Article

Intent Classification and Slot Filling Model for In-Vehicle Services in Korean

Jungwoo Lim [1,†] , Suhyune Son [1,†], Songeun Lee [2], Changwoo Chun [2], Sungsoo Park [2], Yuna Hur [1,*] and Heuiseok Lim [1,*]

1 Department of Computer Science and Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 06289, Republic of Korea
2 Automotive Research & Develpment Division, Hyundai Motor Group, Seoul 06289, Republic of Korea
* Correspondence: yj72722@korea.ac.kr (Y.H.); limhseok@korea.ac.kr (H.L.)
† These authors contributed equally to this work.

Abstract: Since understanding a user's request has become a critical task for the artificial intelligence speakers, capturing intents and finding correct slots along with corresponding slot value is significant. Despite various studies concentrating on a real-life situation, dialogue system that is adaptive to in-vehicle services are limited. Moreover, the Korean dialogue system specialized in an vehicle domain rarely exists. We propose a dialogue system that captures proper intent and activated slots for Korean in-vehicle services in a multi-tasking manner. We implement our model with a pre-trained language model, and it includes an intent classifier, slot classifier, slot value predictor, and value-refiner. We conduct the experiments on the Korean in-vehicle services dataset and show 90.74% of joint goal accuracy. Also, we analyze the efficacy of each component of our model and inspect the prediction results with qualitative analysis.

Keywords: intent classification; slot filling; in-vehicle services domain; Korean dialogue system

1. Introduction

Along with the increasing significance of artificial intelligence (AI) speakers, their understanding capability of a user's request becomes a critical task. In order to build a smart AI speaker such as Google Home or Amazon Echo, natural language understanding (NLU) is an important component to accomplish the goal through human interaction. In specific, NLU in a task-oriented spoken dialogue system typically is divided into intent classification and slot filling [1]. Intent classification is a task to distinguish the intention of the dialogue, slot filling finds the correct value for the corresponding slots [2].

Unlike previous studies that focus on developing independent models for each task [3–5], recent studies on task-oriented dialogue systems introduce the models that jointly learn diverse tasks simultaneously [1,6–9]. The models built with joint learning on intent classification and slot filling in English show improved performance on the task-oriented dialogue datasets as well [8,9]. The datasets for the task-oriented dialogue system include diverse domains such as booking airlines, restaurants, and hotels or playing music [10–14].

Despite the substantial performance of joint models, the above models assume users at home or work so that they cannot cover the users who are in a specific environment like in-vehicle. Since in-vehicle services usually maintain distinct intents and slots compared to normal dialogue at home, it is critical to consolidate the robustness of the model with car-adaptive data. Moreover, they are built with the assumption that the user's request is limited to a small number of intents, such as "find", or "book" compared to our expectation from in-vehicle AI speakers. To avoid the limited coverage and the small number of intents of in-vehicle AI speaker services in English, the studies that consider the users who are under the condition of in-vehicle appear as well [15–17].

On the contrary, Korean language models still suffer from the narrow range of task-oriented dialogue data to build in-vehicle AI speaker service. Even though the datasets for task-oriented dialogue are available (https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71265, accessed on 3 November 2022) [18–20], the dataset that covers in-vehicle situation does not exist. The limited number of datasets leads to the absence of models that learn intent classification and slot filling in a multi-tasking manner.

In this paper, we introduce a model that captures the suitable intents and finds slot value for proper slots simultaneously with the Korean in-vehicle services dataset. Our model encodes the user's utterance with a pre-trained language model to find the correct intent of the request. Then, the model finds the proper slots and fills them with predicted slot values. In the end, the value-refiner post-processes the slot value with database-matching module and value-matching module for precise slot values. From the experiments, we find that our model show improved performances on in-vehicle services in Korean. We also conduct qualitative analysis for detailed investigation of regarding intents and slots. Moreover, we show an ablation study to show the effectiveness of the value-refiner with two database-matching module and value matching module.

- We propose a model that learns in-vehicle services situations with diverse domains in Korean that are jointly trained with intent classification and slot-filling.
- To show our model's effectiveness, we conduct experiments on a mobility domain dataset and show comparable performances on the dataset.
- We show the efficacy of the value-refiner through an ablation study and demonstrate the error types from the model prediction.

