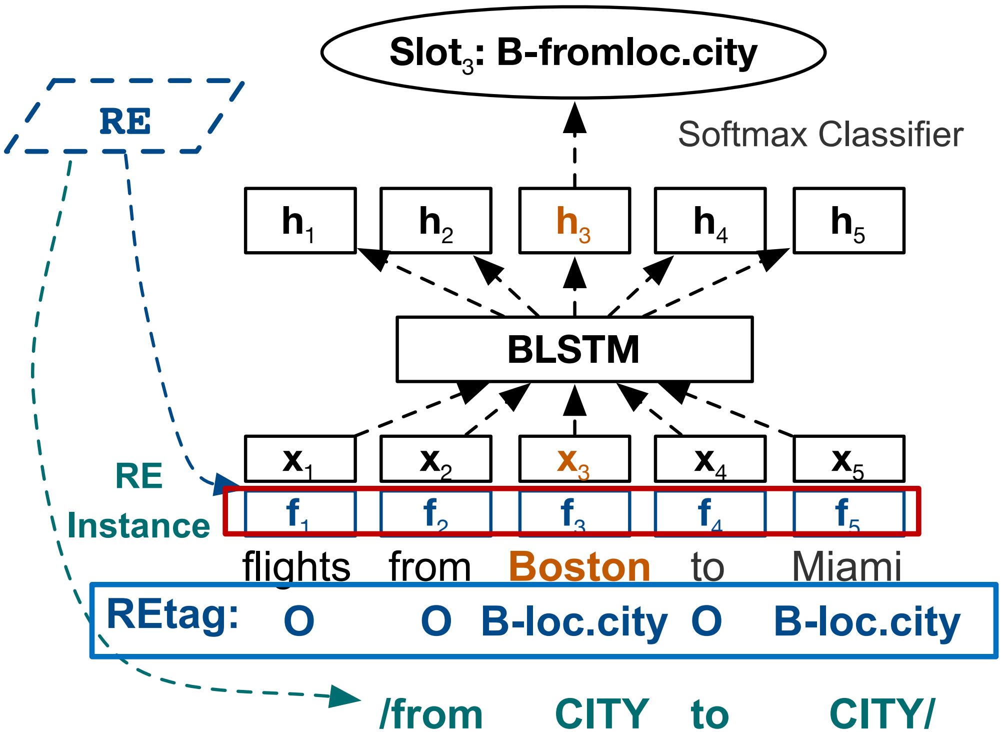
Intent Detection



Method 1: RE Output - As Features

- Embed the REtag, append to input

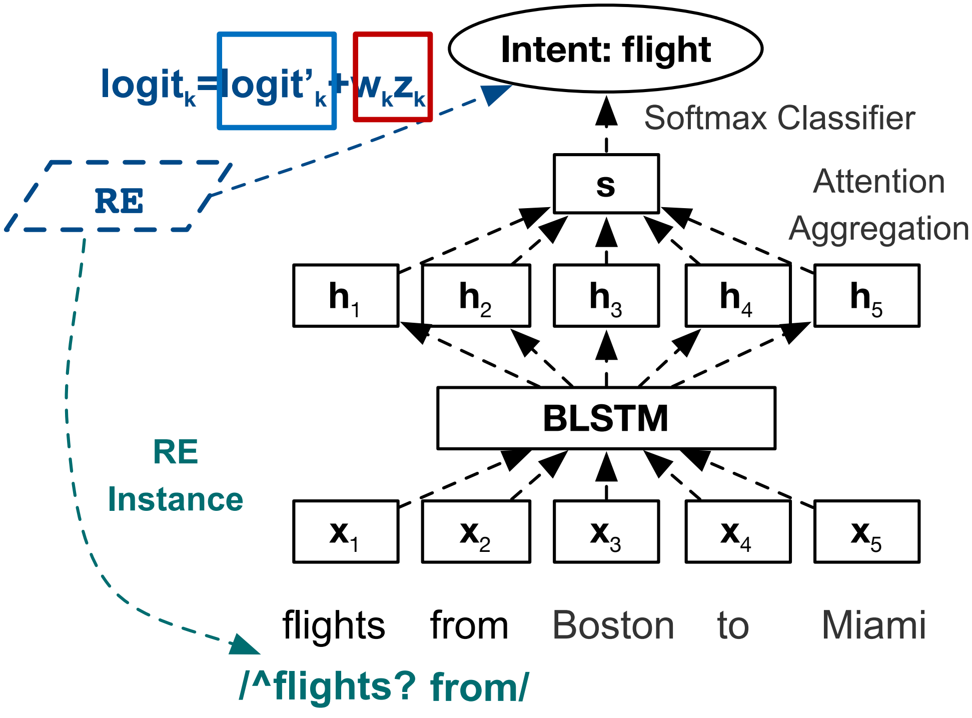
Slot Filling



Method 2: RE Output - Fusion in Output

- $logit_k = \boxed{logit'_k} + \boxed{w_k z_k}$
  - $logit'_k$ is the NN output score for class k (before softmax)
  - $z_k \in \{0, 1\}$, whether regular expression predict class k

