A overview of Nature Language Understanding (NLU) and SoTA Architectures so far

phucpx@ftech.ai

Research Topic	Nature Language Understanding - NLU
Keywords	NLP, NLU, Chatbot, Intent Classification, NER, ...
Author	PhucPX
Duration	15/08/2021 - Now
Planning	Plan
Achievement	- Literature review; - Proposed 3 approaches; - Training/ provided Language Model with Conversation Dataset (CoBerta Dataset); - Implemented the proposed approaches; - Experiment and Report the results of 3 approaches with 4 benchmark dataset; - Wrap to a library. [**Repository**]

1. Introduction

Evolving from basic button/ menu architecture and the keyword recognition, chatbots have now entered the domain of contextual conversation. They don't just translate but understand the speech/ text input, get smarter and sharper with every conversation and pick up on chat history and patterns. With the general advancement of linguistics, chatbots can be deployed to discern not just intents and meanings but also to better understand sentiments, sarcasm, and even tone of voice.

Natural Language Processing (NLP) is an engine for the chatbot to understand the user's intent in the message and fetch the most appropriate response from its database. Regardless of which language a computer is learning, NLP understands the syntax, semantics, discourse, and purpose of the message to engage in a human-like conversation.
There are three components of an NLP system - Natural Language Understanding (NLU), Dialogue Manager (DM) and Natural Language Generation (NLG). When you

input a text into an NLP engine, the meaning or context of the user is deciphered by the NLP construct, navigate an action by DM and the response is generated by NLG.

Process flow:
User text -> Chatbot -> NLU -> DM -> NLG -> Response Generated.

What is Natural Language Understanding (NLU) in Chatbot?
NLU can be used in a variety of applications including chatbots, customer service, sales, and virtual assistants, etc. Many bigcom are using NLU, e.g. Amazon uses it to help users complete purchases, Google uses it in its search engine and Netflix uses it to recommend movies or TV shows. NLU helps provide a better user experience to customers.
NLU is understanding the meaning of the user's input. Primarily focused on machine reading comprehension, NLU gets the chatbot to comprehend what a body of text means. NLU is nothing but an understanding of the text given and classifying it into proper intents.
Intent Classification and Slot Filling (in other words, Named Entities Recognition - NER) are two essential tasks for NLU to form a semantic parse for user utterances. IC focuses on predicting the intent of the query, while NER extracts semantic concepts.

2. Related approaches

Deep learning models have been extensively explored in NLU. We categorize NLU models into three neural architectures: *Independent models*, which IC and NER separately; *jointly models* exploit the mutual benefit of two tasks simultaneously, and *transfer learning models*, that scale the model to new domains.

Early research on IC and NER tackle these two tasks independently, where IC and NER are formulated as utterance classification and sequence labeling problems. Recent neural networks models typically for IC include: based on CNN [1] [2]; attention-based CNN [3]; hierarchical attention networks [4]; adversarial multi-task learning [5]; ULMFIT [6], among others. Approaches for NER include: based on CNN [7]; deep LSTM [8]; RNN-EM [9]; encoder-labeler LSTM [10]; flair framework [11]; among others.
Currently with these approaches, Denver-a toolbox for Natural Language Understanding has provided two main models including ULMFIT for IC and FlairSequenceTagger for NER. Maybe said that, these are the models that have achieved the best results so far in the progress of our research towards this approach for each respective problem. You can use it at the Repository and Documentation here [12].

Takeaways on independent IC and NER models:

➔ The performance of RNN encoder (unidirectional) are Jordan ≤Elman ≤ LSTM. Bi-directional encoding is additive to the performance of each encoder.
➔ Incorporating more context information is better for NER performance. Using global information, such as sentence-level representation, and attention mechanisms boosts performance of bi-directional encoding even further.
➔ The main disadvantage of independent models is that they don't exploit the interaction between intent and entities and may introduce error propagation when they are used in a pipeline.

Recent studies have shown that jointly learning intent classification and named entities recognition produces significant performance improvements over independent models. A survey of joint Intent Classification and Named Entities Recognition models in Nature Language Understanding, you can refer to [13].

In summary, two joint training strategies. The first strategy is through parameter and hidden state sharing, employing a shared BiLSTM/ BERT encoder, and two separate decoders for Intent Classification and Named Entities Recognition that are structured on top of the encoder. In particular, several research works:

- OneNet [14], already deployed in the Denver toolbox, which use a shared bi-LSTM encoder and a joint loss function between IC and NER;
- Convolutional neural network based triangular crf for joint intent detection and slot filling [15], which performs parameters sharing and captures the relation between IC and NER through Tri-CRF. The model uses CNN as a shared encoder for both tasks and the produced hidden states are utilized for IC and NER;
- Joint semantic utterance classification and slot filling with recursive neural networks [16], which a similar approach shares the node representation produced by Recursive Neural Networks (RecNN) which operates on the syntactic tree of the utterance. The node's representation is shared among IC and NER tasks;
- Attention-based recurrent neural network models for joint intent detection and slot filling [17] use a neural sequence to sequence (encoder-decoder) model with attention mechanism. The shared encoder is a bi-directional LSTM and the last hidden state of the encoder is then used by the decoder to generate a sequence of entity labels while for IC there is a separate decoder. The attention mechanism is used to learn alignments between entity labels in the decoder and words in the encoder;
- A bi-model based RNN semantic frame parsing model for intent detection and slot filling [18] proposes a bi-model based structure to learn the cross-impact between IC and NER.
- BERT for joint Intent Classification and Slot Filling [19]

The second strategy extends the first one to model the relationship between entities and intent labels. These approaches present attention mechanism to compute the correlation between a global intent context representation and each entity vector outputted by the encoder or the first learn an utterance representation (i.e. equivalent to the global intent context representation) through self-attention and concatenate this representation with each of the encoder's vector output, before feeding the concatenated vectors into a NER decoder. In particular several research works:

- Slot-Gated Modeling for Joint Slot Filling and Intent Prediction [20];
- A Self-Attentive Model with Gate Mechanism for Spoken Language Understanding [21];
- A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling [22];
- A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Lan- guage Understanding [23];
- A Joint Learning Framework With BERT for Spoken Language Understanding [24].

A comparison of the performance of the joint models in Figure 1 on ATIS and SNIPS dataset.

Method		Model	ATIS		SNIPS	
			Slot F1	Intent Acc/Err	Slot F1	Intent Acc/Err
Parameter & State Sharing						
Xu and Sarikaya (2013)	[15]	CNN + Tri-CRF	95.42	-/5.91	-	-
Guo et al. (2014)	[16]	Recursive NN	93.96	95.40	-	-
Zhang and Wang (2016)		Joint Multi-Task,Bi-GRU	95.49	98.10	-	-
Liu and Lane (2016)	[17]	Seq2Seq + Attention	94.20	91.10	87.80	96.70
Hakkani-Tür et al. (2016)		Bi-LSTM	94.30	92.60	87.30	96.90
Qin et al. (2019)	[23]	Token-Level IC + Self-Attention	95.90	96.90	94.20	98.00
Chen et al. (2019)	[19]	Transformer (BERT)	96.10	97.50	97.00	98.60
State Sharing						
Wang et al. (2018)	[18]	Bi-model, BiLSTM	96.89	98.99	-	-
Slot-Intent Gating						
Goo et al. (2018)	[20]	Slot-Gated Full Attention	94.80	93.60	88.80	97.70
Li et al. (2018)	[21]	BiLSTM + Self-Attention	96.52	-/1.23	-	-
E et al. (2019)	[22]	SF-ID Network	95.75	97.76	91.43	97.43
Hybrid Param Sharing + Gating						
Zhang et al. (2019)	[24]	BERT + Intent-Gate	98.75	99.76	98.78	98.96

Figure 1: Performance comparison of joint models for IC and NER on ATIS and SNIPS.

Takeaways on Joint IC and NER models:

- ➔ The overall performance of joint models for IC and NER is competitive with independent models. The advantages of joint models is that they have relatively less parameters than independent models, as both tasks are trained on a single model.
- ➔ Fine-tuning a pre-trained model such as BERT is the way to go for maximum IC and NER performance. Hybrid methods combining parameter and state sharing + intent gating yield the best performance.

3. Proposed approach

Transformer-based models have shown significantly better performance than the previous neural networks models (see Figure 1). Currently, Denver has provided two independent models for each IC and NER tasks, one joint model using bi-LSTM. With the outperformed results of joint Transformer-based models for NLU tasks. We propose three steps to approach the problem as follows:

- Propose experiment with Joint model based on BERT/ RoBERTa + CRF (Proposed approach 1).
- Propose experiment with Joint model based on BERT/ RoBERTa + CRF with an intent-slot attention layer to explicitly convey the intent context information via the soft intent label embedding into entity extraction (Proposed approach 2).
- Propose a pre-trained language model based on BERT/ RoBERTa architecture for Conversation AI dataset (Proposed approach 3).

