A Token-pair Framework for Information Extraction from Dialog Transcripts in SereTOD Challenge

Chenyue Wang[1,2], Xiangxing Kong[2], Mengzuo Huang[2], Feng Li[2], Jian Xing[2], Weidong Zhang[2], and Wuhe Zou[2]

[1]Peking University, Beijing, China
[2]NetEase Games AI Lab, Hangzhou, China

Abstract

This paper describes our solution for SereTOD Challenge Track 1: Information extraction from dialog transcripts. We propose a token-pair framework to simultaneously identify entity and value mentions and link them into corresponding triples. As entity mentions are usually coreferent, we adopt a baseline model for coreference resolution. We exploit both annotated transcripts and unsupervised dialogs for training. With model ensemble and post-processing strategies, our system significantly outperforms the baseline solution and ranks first in triple f1 and third in entity f1.

1 Introduction

Task-oriented dialogs cover a wide range of daily application, such as ordering food, booking tickets, and querying services. With the development of deep learning and natural language processing, AI assistants start to replace human operators in a few basic scenarios. However, correctly extracting key information in complicated contexts and generating human-like yet informative responses remain a challenge for both academia and industry.

The SereTOD 2022 Workshop introduces a challenge on mobile customer-service scenario with real-world dialog dataset (Ou et al., 2022). We mainly participate in Track 1: Information extraction from dialog transcripts, and present our token-pair framework based solution in this paper.

2 Background

The challenge provides around 100k dialog transcripts between mobile service users and staff, titled as MobileCS dataset, of which 10k are annotated while the rest are unlabeled. The annotation includes service entities and the attributes or values of the service (e.g. package price) or the user (e.g. account balance) mentioned in the dialog. As the dialogs are generally colloquial, co-references are required to be resolved for entity mentions. Moreover, values for an entity may scatter in multi-turn dialogs or nested inside the verbal expression of entities.

Track 1 is mainly formulated as an information extraction problem and contains two sub-tasks: (1) entity extraction, i.e., to extract entity mentions with their corresponding entity types as defined in the schema; (2) slot filling, i.e., to extract values for entity attributes and to match the slot-value pairs with the corresponding entity concepts. F1 score is the metric for system evaluation.

3 System Overview

3.1 Model Design

The submitted system consists of two models: an information extractor for both entity extraction and slot filling, and a co-reference resolution model for value-entity assignment.

3.1.1 Information Extractor

Recent works (Wang et al., 2020; Su et al., 2022; Li et al., 2022) on named entity recognition and information extraction shift from the conventional sequence labeling method into the token-pair approach. A token-pair based model outputs logits in the shape of $c \times n \times n$, where c denotes the number of types and n denotes sequence length, predicting over possible spans in the sequence for all the types.

Compared with previous methods, the token-pair approach has the following advantages. First and foremost, it supports nested and multilabel entities. In the MobileCS dataset, target entities are often nested due to the colloquial references. For example, the price entity *38-yuan* is nested inside the package entity *that 38-yuan package*. In addition, an entity may belong to multiple types, as defined in the schema. Such cases cannot be properly handled by sequence labeling method as a token-level classification task. Secondly, the token-pair

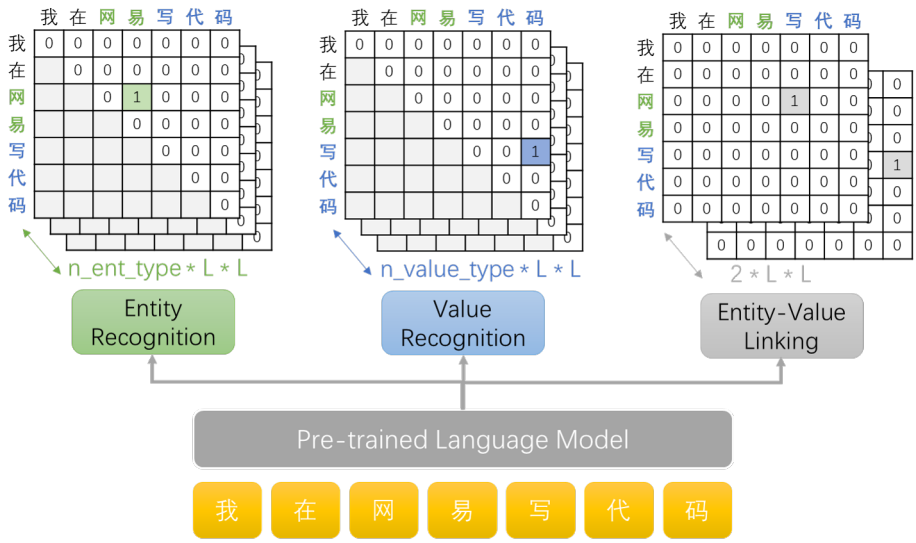

Figure 1: illustration of information extractor model, composed of an entity recognition module, a value recognition model, and an entity-value linking module. Each module is a token-pair based structure that consider all the possible spans in the input text and identify the types or relations for the recognized spans.