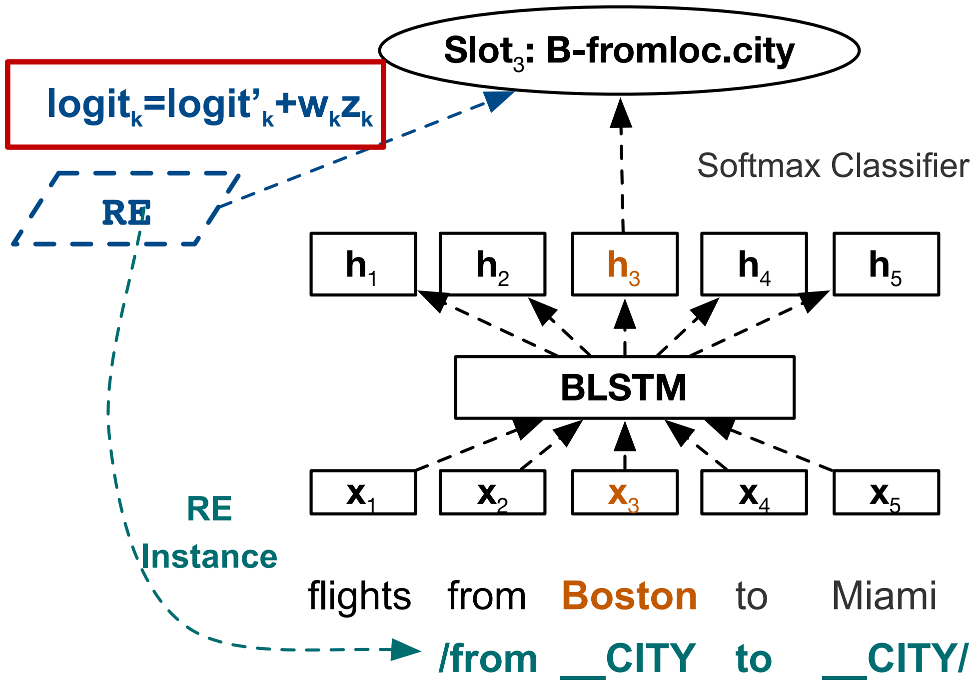
Intent Detection



Method 2: RE Output - Fusion in Output

- $logit_k = logit'_k + w_k z_k$
  - $logit'_k$ is the NN output score for class k (before softmax)
  - $z_k \in \{0, 1\}$, whether regular expression predict class k