3.1 Proposed approach 1

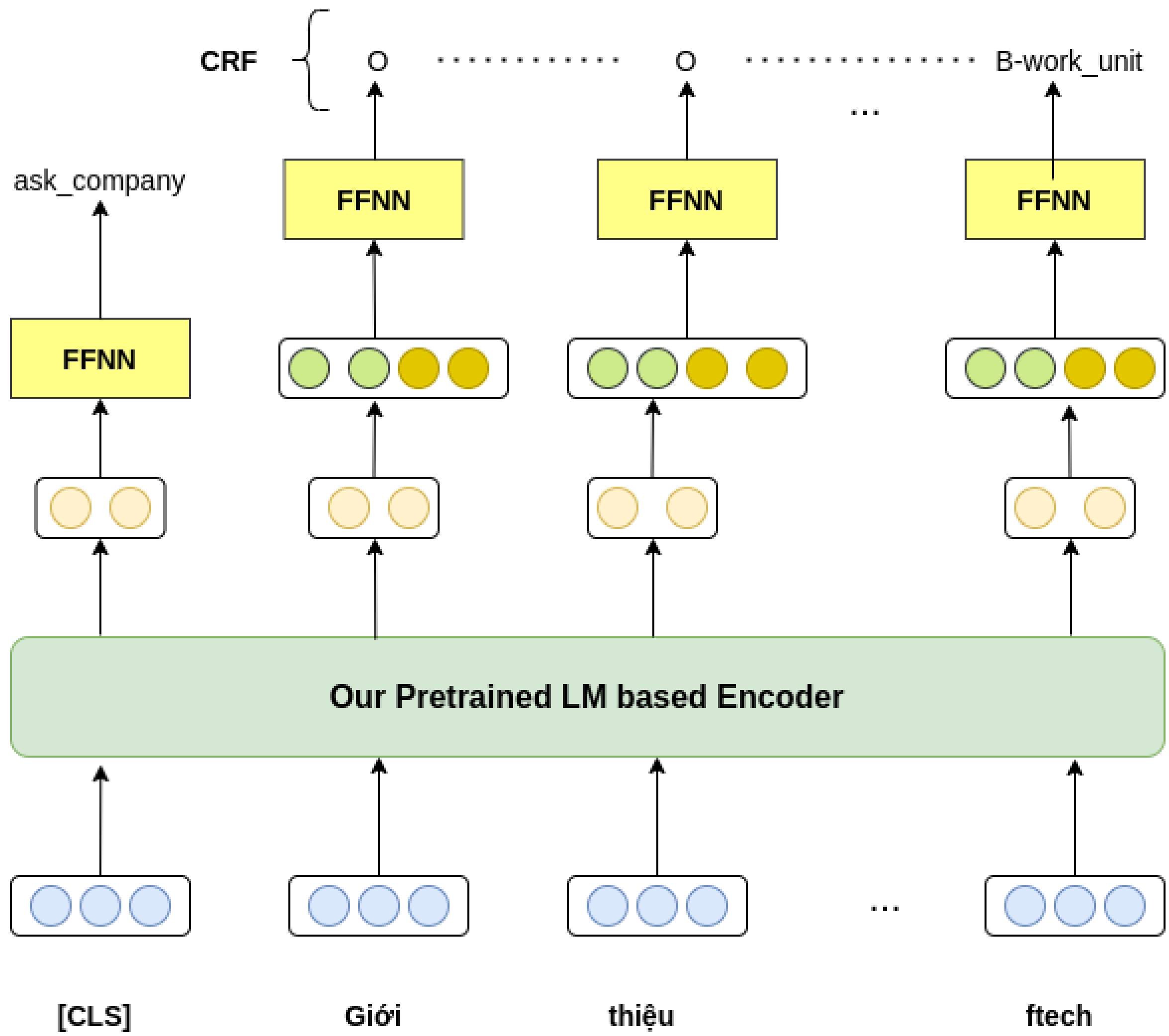

Figure 2: Illustration of proposed approach 1 model Jointly Transformer-based NLU

Figure 2 illustrates the architecture of the proposed approach 1 model, that consists of three layers including:

- An encoder layer based on BERT/ RoBERTa architecture;
- And two decoding layers of Intent classification and slot filling.

The pre-trained BERT/ RoBERTa based model provides a powerful context-dependent sentence representation and can be used for various tasks, i.e. IC and NER, through fine-tuning procedure. We can absolutely use PhoBERT - the recent state-of-the-art pre-trained LM that supports Vietnamese as this pre-trained language model. PhoBERT, a monolingual variant of RoBERTa for Vietnamese, is pre-trained on 20GB of word-level texts.

BERT can be easily extended to a joint IC and NER model. Based on the hidden state of the first special token ([CLS]), denoted h_1, the intent is predicted as:

$$y^i = softmax(W^i h_1 + b^i)$$

For slot filling, we feed the final hidden states of other tokens $h_2, \ldots, h_T$ into a softmax layer to classify over the slot filling labels. To make this procedure compatible with the WordPiece tokenization, we feed each tokenized input word into a WordPiece tokenizer and use the hidden state corresponding to the first sub-token as input to the softmax classifier.

$$y^s_n = softmax(W^s h_n + b^s),\ n \in 1 \ldots N$$

Where h_n is the hidden state corresponding to the first sub-token of word x_n.

The loss function combines the cross-entropy loss between IC and NER tasks.

3.2 Proposed approach 2

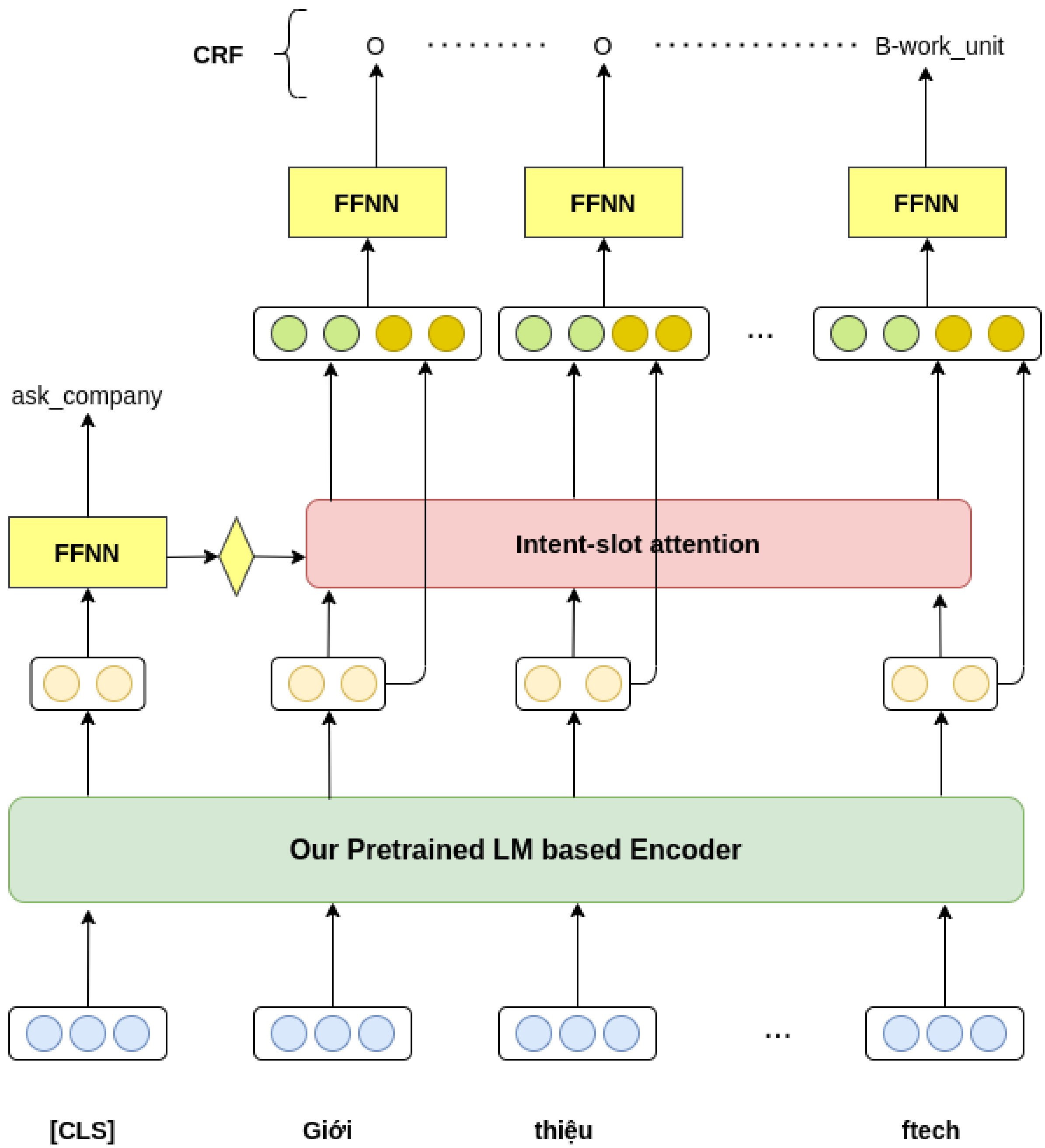

Figure 3: Illustration of proposed approach 2 model Jointly Transformer-based NLU

Figure 3 illustrates the architecture of the proposed approach 2 model, that consists of four layers including:

- An encoder layer;
- An intermediate intent-slot attention layer;
- And two decoding layers of Intent classification and slot filling.

Encoding layer:

Given an utterance consisting of n tokens w_1, w_2, ..., w_n, insert a special classification token [CLS] at the front of the utterance, resulting in an input utterance of $n + 1$token w_0, w_1, w_2, ..., w_n for the encoding layer. The encoding layer employs a pre-trained Transformer-based LM, e.g. BERT, PhoBERT, to procedure contextualized latent feature embedding $c_i \in R^{d_e}$ each representation the i^{th} token w_i:

$$c_i = PretrainedLM(w_{o:n}, i)$$

Here, d_e is the embedding size of the encoder's contextualized embedding outputs.

Intent classification layer:

The intent classification layer is a linear prediction layer that is appended on top the contextualized embedding c_0 of classification token [CLS]. In particular, the intent detection layer feed c 0 into a single-layer feed-forward network (FFNN) followed by a softmax predictor for intent prediction:

$$p = softmax(FFNN_{IC}(c_0))$$

Where, the output size of $FFNN_{IC}$is k being the total number of intent labels. Based on the probability vector $p \in R^k$, a cross-entropy objective loss L_{IC} is calculated for intent classification during training.