2. Related Work

2.1. Task-Oriented Dialogue System

In a task-oriented dialogue system, the ability to understand users' requests is important. Extracting the useful information from the request is processed with two tasks: Intent Classification and Slot Filling. Previous studies are generally divided into the following two approaches: Multi-task framework and using intent information to guide the slot-filling task. In a multi-task framework, they share helpful information between two tasks [21]. For this, one applies a deep bi-directional recurrent neural network (RNN) [22] with long short-term memory (LSTM) [23] and gated recurrent unit (GRU) to learn the representations shared by the tasks [24]. An attention-based neural network model is also used for joint intent detection and slot filling [25]. Another approach is that the model train two tasks jointly using a framework that incorporates intent information to guide slot filling [26,27]. Another study directly uses intent information as input for the slot filling task to make the model perform the slot filling based on intent semantic knowledge [28]. Because the two tasks are closely related and one task affects the other, they are trained jointly in previous approaches [28,29].

In the study of the Korean task-oriented dialogue system, studies on intent classification and slot-filling tasks are also conducted. For the study in [30], Korean intent classification data provided by AI Hub (https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=102, accessed on 3 November 2022) are used for intent classification model that matches the most similar intent using semantic textual similarity. Data in constructed on a calendar domain for slot-filling task research in [31], and the highest probability slot and value are predicted by utilizing the LSTM model and applying two methods, greedy-search decoding, and beam-search decoding. Another builds the Korean-based restaurant reservation and recommended conversation system [19] with the dataset provided by the research [32]. It also proposes a joint model that can predict intents and slot filling using KoBERT simultaneously. In Korean research, task-oriented dialogue system data from various domains are opened, but the data on in-vehicle services is still limited. In particular, joint models on intent classification

and slot filling for mobility services are insufficient compared to the studies on English regarding the numbers of studies.

2.2. Pre-Trained Language Models for Dialogue Systems

Recently, pre-trained language models (PLMs) in English have shown considerable performance in natural language processing (NLP) tasks. BERT [33], which is pre-trained with a large corpus, improves the performance in a wide range of tasks in NLP. Other PLMs such as RoBERTa [34], ELECTRA [35], and ALBERT [36] also show enhanced performance on downstream tasks. Due to the powerful performance of PLMs, Korean PLMs such as KoBERT (https://github.com/SKTBrain/KoBERT, accessed on 3 November 2022), KorBERT (https://aiopen.etri.re.kr/dataDownload, accessed on 3 November 2022), and KLUE-BERT [18] have been released. Among them, we conduct experiments using KoBERT, KLUE-RoBERTa [18], and mBERT [33]. KoBERT is the first Korean PLMs trained with 5M sentences and 54M words from Korean Wikipedia (https://ko.wikipedia.org/, accessed on 3 November 2022). KoBERT is trained with masked language modeling (MLM) and next sentence prediction (NSP) same as BERT. Sentencepiece [37] is used as a tokenizer with a vocabulary size of 8002 words. KLUE-RoBERTa is trained with 63GB of a corpus consisting of Modu corpus (https://corpus.korean.go.kr/, accessed on 3 November 2022), CC-100-Kor (http://data.statmt.org/cc-100/, accessed on 3 November 2022), Namu Wiki (https://namu.wiki/, accessed on 3 November 2022), and other web sources. KLUE-RoBERTa uses a dynamic masking strategy for pre-training. The tokenizer is a morpheme-based subword tokenizer with a vocabulary size of 32,000 words. mBERT is a multilingual pre-trained model, which is trained with the same strategies as BERT. mBERT is trained on 15 G of the Wikipedia dataset and uses a wordpiece tokenizer [38] with a vocabulary size of 119,547 words.