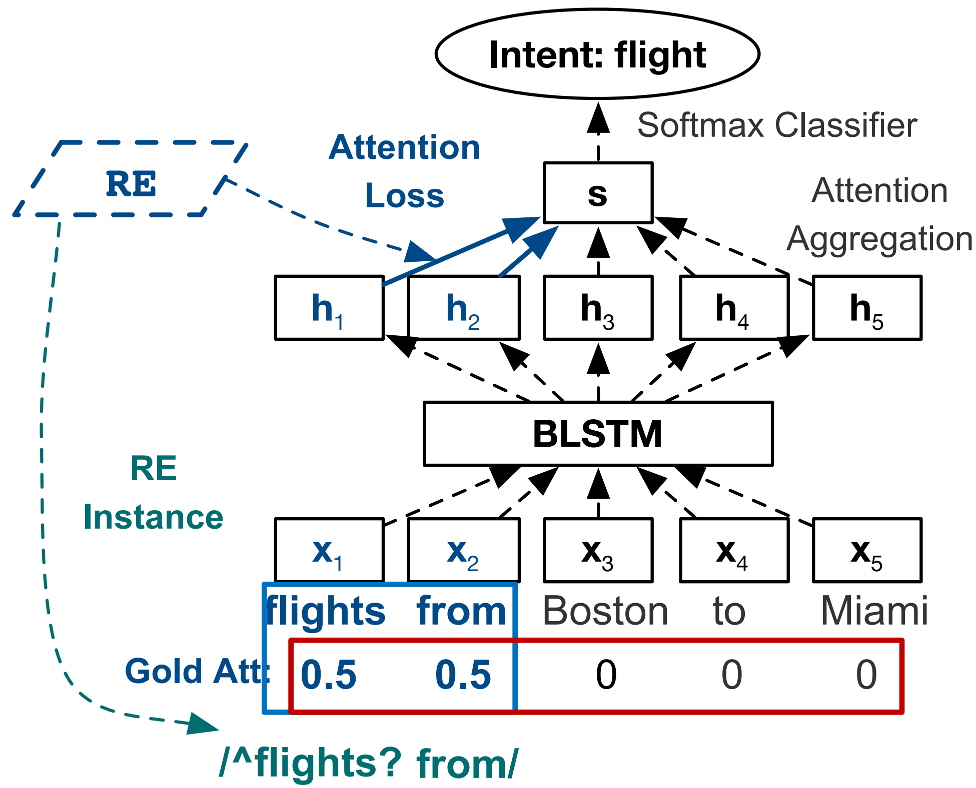
Slot Filling



Method 3: Clue Words - Guide Attention

- **Attention should match clue words**
  - **Cross Entropy Loss**

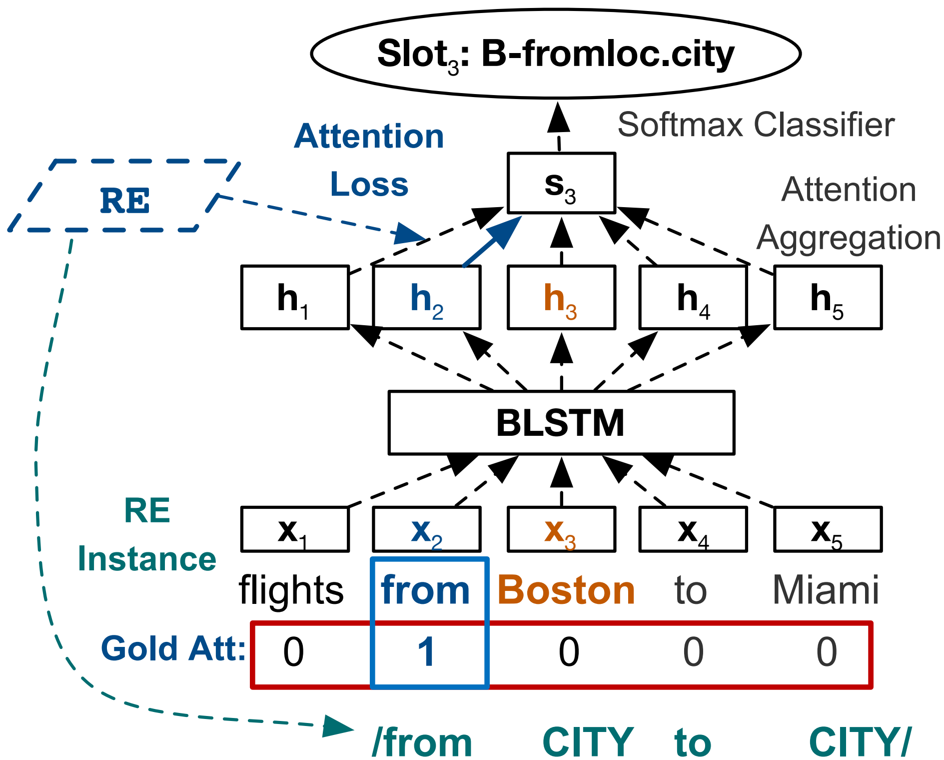
Intent Detection



Method 3: Clue Words - Guide Attention

- **Attention should match clue words**
  - **Cross Entropy Loss**

Slot Filling



Method 3: Clue Words - Guide Attention

- **Positive Regular Expressions (REs) & Negative REs**
  - **REs can indicate the input belong to class k, or does not belong to class k**
  - **Correction of wrong predictions**

/^how long/

How long does it take to fly from LA to NYC?

intent: *abbreviation*

Method 3: Clue Words - Guide Attention

- **Positive Regular Expressions (REs) & Negative REs**
  - **Corresponding to positive / negative REs**
  - $logit_k = logit_{k;\, positive} - logit_{k;\, negative}$

/^how long/

How long does it take to fly from LA to NYC?

intent: *abbreviation*

Method 3: Clue Words - Guide Attention

- **Positive REs and Negative REs interconvertible**
  - **A positive RE for one class can be negative RE for other classes**

/^flights? from/

flights from Boston to Tokyo

- ✔ **intent: *flight***
- ✘ **intent: *abbreviation***
- ✘ **intent: *airfare***
- ✘ **…**