Intent-slot attention layer:

The attention mechanism to align the importance of the intent information with each of the original utterance's tokens. In particular, the intent-slot attention layer takes the outputs from the encoding layer and the intent detection layer to produce intent-specific vectors that are then used as part of the input for the slot filling layer. Formally, this attention layer first creates a "soft" intent label embedding $w \in R^{d_e}$ by multiplying a label weight matrix $W \in R^{d_e \times k}$ with the probability vector $p \in R^k$. Then it uses the intent label embedding w and the contextualized embeddings c_i to generate the intent-specific vectors $s_i (i \in \{1, 2, ..., n\})$ as follows:

$$w = Wp$$

$$\alpha_i = \frac{exp(w^T c_i)}{\sum_{j=1}^{n} exp(w^T c_j)}$$

$$s_i = \alpha_i w$$

Slot filling layer:

The slot filling layer formulates the slot filling task as a BIO-based sequence labeling problem. First, it creates a sequence of vectors $v_{1:n}$ in which each v_i is resulted in by concatenating the intent-specific vector s_i and the corresponding contextualized embedding c_i:

$$v_i = s_i \circ c_i$$

Then pass each vector v_i into another $FNNN_{NER}$:

$$h_i = FFNN_{NER}(v_i)$$

Where, the output layer size of $FFNN_{NER}$ is the number of BIO- based slot types. Lastly, the slot filling layer feeds the output vectors h_i into a linear-chain CRF predictor for slot type prediction. Cross-entropy loss L_{NER} is computed for slot filling during training while the Viterbi algorithm is used for inference.

The training objective loss L of joint model is weighted sum of the intent classification loss L_{IC} and the slot filling loss L_{NER}:

$$L = \lambda L_{IC} + (1 - \lambda) L_{NER}$$

Where, the hyper-parameter λ is a mixture weight: $0 < \lambda < 1$.

3.3 Propose a Pre-train Language Model from Conversation Corpus

We can crawl data from famous forums, such as: Voz, F319, … (see that), use data from Wikipedia and news, such as: Vietnamnet, baomoi, … to train a pre-trained language model used for Conversation AI.

Named: **Co-Dataset** (~ 5G data) -> Pretrained LM: **CoBERTa**

3.3.1 Crawl data

Crawl data from 8 popular forums, obtained the raw data set stored here. In addition, we look for the amount of data from the wiki, comments on facebook are also included for use.

3.3.2 Preprocessing data

Steps to perform data processing:

- Split each line by character \n;

- Normalize unicode data (NFKC mode);
- Remove special characters like: ^, ~, #, (,), non-Vietnamese or English characters;
- Separates special characters and types with the word adjacent to it by a space;
- For each line is an example (are examples already split or split in an same article by \n)

3.3.3 Training a pretrain Language Model

You can refer to [this documentation].

We release some versions for CoBERTa Pretrained - Language Model:

Model	#params	Arch.	Pre-train data
coberta-tiny-850	5.9M	tiny-900	850MB
coberta-small-850		small-900	850MB
coberta-small-1.4		small-1.4	1.4GB
coberta-small -5.0		small.5.0	5GB
coberta-medium-5.0		medium-5.0	5GB

4. Experiments and Results

4.1 Benchmark Dataset

4.1.1 CometV3 dataset

Download: Train dataset | Test dataset

	TRAIN DATASET			
ID	Intent	Numbers	Entities	Numbers
1	agree	128	ask_confirm#age_of_use	166
2	ask_confirm	1792	ask_confirm#brand	169

3	ask_is_bot	128	ask_confirm#color	177
4	deny	1015	ask_confirm#guarantee	164
5	greet	128	ask_confirm#material	168
6	inform	1262	ask_confirm#object_type	166
7	request#age_of_use	129	ask_confirm#origin	169
8	request#brand	149	ask_confirm#price	172
9	request#color	109	ask_confirm#promotion	83
10	request#guarantee	125	ask_confirm#size	170
11	request#image	130	ask_confirm#weight	160
12	request#link	77	deny#brand	176
13	request#material	127	deny#color	174
14	request#origin	123	deny#material	163
15	request#price	151	deny#object_type	168
16	request#promotion	116	deny#origin	63
17	request#size	128	deny#price	159
18	request#weight	99	deny#size	123
19	thanks	131	inform#age_of_use	170
20			inform#brand	515
21			inform#color	234
22			inform#guarantee	170
23			inform#material	165
24			inform#object_type	1878
25			inform#origin	167
26			inform#price	169
27			inform#size	162
28			inform#weight	122

	TEST DATASET			
ID	Intent	Numbers	Entities	Numbers
1	agree	23	ask_confirm#age_of_use	31
2	ask_confirm	315	ask_confirm#brand	31
3	ask_is_bot	23	ask_confirm#color	22
4	deny	179	ask_confirm#guarantee	32
5	greet	23	ask_confirm#material	26
6	inform	230	ask_confirm#object_type	32
7	request#age_of_use	21	ask_confirm#origin	30
8	request#brand	25	ask_confirm#price	28
9	request#color	18	ask_confirm#promotion	17
10	request#guarantee	22	ask_confirm#size	29
11	request#image	22	ask_confirm#weight	35
12	request#link	13	deny#brand	24
13	request#material	22	deny#color	24
14	request#origin	21	deny#material	28
15	request#price	28	deny#object_type	30
16	request#promotion	19	deny#origin	20
17	request#size	21	deny#price	32
18	request#weight	19	deny#size	22
19	thanks	24	inform#age_of_use	30
20			inform#brand	84
21			inform#color	34
22			inform#guarantee	27
23			inform#material	31
24			inform#object_type	349

25			inform#origin	31
26			inform#price	31
27			inform#size	35
28			inform#weight	15

4.1.2 LTA dataset

Download: Train dataset | Test dataset

4.1.3 Kcloset dataset

Download: Train dataset | Test dataset

	TRAIN DATASET			
ID	**Intent**	**Numbers**	**Entities**	**Numbers**
1	agree	246	ask_availability#color	39
2	ask_about_ship	314	ask_if_sells#color	70
3	ask_availability	875	inform#age	1
4	ask_availability#size	498	inform#color	327
5	ask_if_fit	645	inform#gender	873
6	ask_if_sells	941	inform#height	778
7	ask_if_sells#online	378	inform#location	691
8	ask_if_sells#size	588	inform#material	216
9	ask_is_bot	138	inform#mention	2302
10	greet	268	inform#object_type	3698
11	handoff	903	inform#season	48
12	inform	959	inform#weight	2337

13	order_request	546		
14	request#color	112		
15	request#material	280		
16	request#price	551		
17	request#product	406		
18	request#promotion	435		
19	request_more#color	294		
20	request_shop_address	421		
21	thank	264		

	TEST DATASET			
ID	**Intent**	**Numbers**	**Entities**	**Numbers**
1	agree	43	ask_availability#color	4
2	ask_about_ship	58	ask_if_sells#color	7
3	ask_availability	151	inform#age	346
4	ask_availability#size	88	inform#color	59
5	ask_if_fit	118	inform#gender	184
6	ask_if_sells	177	inform#height	130
7	ask_if_sells#online	64	inform#location	113
8	ask_if_sells#size	101	inform#material	35
9	ask_is_bot	25	inform#mention	415
10	greet	50	inform#object_type	625
11	handoff	146	inform#season	12
12	inform	174	inform#weight	442
13	order_request	91		
14	request#color	19		

15	request#material	52		
16	request#price	96		
17	request#product	67		
18	request#promotion	77		
19	request_more#color	55		
20	request_shop_address	78		
21	thank	46		

4.1.4 PhoATIS dataset

Download: Train dataset | Test dataset

	TRAIN DATASET			
ID	**Intent**	**Numbers**	**Entities**	**Numbers**
1	abbreviation	130	aircraft_code	31
2	aircraft	72	airline_code	126
3	aircraft#flight#flight_no	2	airline_name	639
4	airfare	375	airport_code	25
5	airfare#flight	29	airport_name	35
6	airline	126	arrive_date.date_relative	9
7	airline#flight_no	2	arrive_date.day_name	76
8	airport	17	arrive_date.day_number	42
9	capacity	15	arrive_date.month_name	42
10	city	19	arrive_date.today_relative	1
11	city#flight_time	1	arrive_time.end_time	17
12	distance	18	arrive_time.period_mod	7

13	flight	3283	arrive_time.period_of_day	43
14	flight#flight_no	1	arrive_time.start_time	18
15	flight#flight_time	1	arrive_time.time	188
16	flight_no	11	arrive_time.time_relative	171
17	flight_no#flight_time	1	city_name	204
18	flight_time	48	class_type	191
19	ground_fare	18	connect	36
20	ground_fare#ground_service	1	cost_relative	310
21	ground_service	237	day_name	3
22	meal	7	day_number	2
23	quantity	42	days_code	2
24	restriction	22	depart_date.date_relative	76
25			depart_date.day_name	787
26			depart_date.day_number	357
27			depart_date.month_name	340
28			depart_date.today_relative	75
29			depart_date.year	21
30			depart_time.end_time	24
31			depart_time.period_mod	69
32			depart_time.period_of_day	497
33			depart_time.start_time	24
34			depart_time.time	331
35			depart_time.time_relative	284
36			economy	34
37			fare_amount	46
38			fare_basis_code	64
39			flight_days	37

40			flight_mod	292
41			flight_number	78
42			flight_stop	143
43			flight_time	57
44			fromloc.airport_code	14
45			fromloc.airport_name	70
46			fromloc.city_name	3890
47			fromloc.state_code	41
48			fromloc.state_name	35
49			meal	43
50			meal_code	6
51			meal_description	49
52			mod	29
53			month_name	2
54			or	67
55			period_of_day	2
56			restriction_code	21
57			return_date.date_relative	7
58			return_date.day_name	1
59			return_date.day_number	2
60			return_date.month_name	2
61			return_date.today_relative	1
62			return_time.period_mod	2
63			return_time.period_of_day	24
64			round_trip	322
65			state_code	7
66			state_name	1

67			stoploc.airport_name	1
68			stoploc.city_name	217
69			stoploc.state_code	5
70			time	2
71			time_relative	1
72			today_relative	2
73			toloc.airport_code	16
75			toloc.airport_name	36
76			toloc.city_name	3918
77			toloc.country_name	3
78			toloc.state_code	75
79			toloc.state_name	71
80			transport_type	40

	TEST DATASET			
ID	**Intent**	**Numbers**	**Entities**	**Numbers**
1	abbreviation	51	aircraft_code	33
2	aircraft	10	airline_code	35
3	airfare	47	airline_name	101
4	airfare#flight	9	airport_code	9
5	airline	17	airport_name	22
6	airline#flight_no	0	arrive_date.date_relative	2
7	airport	18	arrive_date.day_name	11
8	capacity	20	arrive_date.day_number	6
9	city	3	arrive_date.month_name	6
10	distance	9	arrive_time.end_time	8