3. Method

In order to enable the model to capture the correct intent of the utterance and finds the proper slot value for the corresponding slot simultaneously, we implement the model with intent classifier, slot classifier, and slot value predictor as in Figure 1. Moreover, we additionally integrate a value-refiner with a value predictor for precise slot value. In detail, when the human's utterance comes in, the model encodes the request with a pre-trained language model. Then, the model classifies the intent and activated slots based on the descriptions representation of intents and slots. When the slot is predicted to be active, the slot value predictor finds the proper value to fill in from the utterance along with the value-refiner. A detailed explanation is illustrated in the below sections.
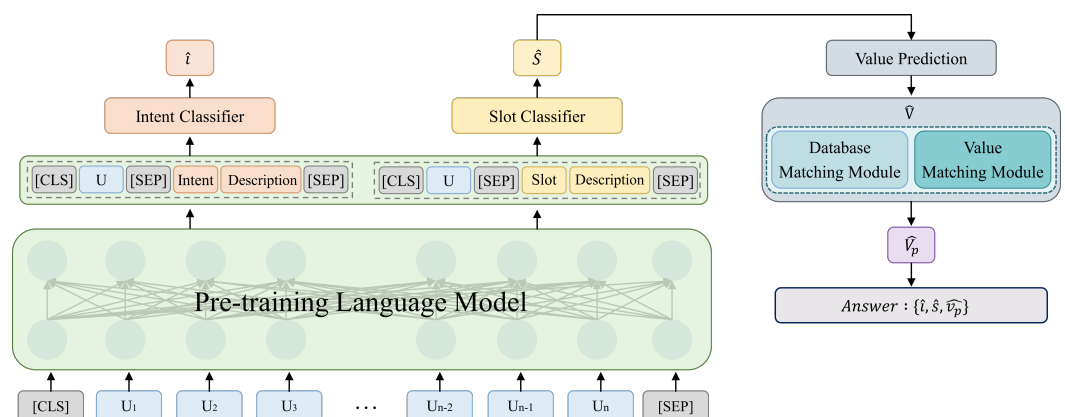

Figure 1. Overview of our model.

3.1. Intent Classifier

The intent classifier receives the user's utterance and encodes them with the pre-trained language model. For the enhanced understanding capability of the intention of the

request, we add the intent description for the model input. Therefore, the input is formalised as the concatenation of [CLS], utterance, the name of the intent, and its description. The pre-trained language model then predicts the proper intention of the user's request with [CLS] token. Since each utterance has only one intention, we use cross-entropy [39] for the intention loss $\mathcal{L}_I$ as in Equation (1) where the l_j is the label of the intention of jth utterance.

$$\mathcal{L}_I = -\sum_j l_j \cdot \log \hat{l}_j \tag{1}$$

3.2. Slot Classifier

Similarly, the activated slot is classified by slot classifier. The input of the slot classifier is also depicted as the concatenation of [CLS], utterance, the name of the slot, and its description for a comprehensive understanding of the request. The output embeddings of the [CLS] token from the final hidden layer are then used to predict whether the given slot should be activated or not. Since one or more slots can be assigned to a single utterance, we utilize binary cross-entropy for the slot loss $\mathcal{L}_S$ as in Equation (2) where the s_j is the ground truth of the activated slots of jth utterance.

$$\mathcal{L}_S = -\sum_j s_j \cdot \log \hat{s}_j + (1 - s_j) \cdot \log(1 - \hat{s}_j) \tag{2}$$

3.3. Slot Value Predictor

When the activated slot is decided by the slot classifier, then the slot value predictor predicts the slot value according to the type of slots. The slots can be divided into categorical slots and non-categorical slots. When the activated slot belongs to categorical slots, the model is trained to classify which value should be selected from the value pool, which is pre-defined. If the non-categorical slot is activated, then the slot value predictor points out the start and end position of the slot value from the utterance. In detail, the start vector and end vector are trained to predict the value for the activated non-categorical slot. The probability of the start token in the utterance is computed as a dot product between the final hidden state of the utterance and the start vector, followed by a softmax over all of the words in the utterance. The end token of the span is also calculated similarly. Then, the maximum scoring span from the list of candidate spans is predicted as slot value [33]. The sum of the log-likelihood [40] of start and end positions is used as a loss function.

3.4. Value Refiner

For accurate and precise slot-filling, we additionally refine the predicted value. It includes two modules, which are the database-matching module and the value-matching module. They are applied differently according to the type of activated slots. For the categorical slots, the database matching module first extracts the pre-defined values of the predicted slot from the slot value database if the predicted value does not exist in the utterance. Then, the value-matching module finds the longest value from the values by checking the possible value in the utterance to find the exact slot value. For non-categorical slots, the database matching module only brings the database values. Then, the value-matching module utilizes fuzzy-matching to predict proper values by calculating the Levenshtein distance between the predicted value and possible value from the database matching module.

4. Experiments

4.1. Data

The given in-vehicle domain dialogue data consists of 492,000 examples of the train split and 260,991 examples of the test split as in Table 1. The dataset has 25 slots and 267 intents. Each example contains a user's utterance and activated slots. Also, the corresponding slot value is labeled. The test split contains the slot value that does not appear in the training

set, and both the training and test split include all domains. The name and description of each domain are listed in Table 2. Each intent of the utterance can be categorized according to the domain, and the number of the intents that belongs to each domain is indicated in Table 2 as well. The statistics of each slot are denoted in Table 3.