method directly optimizes the span-level metric and outputs straightforward result, while previous methods only focus on the token-level and require extra decoding modules such as CRF. Last but not least, the token-pair method is more versatile and flexible. Apart from NER task, it can be applied to joint information extraction and potentially other related tasks with simple modification. However, for token-pair framework, its output logits are large in quantity and extremely sparse, raising issues in model training and ensemble. Fortunately, this drawback could be alleviated.

Su et al. (2022) proposes GlobalPointer, a model structured on the token-pair framework. The encoder outputs, denoted as $[\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n]$, are transformed into queries and keys as $\mathbf{q}_i = \mathbf{W}_q\mathbf{h}_i$ and $\mathbf{k}_i = \mathbf{W}_k\mathbf{h}_i$. The score for span from i to j for type t is calculated as:

$$s_t(i,j) = \mathbf{q}_i^\mathsf{T}\mathbf{k}_j + \mathbf{w}_t^\mathsf{T}[\mathbf{q}_i; \mathbf{k}_i; \mathbf{q}_j; \mathbf{k}_j]$$

where $\mathbf{w}_t$ is a type-specific transformation.

A multi-label class-imbalance loss is proposed for countering severe class imbalance issue in the token-pair setting, where Ω_{neg} and Ω_{neg} are negative samples and positive samples, s_i and s_j are the scores for negative and positive sample:

$$\log\left(1 + \sum_{i\in\Omega_{neg}} e^{s_i}\right) + \log\left(1 + \sum_{j\in\Omega_{pos}} e^{-s_j}\right)$$

We adopt both the structure and loss design in our information extractor model.

Previous token-pair based IE models, such as GPLinker (Su, 2022) and TPLinker (Wang et al., 2020), formulate joint extraction as a token pair linking problem and introduce tagging schemes that align the boundary tokens of entity pairs under each relation type. However, entity types are not considered in the schemes.

Alternatively, we decompose the extractor model into three modules to simultaneously extract and link the entity and value mention together with their types: (1) entity recognition, (2) value recognition, and (3) entity-value linking, as illustrated in Figure 1. For a candidate span, denoted by the its start and end positions as [*i, j*], the first two modules predict whether the span text is an entity or value that belongs to the current type, while the linking module predicts the head-to-head (and tail-to-tail) matching for an entity and a value mention that starts (and ends) at position *i* and *j*, respectively. The entity and triple results can be obtained by combining the outputs of the three modules.

The extractor model is trained with a multitask loss, where $\mathcal{L}_{ent}$, $\mathcal{L}_{val}$, and $\mathcal{L}_{link}$ are the multi-label class-imbalance loss for each module:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{ent} + \lambda_2\mathcal{L}_{val} + \lambda_3\mathcal{L}_{link}$$

For simplicity, we set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ without further tuning.

3.1.2 Co-reference Resolution Model

As is required, each value should match a resolved entity concept. We adopt the entity co-reference

model from the baseline solution (Liu et al., 2022). The model transforms the embedding of the predicted entity tokens into corresponding representations by average pooling, scores candidate entity pairs, and groups them into concept clusters.

In the predicted triples, a value may correspond to multiple entity mentions. Once the concept groups for the entities are determined, we select the most matched entity's concept for each value.

We also attempt other approaches for value-entity resolution. One solution is to group the entity mentions that correspond to the same value into the same concept group. By using disjoint set, we can connect local groups into the global ones. However, this process is significantly affected by mismatched triples and achieves relatively low triple metric.

The co-reference resolution model could be integrated into the IE model and share the same encoder. Nonetheless, the co-reference resolution metrics are not stable during training and the valid result is much lower. How to integrate the co-reference resolution into the token-pair based framework remains further investigation.

3.2 Training

Our system is mainly trained on the annotated dialogs with exploitation of the unlabeled data.