11	flight	641	arrive_time.period_mod	0
12	flight_no	10	arrive_time.period_of_day	7
13	flight_time	1	arrive_time.start_time	8
14	ground_fare	6	arrive_time.time	35
15	ground_service	36	arrive_time.time_relative	30
16	meal	4	city_name	54
17	quantity	8	class_type	24
18	restriction	3	connect	6
19			cost_relative	37
20			day_name	2
21			days_code	1
22			depart_date.date_relative	17
23			depart_date.day_name	212
24			depart_date.day_number	55
25			depart_date.month_name	56
26			depart_date.today_relative	9
27			depart_date.year	3
28			depart_time.end_time	3
29			depart_time.period_mod	10
30			depart_time.period_of_day	119
31			depart_time.start_time	3
32			depart_time.time	64
33			depart_time.time_relative	65
34			economy	6
35			fare_amount	2
36			fare_basis_code	17
37			flight_days	10

38			flight_mod	24
39			flight_number	11
40			flight_stop	21
41			flight_time	1
42			fromloc.airport_code	5
43			fromloc.airport_name	12
44			fromloc.city_name	705
45			fromloc.state_code	23
46			fromloc.state_name	17
47			meal	16
48			meal_code	1
49			meal_description	10
50			mod	2
51			or	7
52			period_of_day	2
53			restriction_code	4
54			return_date.date_relative	3
55			return_date.day_name	2
56			return_date.today_relative	0
57			round_trip	73
58			state_code	1
59			state_name	10
60			stoploc.city_name	20
61			toloc.airport_code	4
62			toloc.airport_name	3
63			toloc.city_name	717
64			toloc.country_name	1

65			toloc.state_code	18
66			toloc.state_name	28
67			transport_type	10

4.2 Results

4.2.1 CometV3 dataset

Models	Intent (F1)	Entities (F1)
OneNet	0.9793	0.9162
Joint-CoBERTa + PhoBERT	0.9703	0.9273
Joint-CoBERTa + CoBERTa-tiny-850		
JointCoBERTa + PhoBERT + Use_intent_context_concat (hard)	0.9786	0.9345
JointCoBERTa + PhoBERT + use_intent_context_attention (soft)	0.9777	0.9369
Joint-CoBERTa + CoBERTa-tiny-850 + use_intent_context_attention		

4.1.2 LTA dataset

Models	Intent (F1)	Entities (F1)
OneNet		
Joint-CoBERTa + PhoBERT-LM		
Joint-CoBERTa + CoBERTa-LM		
JointCoBERTa + PhoBERT + use_intent_context_concat (hard)		

JointCoBERTa + PhoBERT + Use_intent_context_attention (soft)		
Joint-CoBERTa + CoBERTa + use_intent_context_attention		

4.1.3 Kcloset dataset

Models	Intent (F1)	Entities (F1)
OneNet	0.9347	0.9159
Joint-CoBERTa + PhoBERT	0.9412	0.9174
Joint-CoBERTa + CoBERTa-tiny-850	0.8586	0.8320
Joint-CoBERTa + PhoBERT + Use_intent_context_concat (hard)	0.9425	0.9154
JointCoBERTa + PhoBERT + use_intent_context_attention (soft)	0.9369	0.9231
Joint-CoBERTa + CoBERTa + use_intent_context_attention		

4.1.4 PhoATIS dataset

Models	Intent (F1)	Entities (F1)
OneNet	0.9538	0.9368
Joint-CoBERTa + PhoBERT	0.9748	0.9432
Joint-CoBERTa + CoBERTa-tiny-850		
JointCoBERTa + PhoBERT + use_intent_context_concat (hard)		
JointCoBERTa + PhoBERT +		

use_intent_context_attention (soft)		
Joint-CoBERTa + CoBERTa + use_intent_context_attention		

References

[1] Yoon Kim, "*Convolutional neural networks for sentence classification*". In EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1746– 1751. ACL.

[2] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun, "*Character-level convolutional networks for text classification*". In NIPS 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 649–657.

[3] Zhiwei Zhao and Youzheng Wu, "*Attention- based convolutional neural networks for sentence classification*". In Interspeech 2016, San Francisco, CA, USA, September 8-12, 2016, pages 705–709. ISCA.

[4] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy, "*Hierarchical attention networks for document classification*". In NAACL HLT 2016, San Diego California, USA, June 12-17, 2016, pages 1480–1489. The Association for Computational Linguistics.

[5] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang, "*Adversarial multi-task learning for text classification*". In ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pages 1–10. As- sociation for Computational Linguistics.

[6] Jeremy Howard, Sebastian Ruder, "*Universal Language Model Fine-tuning for Text Classification*". In arXiv:1801.06146 2018.

[7] Ngoc Thang Vu, "*Sequential convolutional neu- ral networks for slot filling in spoken language understanding*". In Interspeech 2016, 17th Annual Con- ference of the International Speech Communication Association, San Francisco, CA, USA, September 8- 12, 2016, pages 3250–3254. ISCA.

[8] Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Ge- offrey Zweig, and Yangyang Shi, "*Spoken lan- guage understanding using long short-term memory neural networks*". In 2014 IEEE Spoken Language Technology Workshop, SLT 2014, South Lake Tahoe, NV, USA, December 7-10, 2014, pages 189–194.

[9] Baolin Peng, Kaisheng Yao, Li Jing, and Kam-Fai Wong, "*Recurrent neural networks with ex- ternal memory for spoken language understanding*". In NLPCC 2015, Nanchang, China, October 9-13, 2015, Proceedings, pages 25–35.

[10] Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu, "*Leveraging sentence-level information with encoder LSTM for natural language understanding*". CoRR, abs/1601.01530.

[11] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, Roland Vollgraf, "*FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP*". In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.

[12] Phuc Phan Xuan, "*Denver - a simple, easy-to-use library and toolbox for Language Understandings*". In Ftech 2021.

[13] H. Weld, X. Huang, S. Long, J. Poon, and S. Han, "*A survey of joint intent detection and slot-filling models in natural language understanding*". In arXiv preprint, vol. arXiv:2101.08091, 2021.
[14] Young-Bum Kim, Sungjin Lee, Karl Stratos, "*OneNet: Joint Domain, Intent, Slot Prediction for Spoken Language Understanding*". In arXiv:1801.05149, 2018.
[15] Puyang Xu and Ruhi Sarikaya, "*Convolutional neural network based triangular crf for joint intent detection and slot filling*". In 2013 ieee workshop on automatic speech recognition and understanding, pages 78–83. IEEE.
[16] Daniel Guo, Gökhan Tür, Wen-tau Yih, and Geoffrey Zweig, "*Joint semantic utterance classification and slot filling with recursive neural networks*". In 2014 IEEE Spoken Language Technology Workshop, SLT 2014, South Lake Tahoe, NV, USA, December 7-10, 2014, pages 554–559. IEEE.
[17] Bing Liu and Ian Lane, "*Attention-based recurrent neural network models for joint intent detection and slot filling*". In Nelson Morgan, editor, Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016, pages 685–689. ISCA.
[18] Yu Wang, Yilin Shen, and Hongxia Jin, "*A bi-model based RNN semantic frame parsing model for intent detection and slot filling*". In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics.
[19] Q. Chen, Z. Zhuo, and W. Wang, "*BERT for Joint Intent Classi- fication and Slot Filling*". In arXiv preprint, vol. arXiv:1902.10909, 2019.
[20] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, "*Slot-Gated Modeling for Joint Slot Filling and Intent Prediction*". In Proceedings of NAACL, 2018, pp. 753–757.
[21] C. Li, L. Li, and J. Qi, "*A Self-Attentive Model with Gate Mech- anism for Spoken Language Understanding*". In Proceedings of EMNLP, 2018, pp. 3824–3833.
[22] H. E, P. Niu, Z. Chen, and M. Song, "*A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling*". In Proceedings of ACL, 2019, pp. 5467–5471.
[23] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "*A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Lan- guage Understanding*". In Proceedings of EMNLP-IJCNLP, 2019, pp. 2078–2087.
[24] Z. Zhang, Z. Zhang, H. Chen, and Z. Zhang, "*A Joint Learning Framework With BERT for Spoken Language Understanding*". In IEEE Access, vol. 7, pp. 168 849–168 858, 2019.

Appendix

Additional comparison

<Based on when training Kcloset dataset>

Models	Training Time	Inference	Size of Model
`OneNet`	0:15:08.46 (s)	0:00:08.02 (s)	6.1M
`Joint-CoBERTa + PhoBERT + *`[1]	0:57:38.25 (s)	0:00:38.83 (s)	516M
`Joint-CoBERTa + CoBERTa-tiny-850`	1:44:26 (s)	0:00:07.72 (s)	23M
`Joint-CoBERTa + CoBERTa-small`			

OneNet

Parameters

```
hyper_params:
  # The learning rate
  learning_rate: 0.001
  # The batch size
  batch_size: 128
  # The number epochs to training
  num_epochs: 150
```

CometV3 Dataset

[1] Maybe have intent context embedding.

```
Intent results:
+----------+---------+-----------+--------+
| accucary | f1-score | precision | recall |
+----------+---------+-----------+--------+
|  0.9794  |  0.9793  |  0.9795   | 0.9794 |
+----------+---------+-----------+--------+
Intent detailed results:
                    precision    recall  f1-score   support

              agree       1.00      1.00      1.00        23
        ask_confirm       0.98      0.99      0.98       316
         ask_is_bot       1.00      1.00      1.00        23
               deny       0.99      1.00      1.00       179
              greet       1.00      1.00      1.00        23
             inform       0.97      0.97      0.97       223
request#age_of_use       0.96      0.96      0.96        23
      request#brand       0.93      0.96      0.95        27
      request#color       1.00      0.95      0.97        19
  request#guarantee       1.00      1.00      1.00        22
      request#image       1.00      0.91      0.95        23
       request#link       1.00      1.00      1.00        13
   request#material       1.00      1.00      1.00        22
     request#origin       0.95      0.95      0.95        21
      request#price       0.93      0.96      0.95        27
  request#promotion       0.95      0.86      0.90        21
       request#size       0.95      0.91      0.93        22
     request#weight       1.00      1.00      1.00        18
             thanks       1.00      1.00      1.00        23

           accuracy                           0.98      1068
          macro avg       0.98      0.97      0.97      1068
       weighted avg       0.98      0.98      0.98      1068
```