Table 1. Dataset Statistics.

	# of Examples	# of Domains
Train	492,000	23
Test	260,991	23

Table 2. Statistics and description of each domain. The number of intents that belong to the domain is also listed.

Domain	Domain Description	# of Intent
AVNT	Waypoint	2
BT	Bluetooth	12
chitchat	Inconsequential conversation	8
cluster	Dashboard	6
embedded	Systems built into car	45
fatc	Car control	26
glass	Window and side mirror control	6
hipass	Hipass	6
ma	Mobile application	3
music	Music	3
navi	Navigation	2
others	Others	2
portal	Portal Search	17
QA	Question and answering	22
seat	Seat control	38
settings	Setting control	17
simple	Simple setting control	6
sunroof	Sunroof control	2
trunk	Trunk control	2
vehicle	Charger control	2
weather	Weather check	12
wheel	Wheel heating control	2
wind	Wind control	16

Table 3. Slot statistics and descriptions. The number of slots is also indicated that exist in the dataset.

Slot Name	Slot Description	# of Slots
Categorical Slots		
AboutDisplay	In-vehicle display devices, instrument clusters, heads-up displays, lights, lamp	6328
SettingBar	Settings change control button	14,475
Non-Categorical Slots		
AlbumName	Search music, song album title when playing music	453

Table 3. *Cont.*

Slot Name	Slot Description	# of Slots
AMSetting	Radio AM frequency range	2121
BroadcastStation	The name of the broadcasting station that transmits radio broadcasts and the radio program broadcast from the broadcasting station	3269
CallTarget	A call target with a dialing function through Bluetooth cell phone linkage	18,166
Consumables	Automobile interior parts that are worn out and need replacement or continuous inspection	42,097
Date	Date and time	33,264
FMSetting	Radio FM frequency range	2112
GenreName	Music search, music playback, music genre, classification	449
Region	Address unit, special city, province, city, county, street name, street name	22,882
SearchPlace	Search place, POI, school, restaurant, gas station, subway station, train station, place, address, frequent places	4745
SearchRange	Search area, nearby subway station, train station, airport, park, POI	1877
SettingCheck	Change Settings checkbox	3819
SettingColor	Setting color to change mood light color	3629
SettingTarget	Settings Classification Menu	9869
SettingValue	Setting change value	10,995
SingerName	Search for music, name of singer, artist when playing music	1233
SongName	Search music, song title when playing music	1418
SpecialPlace	Schools, educational institutions, companies, shops, shops, restaurants, gas stations, parking lots, buildings, lakes, complexes, addresses, entrances of apartments, etc.	1407
Switchgear	Associated devices capable of controlling opening and closing	9831
System	Safety device system, driving device, system, alarm, mode	13,022
TemperatureValue	Temperature setting values for air conditioning, air conditioning, and heater inside the vehicle	3576
Update	Software update	12,866
WarningLight	Lights up to warn users when the operating status of equipment in the car is abnormal, etc. This is shown in the car cluster.	3633

4.2. Experimental Setup

We utilize KoBERT-base, KLUE-RoBERTa-base, and mBERT-base, which have 12 layers and 12 attention heads with a 768 embedding size. For KoBERT and mBERT, we use a batch size of 32, and the learning rate is set as 5×10^{-7}. When we train our model with KLUE-RoBERTa, the learning rate is set to 3×10^{-5} with a batch size of 32. We use adam optimizer for all of our experiments, and we choose the hyperparameters with a manual search. One Quadro RTX-8000 is used for our experiments.

4.3. Evaluation Metrics

We use accuracy as the main metrics for intent classification, slot classification, and categorical slot value prediction. As the model needs to select the value from the value pools, accuracy is suitable for checking the percentage of correctly classified observations. For non-categorical value prediction, we use exact-match (EM) and F1 scores for the evaluation metrics. As EM and F1 are two widely used metrics for question and answering, we utilize the metrics due to their similarity in span prediction. Moreover, we use joint goal accuracy (JGA) to measure the number of examples that the model predicts correct intent, slots, and slot values.