3.2.1 Labeled Data

As the triple annotation only marks the value mention without detailed positions, we directly match all the value mentions in the turn utterances and add position information. Dialogs are then split into segments every 3 turns, since a majority of values have their entity mentions appear in this range. In each segment, we further supplement triples by matching values and entities that belong to the same entity group. Utterances are joined by the [SEP] token as input texts. We add specific user tokens at the beginning of each text segment, which serve as the head entity for user-attribute values. The positions and types of the entity and value spans are used as supervised signals for training the IE model.

For the co-reference resolution model, we segment the sessions with a token length of 512 and consider the inter-segment entity co-references. The details are the same as the baseline solution.

3.2.2 Unlabeled Data

We conduct domain adaptive pretraining (Gururangan et al., 2020) on the unlabeled dialog utterances to further fit the language model into the mobile service scenarios. In addition, we infer on the first 10k unlabeled dialogs with models trained on labeled data and adopt the predictions as pseudo annotations. These pseudo-labeled dialogs are then used as training data for a part of the ensemble models.

3.3 Inferring

Different from the data construction strategies in training stage, we infer on each dialog in a sliding window manner with a size of 3 turns. For the predicted triples, which are in the format of *(entity, prop, value)*, we record all distinct value mentions and their matched entities. The predicted entities are fed into the co-reference resolution model and assigned with group ids. Finally, each distinct value mention shares the same group id as its most matched entity.

In our submitted system, we ensemble a dozen of models of different pretraining methods (RoBERTa by Liu et al., 2019, MacBERT by Cui et al., 2020, etc., with or without DAPT), model scale (base or large), and training data (with or without pseudo-labels), by averaging their logits during inference. Invalid and repeated predictions are filtered during this process.

4 Discussions

4.1 Experiments and Results

We present key experiment results on validation set in Table 1 and briefly discuss the effects of the proposed strategies. Our scores and rankings in the official evaluation result are reported in Table 2. The triple-f1 ranks first among all the teams while the ent-f1 ranks third. Our averaged f1 only keeps a minor gap with the top-2 solutions.

4.1.1 Token-pair Framework

Compared with the baseline solution, our system obtains 19.48 percent absolute improvement in entity f1 and 17.06 percent in triple f1 using the same backbone model. This result proves the effectiveness of our token-pair framework. We argue that the improvement derives from better NER result, particularly for the nested and multilabel entities, as well as the joint extraction, alleviating error accumulation as in the pipeline solution. Moreover, our system is more efficient than the baseline, since we integrate three steps (i.e., named entity recognition, slot recognition, and entity slot alignment) into one IE module that shares the same encoder.

methods	entity metrics (p/r/f1)	triple metrics (p/r/f1)	#entity	#triple
RoBERTa$_{large}$ Baseline	- / - / 33.45	- / - / 34.94	-	-
RoBERTa$_{base}$ TPIE	52.87 / 53.37 / 53.12	47.88 / 38.25 / 42.53	6550	8535
w/ coref	52.87 / 53.37 / 53.12	55.07 / 44.00 / 48.92	6550	8535
w/ coref + *DAPT*	55.99 / 51.84 / 53.83	55.04 / 43.08 / 48.33	6000	8362
w/ coref + *pseudo*	57.22 / 53.42 / 55.26	58.93 / 45.15 / 51.13	6048	8185
RoBERTa$_{large}$ TPIE	53.47 / 52.39 / 52.93	52.68 / 40.79 / 45.98	6356	8272
w/ coref	53.47 / 52.39 / 52.93	59.58 / 46.13 / 52.00	6356	8272
w/ coref + *DAPT*	51.35 / 54.33 / 52.80	60.72 / 45.99 / 52.34	6887	8093
w/ coref + *pseudo*	53.81 / 53.36 / 53.58	61.37 / 44.58 / 51.65	6457	7759
Ensemble	**63.27** / 49.67 / 55.65	63.31 / 36.74 / 46.50	5083	6467
w/ coref	**63.27** / 49.67 / 55.65	**70.80** / 41.08 / 51.99	5083	6467
w/ coref + *lower thres.*	56.51 / **56.83** / **56.67**	58.55 / **53.16** / **55.72**	6527	9701

Table 1: evaluation results on dev set. The baseline result is reported in the official implementation. TPIE is our token-pair based information extractor. *DAPT* indicates domain adaptive pretraining on the LM, *pseudo* indicates training with 10k pseudo-labeled dialogs, *lower thres.* indicates adjusting threshold when inferring.

entity f1	entity ranking	triple f1	triple ranking	avg. f1	avg. ranking
55.17	3	56.07	1	55.62	3

Table 2: official evaluation result

4.1.2 Co-reference Resolution Model

As the challenge requires extracted values to be related with an entity concept, it is necessary to train a task-specific co-reference resolution model in place of the error-prone merging strategy solely based on the IE triple results. Experiment results show that better co-reference resolution results improve the triple metric by more than 5 percents.