Classification report of Intent classification with OneNet

```
Tags results:
+----------+----------+--------+
| f1-score | precision | recall |
+----------+----------+--------+
|  0.9162  |  0.9077   | 0.9249 |
+----------+----------+--------+
Tags detailed results:
                         precision    recall  f1-score   support

     inform#object_type       0.90      0.93      0.91       348
ask_confirm#object_type       0.91      0.97      0.94        32
        inform#material       0.97      0.97      0.97        31
      ask_confirm#price       0.90      0.96      0.93        28
           inform#color       0.88      0.85      0.87        34
           inform#price       0.87      0.87      0.87        31
      ask_confirm#brand       1.00      0.94      0.97        31
   ask_confirm#material       0.92      0.88      0.90        26
             deny#price       1.00      0.97      0.98        32
     ask_confirm#origin       1.00      1.00      1.00        30
           inform#brand       0.94      0.95      0.95        84
             deny#brand       0.96      1.00      0.98        24
       ask_confirm#size       0.80      0.83      0.81        29
       inform#guarantee       0.96      1.00      0.98        27
      ask_confirm#color       0.84      0.95      0.89        22
  ask_confirm#guarantee       0.75      0.84      0.79        32
       deny#object_type       0.69      0.73      0.71        30
     ask_confirm#weight       0.97      1.00      0.99        35
          inform#origin       0.97      0.94      0.95        31
      inform#age_of_use       0.93      0.87      0.90        30
            inform#size       0.88      0.83      0.85        35
             deny#color       0.88      0.92      0.90        24
              deny#size       0.96      1.00      0.98        22
  ask_confirm#promotion       0.79      0.88      0.83        17
          inform#weight       1.00      0.93      0.97        15
          deny#material       1.00      1.00      1.00        28
 ask_confirm#age_of_use       0.88      0.90      0.89        31
            deny#origin       1.00      1.00      1.00        20

              micro avg       0.91      0.92      0.92      1159
              macro avg       0.91      0.92      0.92      1159
```

Classification report of NER with OneNet
(CometV3 dataset)

Kcloset Dataset

Intent results:

accucary	f1-score	precision	recall
0.9358	0.9347	0.9364	0.9358

Intent detailed results:

	precision	recall	f1-score	support
agree	0.95	0.98	0.97	43
ask_about_ship	0.92	0.98	0.95	55
ask_availability	0.97	0.97	0.97	154
ask_availability#size	0.97	0.97	0.97	88
ask_if_fit	0.94	0.99	0.97	114
ask_if_sells	0.95	0.99	0.97	166
ask_if_sells#online	0.94	0.93	0.93	67
ask_if_sells#size	1.00	0.97	0.99	104
ask_is_bot	1.00	1.00	1.00	25
greet	0.98	0.98	0.98	47
handoff	0.85	0.70	0.77	159
inform	0.94	0.86	0.90	169
order_request	0.98	0.93	0.95	96
request#color	0.95	0.95	0.95	20
request#material	0.98	1.00	0.99	50
request#price	0.98	0.99	0.98	97
request#product	0.77	0.94	0.85	72
request#promotion	0.88	0.94	0.91	77
request_more#color	0.94	0.96	0.95	52
request_shop_address	0.90	0.97	0.94	74
thank	0.96	0.98	0.97	47
accuracy			0.94	1776
macro avg	0.94	0.95	0.94	1776
weighted avg	0.94	0.94	0.93	1776

Classification report of Intent classification with OneNet (KCloset dataset)

```
Tags results:
+----------+-----------+--------+
| f1-score | precision | recall |
+----------+-----------+--------+
|  0.9159  |   0.907   | 0.925  |
+----------+-----------+--------+
Tags detailed results:
                        precision    recall  f1-score   support

         inform#weight       0.97      0.96      0.96       442
    inform#object_type       0.87      0.92      0.90       625
         inform#height       0.88      0.94      0.91       130
         inform#gender       0.95      0.95      0.95       184
            inform#age       0.97      0.97      0.97       346
          inform#color       0.84      0.81      0.83        59
       inform#location       0.68      0.72      0.70       113
        inform#mention       0.93      0.93      0.93       415
       inform#material       0.76      0.83      0.79        35
ask_availability#color       1.00      1.00      1.00         4
         inform#season       0.75      0.75      0.75        12
    ask_if_sells#color       0.75      0.86      0.80         7

             micro avg       0.91      0.92      0.92      2372
             macro avg       0.91      0.92      0.92      2372
```

Classification report of NER with OneNet (KCloset dataset)

PhoATIS Dataset

```
Intent results:
+----------+---------+----------+--------+
| accucary | f1-score | precision | recall |
+----------+---------+----------+--------+
|  0.9608  |  0.9538  |  0.9572   | 0.9608 |
+----------+---------+----------+--------+
Intent detailed results:
                   precision    recall  f1-score   support

     abbreviation       0.98      0.98      0.98        52
         aircraft       1.00      0.67      0.80         9
  aircraft#flight       0.00      0.00      0.00         1
          airfare       0.92      1.00      0.96        45
   airfare#flight       1.00      0.35      0.52        17
          airline       1.00      1.00      1.00        17
   airline#flight       0.00      0.00      0.00         1
airline#flight_no       0.00      0.00      0.00         1
          airport       1.00      1.00      1.00        18
         capacity       0.87      0.95      0.91        21
             city       0.75      0.50      0.60         6
         day_name       0.00      0.00      0.00         2
         distance       1.00      0.50      0.67        10
           flight       0.97      1.00      0.98       626
        flight_no       1.00      1.00      1.00         8
      flight_time       1.00      1.00      1.00         1
      ground_fare       1.00      0.86      0.92         7
   ground_service       0.90      1.00      0.95        36
             meal       1.00      0.67      0.80         6
         quantity       0.78      1.00      0.88         7
      restriction       0.67      1.00      0.80         2

         accuracy                           0.96       893
        macro avg       0.75      0.69      0.70       893
     weighted avg       0.96      0.96      0.95       893
```

Classification report of Intent classification with OneNet (PhoATIS dataset)

Tags results:

f1-score	precision	recall
0.9368	0.9354	0.9381

Tags detailed results:

	precision	recall	f1-score	support
round_trip	0.97	0.99	0.98	73
depart_time.time	0.97	0.98	0.98	64
toloc.city_name	0.96	0.99	0.97	717
arrive_date.day_name	0.69	0.82	0.75	11
depart_time.period_of_day	0.97	0.97	0.97	119
toloc.state_name	0.92	0.82	0.87	28
fromloc.city_name	0.97	0.99	0.98	705
aircraft_code	0.57	0.61	0.59	33
airline_name	0.94	0.91	0.92	101
fromloc.state_name	0.76	0.76	0.76	17
fromloc.airport_name	0.38	0.67	0.48	12
depart_time.time_relative	0.98	0.97	0.98	65
arrive_time.time_relative	1.00	0.93	0.97	30
meal_description	1.00	0.90	0.95	10
depart_time.period_mod	0.89	0.80	0.84	10
arrive_date.month_name	0.62	0.83	0.71	6
depart_date.day_name	0.98	0.98	0.98	212
flight_stop	1.00	1.00	1.00	21
airport_name	0.80	0.36	0.50	22
depart_date.month_name	0.98	0.95	0.96	56
depart_date.day_number	0.98	0.95	0.96	55
class_type	1.00	1.00	1.00	24
arrive_time.time	0.97	0.94	0.96	35
cost_relative	1.00	0.97	0.99	37
mod	0.25	0.50	0.33	2
airline_code	0.85	0.97	0.91	35
stoploc.city_name	0.91	1.00	0.95	20
city_name	0.81	0.54	0.64	54
fare_basis_code	0.94	0.94	0.94	17
connect	0.83	0.83	0.83	6
flight_number	0.82	0.82	0.82	11
state_name	0.00	0.00	0.00	10
flight_mod	0.65	0.71	0.68	24
toloc.state_code	1.00	1.00	1.00	18
arrive_date.date_relative	1.00	0.50	0.67	2
fromloc.airport_code	0.50	0.20	0.29	5
depart_date.date_relative	0.94	1.00	0.97	17
depart_date.today_relative	1.00	0.78	0.88	9
fromloc.state_code	1.00	1.00	1.00	23
meal	0.88	0.94	0.91	16
restriction_code	1.00	1.00	1.00	4
period_of_day	0.50	0.50	0.50	2
airport_code	0.50	0.22	0.31	9
compartment	0.00	0.00	0.00	1
transport_type	0.82	0.90	0.86	10
or	1.00	0.71	0.83	7
flight_days	0.90	0.90	0.90	10
booking_class	0.00	0.00	0.00	1
arrive_date.day_number	0.62	0.83	0.71	6
arrive_time.end_time	0.89	1.00	0.94	8
arrive_time.period_of_day	0.67	0.86	0.75	7
toloc.airport_name	0.67	0.67	0.67	3
arrive_time.start_time	0.89	1.00	0.94	8
depart_time.start_time	1.00	0.67	0.80	3
day_name	1.00	0.50	0.67	2
economy	0.86	1.00	0.92	6
toloc.airport_code	1.00	0.50	0.67	4
return_date.day_name	1.00	0.50	0.67	2
depart_date.year	1.00	1.00	1.00	3
fare_amount	1.00	1.00	1.00	2
return_date.date_relative	0.00	0.00	0.00	3
depart_time.end_time	0.50	0.33	0.40	3
state_code	1.00	1.00	1.00	1
toloc.country_name	0.00	0.00	0.00	1
days_code	0.33	1.00	0.50	1
stoploc.airport_code	0.00	0.00	0.00	1
flight_time	1.00	1.00	1.00	1
meal_code	1.00	1.00	1.00	1
micro avg	0.94	0.94	0.94	2842
macro avg	0.93	0.94	0.93	2842