Experiment Setup

- **ATIS Dataset**
  - **18 intents, 63 slots**
- **Regular Expressions (RE)**
  - **Written by a paid annotator**
  - **Intent: 54 REs, 1.5 hours**
  - **Slot: 60 REs, 1 hour (feature & output); 115 REs, 5.5 hours (attention)**

Experiment Setup

- **We want to answer the following questions:**
  - **Can regular expressions (REs) improve the neural network (NN) when data is limited (only use a small fraction of the training data)?**
  - **Can REs still improve NN when using the full dataset?**
  - **How does RE complexity influence the results?**

Few-Shot Learning Experiment

- **Intent Detection**
  - **Macro-F1 / Accuracy**
  - **5/10/20-shot: every intent have 5/10/20 sentences**

	5-shot	10-shot	20-shot
base	45.28 / 60.02	60.62 / 64.61	63.60 / 80.52
feat	49.40 / 63.72	64.34 / 73.46	65.16 / 83.20
ouput	46.01 / 58.68	63.51 / 77.83	69.22 / **89.25**
att	**54.86 / 75.36**	**71.23 / 85.44**	**75.58** / 88.80

RE	**70.31 / 68.98**

Regular expressions help

Few-Shot Learning Experiment

- **Intent Detection**
  - **Macro-F1 / Accuracy**
  - **5/10/20-shot: every intent have 5/10/20 sentences**

	5-shot	10-shot	20-shot
base	45.28 / 60.02	60.62 / 64.61	63.60 / 80.52
feat	49.40 / 63.72	64.34 / 73.46	65.16 / 83.20
ouput	46.01 / 58.68	63.51 / 77.83	69.22 / **89.25**
att	**54.86 / 75.36**	**71.23 / 85.44**	**75.58** / 88.80

RE	**70.31 / 68.98**

Using clue words to guide attention performs best for intent detection

Few-Shot Learning Experiment

- Slot Filling
  - Macro/Micro-F1
  - 5/10/20-shot: every intent have 5/10/20 sentences

	5-shot	10-shot	20-shot
base	60.78 / 83.91	74.28 / 90.19	80.57 / 93.08
feat	**66.84 / 88.96**	**79.67 / 93.64**	**84.95 / 95.00**
ouput	63.68 / 86.18	76.12 / 91.64	83.71 / 94.43
att	59.47 / 83.35	73.55 / 89.54	79.02 / 92.22

RE	**42.33 / 70.79**

Few-Shot Learning Experiment

- **Slot Filling**
  - **Macro/Micro-F1**
  - **5/10/20-shot: every intent have 5/10/20 sentences**

	5-shot	10-shot	20-shot
base	60.78 / 83.91	74.28 / 90.19	80.57 / 93.08
feat	**66.84 / 88.96**	**79.67 / 93.64**	**84.95 / 95.00**
ouput	63.68 / 86.18	76.12 / 91.64	83.71 / 94.43
att	59.47 / 83.35	73.55 / 89.54	79.02 / 92.22

RE	**42.33 / 70.79**

Using RE output as feature performs best for slot filling

Full Dataset Experiment

- Use all the training data
  - RE still works!

	Intent	Slot
base	92.50/98.77	85.01/95.47
feat	91.86/97.65	86.70/**95.55**
ouput	92.48/98.77	**86.94**/95.42
att	**96.20/98.99**	85.44/95.27
RE	70.31/68.98	42.33/70.79
SoA (Joint Model)	- / 98.43	-/ 95.98

Complex RE v.s. Simple RE

- **Complex RE: many semantically independant groups**

Complex RE: /(_AIRCRAFT_CODE) that fly/

Simple RE: /(_AIRCRAFT_CODE)/

	Intent		Slot	
	Complex	**Simple**	**Complex**	**Simple**
base	80.52		93.08	
feat	**83.20**	80.40	**95.00**	94.71
ouput	**89.25**	83.09	**94.43**	93.94
att	**88.80**	87.46	-	-

Complex REs yield better results

Complex RE v.s. Simple RE

- **Complex RE: many semantically independant groups**

Complex RE: /(_AIRCRAFT_CODE) that fly/

Simple RE: /(_AIRCRAFT_CODE)/

	Intent		Slot	
	Complex	**Simple**	**Complex**	**Simple**
base	80.52		93.08	
feat	83.20	80.40	95.00	94.71
ouput	89.25	83.09	94.43	**93.94**
att	88.80	**87.46**	-	-

Simple REs also clearly improves the baseline

Conclusion

- **Using REs can help to train of NN when data is limited**
- **Guiding attention is best for intent detection (sentence classification)**
- **RE output as feature is best for slot filling (sequence labeling)**
- **We can start with simple REs, and increase complexity gradually**

Q&A