4.4. Results and Analysis

4.4.1. Main Results

We conduct experiments of intent classification and slot filling on the in-vehicle services dataset. To show the efficacy of multi-tasking in our model, we also include experiments on the results from single training. The single training model only trains one task with the dataset. In Table 4, ICO denotes the model that trains with intent classification only, and SFO indicates the model that is learned with slot-filling only. According to the experimental result, it is found that multitask-learning improved intent accuracy, slot accuracy, and value prediction at the same time. Especially, the score of value prediction is improved while indicating that the learning intent and slot simultaneously affect each other task performance in a positive way.

Table 4. Main Experimental Results. ICO denotes the model that trains with intent classification only, and SFO indicates the model that is learned with slot-filling only. Acc. is the abbreviation of accuracy.

		Intent Acc.	Slot Acc.	Cat Acc.	Non-Cat EM	Non-Cat F1	JGA
ICO	KoBERT	98.50	-	-	-	-	-
	KLUE-RoBERTa	97.14	-	-	-	-	-
	mBERT	96.70	-	-	-	-	-
SFO	KoBERT	-	99.46	51.54	85.51	94.68	-
	KLUE-RoBERTa	-	98.47	51.57	54.00	86.88	-
	mBERT	-	98.45	51.57	78.94	91.27	-
ICO + SFO	KoBERT	98.90	**99.70**	93.71	86.67	95.53	86.55
	KLUE-RoBERTa	**98.98**	99.45	94.29	76.41	92.84	86.42
	mBERT	98.38	99.52	**94.51**	**89.97**	**95.83**	**90.74**

It is also revealed that the mBERT's value prediction is better than that of other Korean-based language models. We assume that the performance difference in value prediction is due to the uniqueness of the in-vehicle service dataset. As mBERT is trained with diverse, multilingual languages, it is more robust than other Korean-based language models on slot value prediction since a substantial number of the slots' values include foreign words, including English.

4.4.2. Ablation Study on Value Refiner

We include the ablation study results on value-refining in Table 5. It is found that our two refiner modules are effective regardless of the type of language model. Also, it is shown that our database matching module and value matching module are necessary for improved performance. The difference between the full model and base model with no refining is more than 4% at least.

Table 5. Ablation Study on value-refiner. DM indicates the Database-Matching module, and VM is Value-Matching module. The number in the table is the result of joint goal accuracy metric.

PP Type	KoBERT	KLUE-RoBERTa	mBERT
DM + VM	**86.55**	**86.42**	**90.74**
DM	85.91	84.82	90.54
VM	83.51	83.35	87.66
-	76.95	80.55	86.38

5. Discussion

5.1. Qualitative Results on Slot Value Prediction

We also conduct qualitative analysis on slot value prediction from our models. In Table 6, an example from the Slot-value predicted result is illustrated. As in the example, when the utterance is given as "Check if a software version update is necessar", the correct answer of the slot label is "Update" and the value label is the "software version". Regardless of the type of language models, the slot is correctly predicted with "Update". However, the prediction for slot value is "Software ver", and it can be seen that the entire word is not predicted in the case of KoBERT and RoBERTa while mBERT correctly predicts "Software version". From this result, it is found that the car domain shows satisfactory results in a model trained in multiple languages such as mBERT rather than a model trained in Korean because most terms are in English. We assume that this result implies that mBERT is more robust to the English-based slot value which are spread in in-vehicle services dataset.

5.2. Error Analysis

To deeply understand the errors in our models, we analyze the inference examples from the value prediction procedure. We categorize them into two types of errors. One is the situation where the model misses out on the address name units as in Table 7. The Korean address system is divided into three hierarchies: -daero, -ro, and -gil [41]. -daero is designated for a road with a width of 40 m or more than eight lanes. -ro is roads or streets smaller than daero. These roads are between eight and two lanes. -gill is typically the smallest road having only one lane. That is, the road name and the corresponding hierarchy appear together in the sentence but should be interpreted separately. Due to the characteristic of our mobility domain, examples related to these addresses occupy a larger proportion compared to other domain datasets. For this reason, we find that the model becomes more confused and shows errors in omitting the address hierarchy.

The other occurs when the model predicts the slot value without omitting the postposition from the utterance as in Table 8. Korean has a linguistic characteristic that is different from English. In Korean, when a root word and a postposition are combined, the word has a specific meaning in the sentence [42]. Postposition includes 은(eun), 는(neun), 이(i), 가 (ga), 를(leul), and 의(ui). That is, the root word has different meanings in the sentence depending on the postposition [43]. For example, if 은(eun) is used as a postposition, the root word has the meaning of the subject in the sentence, and when used with 를(leul), it has the meaning of the object. Therefore, it is important to distinguish the postposition from the root word [42] as the meaning of the slot value also depends on the postposition. Therefore, we find that this error occurs since it learns the meaning of the slot value that depends on the postposition in the sentence.