Classification report of NER with OneNet (PhoATIS dataset)

JointCoBERTa + PhoBert LM

Parameters

```
intent_loss_coef=0.4,
tag_loss_coef=0.6,
use_intent_context_concat=False,
use_intent_context_attention=False,
attention_embedding_dim=200,
max_seq_len=50,
use_attention_mask=False,
train_batch_size=128,
eval_batch_size=256,
learning_rate=3e-5,
n_epochs=50,
```

CometV3 Dataset

-	intent_acc	0.9700
-	intent_f1	0.9703
-	intent_precision	0.9713
-	intent_recall	0.9700
-	tag_f1	0.9273
-	tag_precision	0.9202
-	tag_recall	0.9345

Metrics

	precision	recall	f1-score	support
agree	0.9565	1.0000	0.9778	22
ask_confirm	0.9842	0.9599	0.9719	324
ask_is_bot	1.0000	1.0000	1.0000	23
deny	0.9944	0.9944	0.9944	179
greet	1.0000	1.0000	1.0000	23
inform	0.9731	0.9602	0.9666	226
request#age_of_use	0.9565	1.0000	0.9778	22
request#brand	0.9630	0.9630	0.9630	27
request#color	0.9474	0.9474	0.9474	19
request#guarantee	0.9545	1.0000	0.9767	21
request#image	0.9130	1.0000	0.9545	21
request#link	1.0000	1.0000	1.0000	13
request#material	0.9545	1.0000	0.9767	21
request#origin	0.9048	0.9048	0.9048	21
request#price	0.9259	0.8929	0.9091	28
request#promotion	0.7619	1.0000	0.8649	16
request#size	0.9545	0.9545	0.9545	22
request#weight	0.9444	0.9444	0.9444	18
thanks	0.9565	1.0000	0.9778	22
accuracy			0.9700	1068
macro avg	0.9498	0.9748	0.9612	1068
weighted avg	0.9713	0.9700	0.9703	1068

Classification report of Intent classification with JointCoBERTa + PhoBert LM (KCloset dataset)

	precision	recall	f1-score	support
ask_confirm#age_of_use	0.8788	0.9355	0.9062	31
ask_confirm#brand	1.0000	0.8710	0.9310	31
ask_confirm#color	0.8077	0.9545	0.8750	22
ask_confirm#guarantee	0.8529	0.9062	0.8788	32
ask_confirm#material	0.8889	0.9231	0.9057	26
ask_confirm#object_type	0.8750	0.8750	0.8750	32
ask_confirm#origin	0.9375	1.0000	0.9677	30
ask_confirm#price	0.9000	0.9643	0.9310	28
ask_confirm#promotion	0.8824	0.8824	0.8824	17
ask_confirm#size	0.8667	0.8966	0.8814	29
ask_confirm#weight	0.9722	1.0000	0.9859	35
deny#brand	1.0000	1.0000	1.0000	24
deny#color	0.8750	0.8750	0.8750	24
deny#material	0.9655	1.0000	0.9825	28
deny#object_type	0.7241	0.7000	0.7119	30
deny#origin	1.0000	1.0000	1.0000	20
deny#price	1.0000	0.9688	0.9841	32
deny#size	0.9565	1.0000	0.9778	22
inform#age_of_use	0.8966	0.8667	0.8814	30
inform#brand	0.9419	0.9643	0.9529	84
inform#color	0.8824	0.8824	0.8824	34
inform#guarantee	0.9643	1.0000	0.9818	27
inform#material	0.9091	0.9677	0.9375	31
inform#object_type	0.9185	0.9370	0.9277	349
inform#origin	1.0000	0.9677	0.9836	31
inform#price	0.9333	0.9032	0.9180	31
inform#size	0.9706	0.9429	0.9565	35
inform#weight	1.0000	0.9333	0.9655	15
micro avg	0.9202	0.9345	0.9273	1160
macro avg	0.9214	0.9328	0.9264	1160
weighted avg	0.9210	0.9345	0.9273	1160

Classification report of NER with JointCoBERTa + PhoBert LM (KCloset dataset)

KCloset Dataset

- Evaluation results	
- loss	0.6909
- mean_acc_score	0.0000
- intent_acc	0.9398
- intent_f1	0.9412
- intent_precision	0.9446
- intent_recall	0.9398
- tag_f1	0.9174
- tag_precision	0.9038
- tag_recall	0.9313

Metrics

	precision	recall	f1-score	support
agree	0.9767	0.9333	0.9545	45
ask_about_ship	1.0000	0.9322	0.9649	59
ask_availability	0.9935	0.9503	0.9714	161
ask_availability#size	0.9659	0.9884	0.9770	86
ask_if_fit	1.0000	0.9500	0.9744	120
ask_if_sells	1.0000	0.9432	0.9708	176
ask_if_sells#online	0.9403	0.9692	0.9545	65
ask_if_sells#size	0.9904	0.9810	0.9856	105
ask_is_bot	0.9600	0.9231	0.9412	26
greet	0.9787	0.9787	0.9787	47
handoff	0.6981	0.8409	0.7629	132
inform	0.9349	0.9133	0.9240	173
order_request	0.9271	0.9889	0.9570	90
request#color	0.8000	0.9412	0.8649	17
request#material	0.9800	1.0000	0.9899	49
request#price	0.9691	0.9792	0.9741	96
request#product	0.9028	0.8333	0.8667	78
request#promotion	0.9610	0.9250	0.9427	80
request_more#color	0.9808	0.9273	0.9533	55
request_shop_address	0.9189	0.9444	0.9315	72
thank	0.9149	0.9773	0.9451	44
accuracy			0.9398	1776
macro avg	0.9425	0.9438	0.9421	1776
weighted avg	0.9446	0.9398	0.9412	1776

Classification report of Intent classification with JointCoBERTa + PhoBert LM (KCloset dataset)

	precision	recall	f1-score	support
ask_availability#color	0.0000	0.0000	0.0000	4
ask_if_sells#color	0.5000	0.7143	0.5882	7
inform#age	0.9652	0.9624	0.9638	346
inform#color	0.8689	0.8983	0.8833	59
inform#gender	0.9415	0.9620	0.9516	184
inform#height	0.8699	0.9769	0.9203	130
inform#location	0.6947	0.8053	0.7459	113
inform#material	0.7442	0.9143	0.8205	35
inform#mention	0.9345	0.9277	0.9311	415
inform#object_type	0.8704	0.9136	0.8915	625
inform#season	0.7000	0.5833	0.6364	12
inform#weight	0.9683	0.9683	0.9683	442
micro avg	0.9038	0.9313	0.9174	2372
macro avg	0.7548	0.8022	0.7751	2372
weighted avg	0.9055	0.9313	0.9177	2372

Classification report of NER with JointCoBERTa + PhoBert LM (KCloset dataset)

PhoATIS Dataset

```
-    -------- Evaluation results --------
-                 loss      1.1774
-       mean_acc_score      0.0000
-           intent_acc      0.9698
-            intent_f1      0.9748
-     intent_precision      0.9822
-        intent_recall      0.9698
-               tag_f1      0.9432
-        tag_precision      0.9404
-           tag_recall      0.9460
```

Metrics

	precision	recall	f1-score	support
UNK	0.0000	0.0000	0.0000	0
abbreviation	1.0000	0.9811	0.9905	53
aircraft	1.0000	0.9000	0.9474	10
airfare	1.0000	0.8654	0.9278	52
airfare#flight	0.5294	0.7500	0.6207	12
airline	1.0000	0.9444	0.9714	18
airline#flight_no	0.0000	0.0000	0.0000	0
airport	1.0000	1.0000	1.0000	18
capacity	1.0000	1.0000	1.0000	21
city	0.5000	0.7500	0.6000	4
distance	1.0000	1.0000	1.0000	10
flight	0.9904	0.9810	0.9857	632
flight_no	1.0000	0.8889	0.9412	9
flight_time	1.0000	1.0000	1.0000	1
ground_fare	0.8571	1.0000	0.9231	6
ground_service	1.0000	1.0000	1.0000	36
meal	0.3333	1.0000	0.5000	2
quantity	1.0000	1.0000	1.0000	7
restriction	1.0000	1.0000	1.0000	2
accuracy			0.9698	893
macro avg	0.8005	0.8453	0.8109	893
weighted avg	0.9822	0.9698	0.9748	893

Classification report of Intent classification with OneNet (PhoATIS dataset)

