National Engineering Laboratory for Speech and
Language Information Processing,
University of Science and Technology of China

Multi-Classification Model for Spoken Language Understanding

Chao-Hong Tan, Zhen-Hua Ling

Oct 18, 2019 · Suzhou, China

Outline

- Introduction
- Methods
- Experiments
- Conclusions

Outline

- **Introduction**
- Methods
- Experiments
- Conclusions

□ Introduction

➢ About Task

- Extract a set of tuples (act-slot-value triples or act-slot pairs) from users' utterances in Chinese Audio-Textual Spoken Language Understanding Challenge (CATSLU)
- Audio information and text generated by automatic speech recognition (ASR) was provided
- For example, "What's the weather like in Shunyi District today?" → (inform, area, Shunyi District), (inform, date, today), (request, weather).

➢ Baseline System provided by organizers

- Baseline system 1: a rule-based method, works in a simple string matching [Zhu et al., 2019]
- Baseline system 2: a neural network-based method, a shared utterance encoder, an act type classifier, a slot type classifier, and a value decoder [Zhao et al., 2019]

□ Introduction (Motivation)

- Tag heads
    - Value Decoding is a more difficult problem compared with classification
    - Most values in tuples are the words of input utterance
    - Thus use sentence tagging to get the value

- Pretrained language model, BERT
    - Training the utterance is difficult without enough training data
    - BERT is the state-of-the-art pretrained language model
    - BERT provides lots of prior knowledge for comprehending semantic of the utterance

- Minimum Edit Distance to reduce bias, MED
    - We only use the text from ASR as input, the ASR accuracy will limit the upper bound of our extract systems
    - To reduce this bias, MED was performed to rebuild the triples extracted from utterance

Introduction (Model Description)

- Framework
  - Multitask framework
  - Getting tuples like (act, slot, value) from one utterance, means that different task can share the input semantic
  - Archive the targets means better understand of the utterance

- Sentence encoder
  - Pretrained language model, BERT, to encode input utterance

- Task-specific output layer
  - Three different tag heads to get tuples

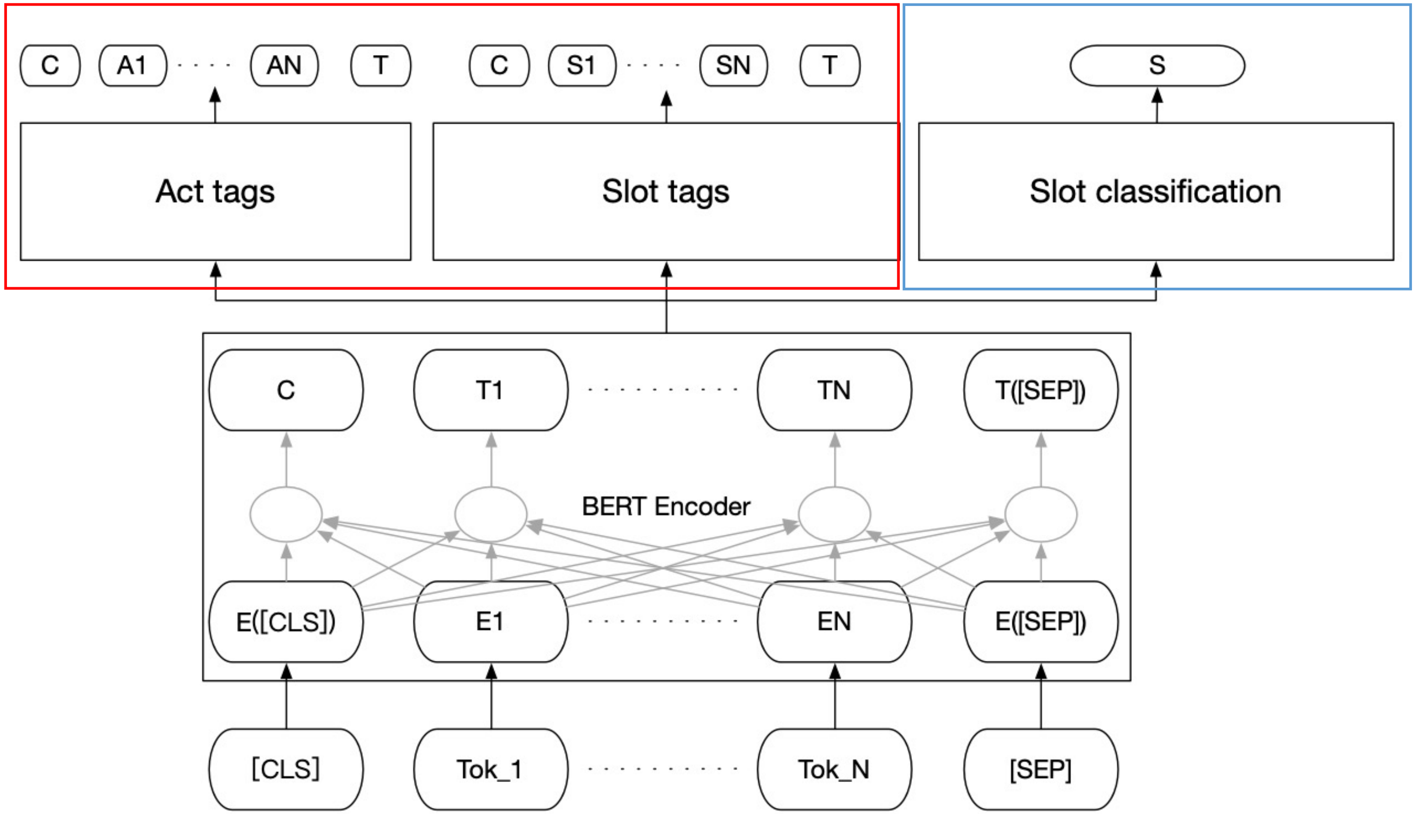
- Post processing
  - MED to repair the output from our system

□ Outline

- Introduction
- **Methods**
- Experiments
- Conclusions

Methods

Model Architecture



3 heads with a Shared Utterance Encoder
Act tags and Slot tags for (act, slot, value) triples
Slot classification for (act, slot) pairs

Methods

Slot Tags

- IOB form

  Each word is tagged as "B$-$slot$_i$", "$\mathrm{I}$$-$slot$_i$" or "O".
- Classification

  2|S| + 1. |S| means the number of slot classes in triple's acts.