4.1.3 Training with Unsupervised Data

Domain adaptive pretraining and pseudo labeling are the two methods for exploiting unsupervised data. As the mobile service domain differs from the general pretraining corpus, we expect DAPT to yield considerable benefit. However, the results suggest otherwise. To our surprise, training with pseudo-labeled data improves entity recognition task. Notably, the triple f1 for RoBERTa base model is significantly boosted with pseudo-labels.

4.1.4 Large Pretrained Model

Using larger pretrained model improves triple f1, which relies more on the entity-value linking module. Compared with named entity and value recognition, entity-value linking task is more complex and challenging. We argue that larger models are capable of solving such harder tasks and contribute to better performance.

4.1.5 Model Ensemble

Directly adopting model ensemble only yields marginal or even negative gains. The numbers of predicted entities and triples drop by a large portion, resulting in higher precision but lower recall. This suggests model ensemble suppresses the averaged logits and the default threshold is no longer suitable. We empirically lower the thresholds for entity and value recognition to balance precision and recall for higher f-scores.

4.2 Noisy Labels

In the initially released dataset, there exist a number of noisy entity type labels. Some clearly defined items are marked with different types. For example, the item *Two City, One Family*, is partially marked as *Long-distance Plan* and partially *Plan*, an ancestor for the former in the entity type hierarchy. Classifying an item as its ancestor type, though not perfect, is somehow acceptable. Therefore, we propose type smoothing to counter type label noise by assigning soft label weight instead of hard one-hot for entity types:

$$label_i^j = \begin{cases} w_1, & \text{if } j \text{ in ancestor types} \\ w_2, & \text{if } j \text{ is annotated type} \\ 0, & \text{otherwise} \end{cases}$$

Type label discrepancies are mostly corrected

by rule-based filtering in the later released dataset during the challenge, thus we do not adopt this strategy in our submitted system.

Span boundary issues also prevail in the annotated transcript as the entity and value mentions are typically colloquial. For example, for the expression *that 38-yuan package*, annotators may neglect *that*. Determiners and attributes as such are tricky for uniform annotations. Some value types, e.g. user demands, package rules, are too flexible to uniformly determine the mention spans.

Boundary smoothing (Zhu and Li, 2022) is a recently proposed technique to handle boundary issues for span-based models. It assigns a portion of probability ϵ from the target span $[i, j]$ to its neighboring spans whose Manhattan distances are within the smoothing size D. However, we discover a large portion of boundary noise also exist in the dev set and urge for cleaner validation samples to verify the effects of label denoising strategies.

5 Conclusion and Further Work

We present our solution for information extraction from dialog transcripts in SereTOD Challenge. The system is trained on both annotated transcripts and unsupervised dialogs. Various strategies and tricks are employed to further boost system performance, with their effects analyzed and discussed. Compared with the baseline implementation, our token-pair solution not only integrates multiple modules into a unified model framework, but also significantly outperforms the baseline result by more than 20 percent. In the official evaluation results, our system ranks first in triple-f1 and third in ent-f1.

For further work, we plan to integrate the coreference resolution model into the token-pair framework. We will evaluate the proposed label denoising methods and expect a well-anotated dataset. Detailed settings, such as multi-task weighting, shall also be tuned for better performance.

References

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Hong Liu, Hao Peng, Zhijian Ou, Juanzi Li, Yi Huang, and Junlan Feng. 2022. Information extraction and human-robot dialogue towards real-life tasks: A baseline study with the mobilecs dataset. *arXiv preprint arXiv:2209.13464*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhijian Ou, Junlan Feng, Juanzi Li, Yakun Li, Hong Liu, Hao Peng, Yi Huang, and Jiangjiang Zhao. 2022. A challenge on semi-supervised and reinforced task-oriented dialog systems. *arXiv preprint arXiv:2207.02657*.

Jianlin Su. 2022. GPLinker: A joint extraction of entities and relations based on GlobalPointer.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. *arXiv preprint arXiv:2204.12031*.