	precision	recall	f1-score	support
NK	0.0000	0.0000	0.0000	15
aircraft_code	0.7442	0.9697	0.8421	33
airline_code	0.8293	0.9714	0.8947	35
airline_name	0.9400	0.9307	0.9353	101
airport_code	0.5000	0.2222	0.3077	9
airport_name	1.0000	0.5000	0.6667	22
arrive_date.date_relative	1.0000	0.5000	0.6667	2
arrive_date.day_name	0.6875	1.0000	0.8148	11
arrive_date.day_number	0.7143	0.8333	0.7692	6
arrive_date.month_name	0.8333	0.8333	0.8333	6
arrive_time.end_time	0.7143	0.6250	0.6667	8
arrive_time.period_of_day	0.7000	1.0000	0.8235	7
arrive_time.start_time	1.0000	0.7500	0.8571	8
arrive_time.time	0.9429	0.9429	0.9429	35
arrive_time.time_relative	0.8710	0.9000	0.8852	30
city_name	0.8108	0.5556	0.6593	54
class_type	1.0000	1.0000	1.0000	24
connect	1.0000	0.8333	0.9091	6
cost_relative	1.0000	0.9730	0.9863	37
day_name	0.0000	0.0000	0.0000	2
days_code	0.0000	0.0000	0.0000	1
depart_date.date_relative	0.8947	1.0000	0.9444	17
depart_date.day_name	0.9858	0.9858	0.9858	212
depart_date.day_number	0.9815	0.9636	0.9725	55
depart_date.month_name	0.9821	0.9821	0.9821	56
depart_date.today_relative	0.8750	0.7778	0.8235	9
depart_date.year	1.0000	1.0000	1.0000	3
depart_time.end_time	0.5000	1.0000	0.6667	3
depart_time.period_mod	0.9091	1.0000	0.9524	10
depart_time.period_of_day	0.9832	0.9832	0.9832	119
depart_time.start_time	0.6000	1.0000	0.7500	3
depart_time.time	0.9403	0.9844	0.9618	64
depart_time.time_relative	0.9841	0.9538	0.9688	65
economy	0.7500	1.0000	0.8571	6
fare_amount	0.3333	1.0000	0.5000	2
fare_basis_code	0.8500	1.0000	0.9189	17
flight_days	1.0000	1.0000	1.0000	10
flight_mod	0.6207	0.7500	0.6792	24
flight_number	0.7857	1.0000	0.8800	11
flight_stop	0.9545	1.0000	0.9767	21
flight_time	1.0000	1.0000	1.0000	1
fromloc.airport_code	0.0000	0.0000	0.0000	5
fromloc.airport_name	0.4000	0.6667	0.5000	12
fromloc.city_name	0.9723	0.9957	0.9839	705
fromloc.state_code	1.0000	1.0000	1.0000	23
fromloc.state_name	0.7778	0.8235	0.8000	17
meal	0.9412	1.0000	0.9697	16
meal_code	0.0000	0.0000	0.0000	1
meal_description	1.0000	1.0000	1.0000	10
mod	0.0000	0.0000	0.0000	2
or	1.0000	1.0000	1.0000	7
period_of_day	0.0000	0.0000	0.0000	2
restriction_code	1.0000	1.0000	1.0000	4
return_date.date_relative	0.5000	0.3333	0.4000	3
return_date.day_name	0.0000	0.0000	0.0000	2
round_trip	1.0000	0.9863	0.9931	73
state_code	1.0000	1.0000	1.0000	1
state_name	0.0000	0.0000	0.0000	10
stoploc.city_name	0.8696	1.0000	0.9302	20
toloc.airport_code	0.6667	0.5000	0.5714	4
toloc.airport_name	0.6667	0.6667	0.6667	3
toloc.city_name	0.9647	0.9902	0.9773	717
toloc.country_name	0.0000	0.0000	0.0000	1
toloc.state_code	1.0000	1.0000	1.0000	18
toloc.state_name	0.8571	0.8571	0.8571	28
transport_type	1.0000	1.0000	1.0000	10
micro avg	0.9404	0.9460	0.9432	2854
macro avg	0.7248	0.7506	0.7260	2854
weighted avg	0.9313	0.9460	0.9361	2854

Classification report of NER with OneNet (PhoATIS dataset)

JointCoBERTa + CoBERTa-tiny-850

Parameters

```
intent_loss_coef=1.0,
tag_loss_coef=1.0,
use_intent_context_concat=False,
use_intent_context_attention=False,
attention_embedding_dim=200,
max_seq_len=50,
use_attention_mask=False,
train_batch_size=128,
eval_batch_size=256,
learning_rate=3e-5,
n_epochs=250,
```

KCloset Dataset

```
-     ----------- Evaluation results -----------
-                 loss      1.2492
-       mean_acc_score      0.0000
-           intent_acc      0.8491
-            intent_f1      0.8586
-     intent_precision      0.8764
-        intent_recall      0.8491
-               tag_f1      0.8320
-        tag_precision      0.8151
-           tag_recall      0.8495
```

Metrics

	precision	recall	f1-score	support
agree	0.7209	0.9394	0.8158	33
ask_about_ship	0.9455	0.9455	0.9455	55
ask_availability	0.9091	0.8284	0.8669	169
ask_availability#size	0.8750	0.9167	0.8953	84
ask_if_fit	0.9035	0.8803	0.8918	117
ask_if_sells	0.9759	0.8182	0.8901	198
ask_if_sells#online	0.9701	0.9155	0.9420	71
ask_if_sells#size	0.9615	0.9091	0.9346	110
ask_is_bot	0.3600	0.9000	0.5143	10
greet	0.9787	0.8364	0.9020	55
handoff	0.5597	0.7236	0.6312	123
inform	0.8166	0.9324	0.8707	148
order_request	0.9375	0.8333	0.8824	108
request#color	0.0000	0.0000	0.0000	0
request#material	0.9800	0.9245	0.9515	53
request#price	0.9485	0.9485	0.9485	97
request#product	0.6389	0.7302	0.6815	63
request#promotion	0.8831	0.8193	0.8500	83
request_more#color	0.8077	0.7778	0.7925	54
request_shop_address	0.9324	0.8023	0.8625	86
thank	0.8511	0.6780	0.7547	59
accuracy			0.8491	1776
macro avg	0.8074	0.8123	0.8011	1776
weighted avg	0.8764	0.8491	0.8586	1776

Classification report of Intent classification with JointCoBERTa + CoBERTa-tiny-900 (KCloset dataset)

	precision	recall	f1-score	support
ask_availability#color	0.0000	0.0000	0.0000	4
ask_if_sells#color	0.0000	0.0000	0.0000	7
inform#age	0.8952	0.9133	0.9041	346
inform#color	0.6753	0.8814	0.7647	59
inform#gender	0.9206	0.9457	0.9330	184
inform#height	0.8582	0.8846	0.8712	130
inform#location	0.4552	0.5841	0.5116	113
inform#material	0.7083	0.4857	0.5763	35
inform#mention	0.8876	0.9133	0.9002	415
inform#object_type	0.8143	0.8912	0.8510	625
inform#season	0.0000	0.0000	0.0000	12
inform#weight	0.7722	0.7670	0.7696	442
micro avg	0.8151	0.8495	0.8320	2372
macro avg	0.5822	0.6055	0.5901	2372
weighted avg	0.8117	0.8495	0.8291	2372

Classification report of NER with JointCoBERTa + CoBERTa-tiny-900 (KCloset dataset)

JointCoBERTa + PhoBERT LM + Intent context embedding (soft)

Parameters

```
intent_loss_coef=0.4,
tag_loss_coef=0.6,
use_intent_context_concat=False,
use_intent_context_attention=True,
attention_embedding_dim=200,
max_seq_len=50,
intent_embedding_type='soft',
use_attention_mask=False,
train_batch_size=128,
eval_batch_size=256,
learning_rate=3e-5,
n_epochs=50 ,
```

CometV3 Dataset

Evaluation results	
loss	0.6169
mean_acc_score	0.0000
intent_acc	0.9775
intent_f1	0.9777
intent_precision	0.9782
intent_recall	0.9775
tag_f1	0.9369
tag_precision	0.9325
tag_recall	0.9414

Metrics

	precision	recall	f1-score	support
agree	1.0000	1.0000	1.0000	23
ask_confirm	0.9873	0.9720	0.9796	321
ask_is_bot	1.0000	0.9583	0.9787	24
deny	0.9944	0.9944	0.9944	179
greet	1.0000	1.0000	1.0000	23
inform	0.9641	0.9729	0.9685	221
request#age_of_use	0.9565	1.0000	0.9778	22
request#brand	0.9630	0.9630	0.9630	27
request#color	0.9474	0.9474	0.9474	19
request#guarantee	1.0000	1.0000	1.0000	22
request#image	0.9565	1.0000	0.9778	22
request#link	1.0000	1.0000	1.0000	13
request#material	1.0000	0.9565	0.9778	23
request#origin	0.8095	0.9444	0.8718	18
request#price	1.0000	0.9643	0.9818	28
request#promotion	1.0000	0.9545	0.9767	22
request#size	0.9545	0.9545	0.9545	22
request#weight	0.9444	1.0000	0.9714	17
thanks	0.9565	1.0000	0.9778	22
accuracy			0.9775	1068
macro avg	0.9702	0.9780	0.9736	1068
weighted avg	0.9782	0.9775	0.9777	1068

Classification report of Intent classification with JointCoBERTa + PhoBERT LM + Intent context embedding (soft) (CometV3 dataset)

	precision	recall	f1-score	support
ask_confirm#age_of_use	0.9062	0.9355	0.9206	31
ask_confirm#brand	0.9643	0.8710	0.9153	31
ask_confirm#color	0.9130	0.9545	0.9333	22
ask_confirm#guarantee	0.8788	0.9062	0.8923	32
ask_confirm#material	0.9200	0.8846	0.9020	26
ask_confirm#object_type	0.9375	0.9375	0.9375	32
ask_confirm#origin	0.9677	1.0000	0.9836	30
ask_confirm#price	0.9000	0.9643	0.9310	28
ask_confirm#promotion	0.8824	0.8824	0.8824	17
ask_confirm#size	0.9259	0.8621	0.8929	29
ask_confirm#weight	0.9722	1.0000	0.9859	35
deny#brand	1.0000	1.0000	1.0000	24
deny#color	0.9167	0.9167	0.9167	24
deny#material	1.0000	1.0000	1.0000	28
deny#object_type	0.7419	0.7667	0.7541	30
deny#origin	1.0000	1.0000	1.0000	20
deny#price	1.0000	0.9688	0.9841	32
deny#size	0.9565	1.0000	0.9778	22
inform#age_of_use	0.8966	0.8667	0.8814	30
inform#brand	0.9647	0.9762	0.9704	84
inform#color	0.9143	0.9412	0.9275	34
inform#guarantee	0.9643	1.0000	0.9818	27
inform#material	0.9375	0.9677	0.9524	31
inform#object_type	0.9209	0.9341	0.9275	349
inform#origin	1.0000	1.0000	1.0000	31
inform#price	0.9333	0.9032	0.9180	31
inform#size	0.9211	1.0000	0.9589	35
inform#weight	1.0000	0.9333	0.9655	15
micro avg	0.9325	0.9414	0.9369	1160
macro avg	0.9370	0.9419	0.9390	1160
weighted avg	0.9329	0.9414	0.9369	1160

Classification report of NER with JointCoBERTa + PhoBERT LM + Intent context embedding (soft) (CometV3 dataset)

KCloset Dataset

Evaluation results	
loss	0.9711
mean_acc_score	0.0000
intent_acc	0.9352
intent_f1	0.9369
intent_precision	0.9416
intent_recall	0.9352
tag_f1	0.9231
tag_precision	0.9106
tag_recall	0.9359