Act Tags

- The number of act classes holding in act-slot-value triples.
- Only classify the words which are tagged with "B$-$slot$_i$" in slot tags.

Slot Type Classifier

- A linear layer over the top of the hidden state associated to the first character of the input ([CLS])
- Add a "NONE" label to this classifier
- |S'|+1. |S'| means the number of slot class in the pair's act

Methods

➢ Minimum Edit Distance to reduce bias (MED)

- Outputs of our system may not be covered in the candidates since the ASR error
- ASR error has a direct influence on *value*, indirect effects on *act* and *slot,* thus assume *act* and *slot* are correct
- Use the MED [Levenshtein, 1966] between values from results and candidates to rebuild the triples
- Perform MED on phonetic space maybe more suitable

Manual	导航到包埠村收费站
	Navigate to Baobu Village Toll Station
ASR best	导航到保铺村收费站
	Navigate to Baopu Village Toll Station
Tuples	[inform, 操作, 导航]，[inform, 终点名称, 包埠村收费站]
	[inform, operation, navigation], [inform, endpoint name, Baobu Village Toll Station]

□ Methods

How to train?

- ✓ Fine-tune the pretrained Chinese Language BERT
- ✓ Summarize the cross entropy loss from three tag heads
- ✓ Cut off
  - Notice that act labels is a little set and classification will be simple
  - Weak encoder of the utterance can also due with this classification
  - We truncate the back propagation of gradient before act tags head

□ Methods

How to get result (infer)?

✓ Feed an input sentence to the model and get three logits from tag heads

✓ Get tuples from the logits

- Perform **argmax** on slot type classifier to get (act, slot) pairs
- Perform **argmax** on slot tags to get (slot, value) pairs and then
- Perform **argmax** on act tags at the position where slots own the prefix "B-" to get (act, slot, value) triples

✓ Perform MED on (act, slot, value) triples

Outline

- Introduction
- Methods
- **Experiments**
- Conclusions

Experiments

CATSLU Dataset

- 4 domain: **map**, **music**, **weather**, **video**
- **map** and **music** domains contained 5,093 and 2,189 training samples
- **weather** and **video** domains contained just 341 and 205 training samples

Model for comparison

- Baseline system 1: a rule-based method, works in a simple string matching [Zhu et al., 2019]
- Baseline system 2: a neural network-based method, a shared utterance encoder, an act type classifier, a slot type classifier, and a value decoder [Zhao et al., 2019]

Experiments

Evaluation

System	Map		Music		Weather		Video	
	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)
Baseline 1	37.92	40.43	77.39	49.26	85.52	75.38	78.25	45.28
Baseline 2	77.61	74.65	81.57	71.15	85.25	78.16	75.18	57.53
Our system	**87.43**	**83.08**	**91.53**	**82.40**	**93.24**	**86.95**	**91.71**	**81.17**

- Our model achieved significantly better results than the baseline systems.
- Even in the **weather** and **video** domain, which lacks of training data, we achieve high score under F1 and accuracy metrics

Experiments

Ablation

System	Map		Music		Weather		Video	
	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)
Our system	87.43	83.08	91.53	82.40	**93.24**	**86.95**	**91.71**	**81.17**
- Cut off	**87.95**	**83.78**	**92.49**	**83.73**	92.65	86.09	90.73	79.10
- MED	81.43	77.63	88.01	78.85	90.99	84.17	84.81	72.46

- On **weather** and **video** domains cutting off performed better while **map** and **music** domains did not.
- **Map** and **music** domains have enough samples so can provide correct information for act tags to adjust the encoder
- Choice appropriate heads to joint calculate the shared encoder

Experiments

Ablation

System	Map		Music		Weather		Video	
	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)
Our system	87.43	83.08	91.53	82.40	**93.24**	**86.95**	**91.71**	**81.17**
- Cut off	**87.95**	**83.78**	**92.49**	**83.73**	92.65	86.09	90.73	79.10
- MED	81.43	77.63	88.01	78.85	90.99	84.17	84.81	72.46

- MED contributed a lot for better performance
- It repaired the gaps between ASR and ground truth

Outline

- Introduction
- Methods
- Experiments
- **Conclusions**

Conclusions

- The main contributions of our proposed method include:
  - Tag heads
  - Pretrained language model BERT
  - Minimum Edit Distance to reduce bias
  - Multi-task Learning framework

- Not all heads' information are beneficial to the shared utterance encoder

Q & A ?

Thanks!

References

1. Jacob Devlin, et al. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In North American Chapter of the Association for Computational Linguistics.
2. Abhinav Rastogi, et al. 2018. "Multitask learning for Joint Language Understanding and Dialogue State Tracking". In Annual Meeting of the Special Interest Group on Discourse and Dialogue. 376–384.
3. Zijian Zhao, et al. 2019. "A Hierarchical Decoding Model For Spoken Language Understanding From Unaligned Data". In International Conference on Acoustics Speech and Signal Processing.
4. Su Zhu, et al. 2019. "CATSLU: The 1st Chinese Audio-Textual Spoken Language Understanding Challenge". In ACM International Conference on Multimodal Interaction. (in press).