Metrics

	precision	recall	f1-score	support
agree	0.9535	0.9762	0.9647	42
ask_about_ship	0.9818	0.9153	0.9474	59
ask_availability	0.9740	0.9554	0.9646	157
ask_availability#size	0.9659	0.9551	0.9605	89
ask_if_fit	0.9737	0.9569	0.9652	116
ask_if_sells	0.9819	0.9532	0.9674	171
ask_if_sells#online	0.9403	0.9545	0.9474	66
ask_if_sells#size	0.9904	0.9810	0.9856	105
ask_is_bot	1.0000	0.9259	0.9615	27
greet	0.9787	0.9388	0.9583	49
handoff	0.6730	0.8917	0.7670	120
inform	0.9231	0.9176	0.9204	170
order_request	0.9375	0.9474	0.9424	95
request#color	0.9500	0.8636	0.9048	22
request#material	0.9800	1.0000	0.9899	49
request#price	0.9588	0.9789	0.9688	95
request#product	0.8889	0.8000	0.8421	80
request#promotion	0.9870	0.8837	0.9325	86
request_more#color	0.9615	0.9615	0.9615	52
request_shop_address	0.9595	0.8765	0.9161	81
thank	0.9574	1.0000	0.9783	45
accuracy			0.9352	1776
macro avg	0.9484	0.9349	0.9403	1776
weighted avg	0.9416	0.9352	0.9369	1776

Classification report of Intent classification with JointCoBERTa + PhoBERT LM + Intent context embedding (soft) (KCloset dataset)

	precision	recall	f1-score	support
ask_availability#color	1.0000	1.0000	1.0000	4
ask_if_sells#color	0.8571	0.8571	0.8571	7
inform#age	0.9594	0.9566	0.9580	346
inform#color	0.9016	0.9322	0.9167	59
inform#gender	0.9415	0.9620	0.9516	184
inform#height	0.8671	0.9538	0.9084	130
inform#location	0.7190	0.7699	0.7436	113
inform#material	0.7895	0.8571	0.8219	35
inform#mention	0.9281	0.9325	0.9303	415
inform#object_type	0.8788	0.9280	0.9027	625
inform#season	0.6875	0.9167	0.7857	12
inform#weight	0.9772	0.9683	0.9727	442
micro avg	0.9106	0.9359	0.9231	2372
macro avg	0.8756	0.9195	0.8957	2372
weighted avg	0.9125	0.9359	0.9238	2372

Classification report of NER with JointCoBERTa + PhoBERT LM + Intent context embedding (soft) (KCloset dataset)

PhoATIS Dataset

Classification report of Intent classification with JointCoBERTa + PhoBERT LM + Intent context embedding (soft) (PhoATIS dataset)

Classification report of NER with JointCoBERTa + PhoBERT LM + Intent context embedding (soft) (PhoATIS dataset

JointCoBERTa + CoBERTa LM + Intent context embedding (hard)

Parameters

```
intent_loss_coef=0.4,
tag_loss_coef=0.6,
use_intent_context_concat=True,
```

```
use_intent_context_attention=False,
attention_embedding_dim=200,
max_seq_len=50,
intent_embedding_type='hard',
use_attention_mask=False,
train_batch_size=128,
eval_batch_size=256,
learning_rate=3e-5,
n_epochs=50,
```

CometV3 Dataset

```
-     ---------- Evaluation results ----------
-                     loss      0.6056
-           mean_acc_score      0.0000
-               intent_acc      0.9785
-                intent_f1      0.9786
-         intent_precision      0.9791
-            intent_recall      0.9785
-                   tag_f1      0.9345
-            tag_precision      0.9345
-               tag_recall      0.9345
```

Metrics

	precision	recall	f1-score	support
agree	1.0000	1.0000	1.0000	23
ask_confirm	0.9842	0.9749	0.9795	319
ask_is_bot	1.0000	0.9583	0.9787	24
deny	0.9944	1.0000	0.9972	178
greet	1.0000	1.0000	1.0000	23
inform	0.9776	0.9604	0.9689	227
request#age_of_use	0.9565	1.0000	0.9778	22
request#brand	0.9630	0.9630	0.9630	27
request#color	0.9474	1.0000	0.9730	18
request#guarantee	1.0000	1.0000	1.0000	22
request#image	0.9565	1.0000	0.9778	22
request#link	1.0000	1.0000	1.0000	13
request#material	1.0000	1.0000	1.0000	22
request#origin	0.8571	0.9474	0.9000	19
request#price	1.0000	0.9643	0.9818	28
request#promotion	0.8571	0.9474	0.9000	19
request#size	0.9091	0.9524	0.9302	21
request#weight	1.0000	1.0000	1.0000	18
thanks	1.0000	1.0000	1.0000	23
accuracy			0.9785	1068
macro avg	0.9686	0.9825	0.9752	1068
weighted avg	0.9791	0.9785	0.9786	1068

Classification report of Intent classification with JointCoBERTa + CoBERTa LM + Intent context embedding (hard) (CometV3 dataset)

	precision	recall	f1-score	support
ask_confirm#age_of_use	0.9062	0.9355	0.9206	31
ask_confirm#brand	1.0000	0.9677	0.9836	31
ask_confirm#color	0.9545	0.9545	0.9545	22
ask_confirm#guarantee	0.8235	0.8750	0.8485	32
ask_confirm#material	0.9231	0.9231	0.9231	26
ask_confirm#object_type	0.9333	0.8750	0.9032	32
ask_confirm#origin	1.0000	0.9667	0.9831	30
ask_confirm#price	0.9643	0.9643	0.9643	28
ask_confirm#promotion	0.8824	0.8824	0.8824	17
ask_confirm#size	0.9259	0.8621	0.8929	29
ask_confirm#weight	1.0000	1.0000	1.0000	35
deny#brand	1.0000	1.0000	1.0000	24
deny#color	0.9167	0.9167	0.9167	24
deny#material	0.9655	1.0000	0.9825	28
deny#object_type	0.6875	0.7333	0.7097	30
deny#origin	1.0000	1.0000	1.0000	20
deny#price	1.0000	0.9688	0.9841	32
deny#size	0.9565	1.0000	0.9778	22
inform#age_of_use	0.8966	0.8667	0.8814	30
inform#brand	0.9880	0.9762	0.9820	84
inform#color	0.8889	0.9412	0.9143	34
inform#guarantee	0.9643	1.0000	0.9818	27
inform#material	0.9375	0.9677	0.9524	31
inform#object_type	0.9304	0.9198	0.9251	349
inform#origin	1.0000	0.9677	0.9836	31
inform#price	0.8710	0.8710	0.8710	31
inform#size	0.9211	1.0000	0.9589	35
inform#weight	0.9333	0.9333	0.9333	15
micro avg	0.9345	0.9345	0.9345	1160
macro avg	0.9347	0.9382	0.9361	1160
weighted avg	0.9354	0.9345	0.9347	1160

Classification report of NER with JointCoBERTa + CoBERTa LM + Intent context embedding (hard) (CometV3 dataset)

KCloset Dataset

	Evaluation results	
-	loss	0.9602
-	mean_acc_score	0.0000
-	intent_acc	0.9414
-	intent_f1	0.9425
-	intent_precision	0.9464
-	intent_recall	0.9414
-	tag_f1	0.9154
-	tag_precision	0.9011
-	tag_recall	0.9300

Metrics

	precision	recall	f1-score	support
agree	0.9767	0.9130	0.9438	46
ask_about_ship	1.0000	0.9322	0.9649	59
ask_availability	0.9740	0.9804	0.9772	153
ask_availability#size	0.9773	0.9773	0.9773	88
ask_if_fit	0.9912	0.9576	0.9741	118
ask_if_sells	0.9940	0.9429	0.9677	175
ask_if_sells#online	0.9701	0.9559	0.9630	68
ask_if_sells#size	0.9904	0.9717	0.9810	106
ask_is_bot	1.0000	1.0000	1.0000	25
greet	0.9787	0.9200	0.9485	50
handoff	0.7170	0.8769	0.7889	130
inform	0.8935	0.9679	0.9292	156
order_request	0.9375	0.9890	0.9626	91
request#color	1.0000	0.7692	0.8696	26
request#material	0.9800	0.9800	0.9800	50
request#price	0.9794	0.9596	0.9694	99
request#product	0.9028	0.8025	0.8497	81
request#promotion	0.9740	0.9036	0.9375	83
request_more#color	0.9038	0.9792	0.9400	48
request_shop_address	0.9595	0.9103	0.9342	78
thank	0.9574	0.9783	0.9677	46
accuracy			0.9414	1776
macro avg	0.9551	0.9365	0.9441	1776
weighted avg	0.9464	0.9414	0.9425	1776

Classification report of Intent classification with JointCoBERTa + CoBERTa LM + Intent context embedding (hard) (KCloset dataset)

	precision	recall	f1-score	support
ask_availability#color	0.0000	0.0000	0.0000	4
ask_if_sells#color	0.5000	0.8571	0.6316	7
inform#age	0.9511	0.9566	0.9539	346
inform#color	0.9310	0.9153	0.9231	59
inform#gender	0.9149	0.9348	0.9247	184
inform#height	0.8803	0.9615	0.9191	130
inform#location	0.6984	0.7788	0.7364	113
inform#material	0.7949	0.8857	0.8378	35
inform#mention	0.9236	0.9325	0.9281	415
inform#object_type	0.8784	0.9248	0.9010	625
inform#season	0.3333	0.4167	0.3704	12
inform#weight	0.9684	0.9706	0.9695	442
micro avg	0.9011	0.9300	0.9154	2372
macro avg	0.7312	0.7945	0.7580	2372
weighted avg	0.9028	0.9300	0.9158	2372

Classification report of NER with JointCoBERTa + CoBERTa LM + Intent context embedding (hard) (KCloset dataset)

PhoATIS Dataset

Classification report of Intent classification with JointCoBERTa + CoBERTa LM + Intent context embedding (hard) (PhoATIS dataset)

Classification report of NER with JointCoBERTa + CoBERTa LM + Intent context embedding (hard) (PhoATIS dataset)