Table 6. Qualitative result for slot-value predicted.

Utterance	
소프트웨어 버전 업데이트 필요한지 봐줘 *Watch if a software version update is necessary*	
Slot Label	
업데이트 *Update*	
Slot Prediction	
KoBERT	업데이트 *Update*
RoBERTa	업데이트 *Update*
mBERT	업데이트 *Update*
Value Label	
소프트웨어 버전 *Software version*	
Value Prediction	
KoBERT	소프트웨어 버 *Software ver*
RoBERTa	소프트웨어 버 *Software ver*
mBERT	소프트웨어 버전 *Software version*

Table 7. Examples for missing address name units error.

Ground-Truth	Slot Value Prediction
도두리로 Doduri-ro	도두리 Doduri
우천산업단지로 Ucheonsaneopdanji-ro	우천산업단지 Ucheonsaneopdanji
녹산산단 153로 Noksansandan 153-ro	녹산산단 153 Noksansandan 153
학동 11로 Hakdong 11-ro	학동 11 Hakdong 11

Table 8. Examples for prediction including postposition. Bold denotes the postposition in each example.

Ground-Truth	Slot Value Prediction
우리 아파트 어린이집 선생님 uri apateu eorinijip seonsaengnim	우리 아파트 어린이집 선생님**에** uri apateu eorinijip seonsaengnim**E**
ΔΔ국제대 ΔΔgukjedae	ΔΔ국제대**가** ΔΔgukjedae **Ga**
1183 1183	1183**로** 1183**ro**
유O은 YooOEun	유O은**이** YooOEun**Yi**

6. Conclusions

In this study, we introduce intent classification and slot-filling model for in-vehicle services in Korean. To build the dialogue system, we utilize the pre-trained language model and train in a multi-tasking manner with value-refiner. In the experiments, we show that our model shows improved performance on intent classification and slot classification than other single training models. Moreover, we found that mBERT's value prediction is better than that of Korean-based language models due to mBERT's robust performance on the unique in-vehicle dataset. Also, we conduct the ablation study for the value-refiner to show its efficacy across the types of language models. From additional analyses, we find two error patterns from the model prediction, and this implies that future work is needed to improve the proposed dialogue system.

References

1. Zhang, Z.; Zhang, Z.; Chen, H.; Zhang, Z. A joint learning framework with bert for spoken language understanding. *IEEE Access* **2019**, *7*, 168849–168858. [CrossRef]
2. Louvan, S.; Magnini, B. Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 480–496.
3. Mesnil, G.; He, X.; Deng, L.; Bengio, Y. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; pp. 3771–3775.
4. Mesnil, G.; Dauphin, Y.; Yao, K.; Bengio, Y.; Deng, L.; Hakkani-Tur, D.; He, X.; Heck, L.; Tur, G.; Yu, D.; et al. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *23*, 530–539. [CrossRef]
5. Liu, B.; Lane, I. Recurrent neural network structured output prediction for spoken language understanding. In Proceedings of the NIPS Workshop on Machine Learning for Spoken Language Understanding and Interactions, Montreal, QC, Canada, 11 December 2015.
6. Zhang, C.; Li, Y.; Du, N.; Fan, W.; Philip, S.Y. Joint Slot Filling and Intent Detection via Capsule Neural Networks. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5259–5267.
7. Wang, Y.; Shen, Y.; Jin, H. A Bi-Model Based RNN Semantic Frame Parsing Model for Intent Detection and Slot Filling. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 309–314.
8. Lin, Z.; Madotto, A.; Winata, G.I.; Fung, P. MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 3391–3405.
9. Wu, C.S.; Hoi, S.C.; Socher, R.; Xiong, C. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 917–929.
10. Hemphill, C.T.; Godfrey, J.J.; Doddington, G.R. The ATIS Spoken Language Systems Pilot Corpus. In Proceedings of the Speech and Natural Language, St. Louis, PA, USA, 24–27 June 1990.
11. Coucke, A.; Saade, A.; Ball, A.; Bluche, T.; Caulier, A.; Leroy, D.; Doumouro, C.; Gisselbrecht, T.; Caltagirone, F.; Lavril, T.; et al. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. *arXiv* **2018**, arXiv:1805.10190.

12. Schuster, S.; Gupta, S.; Shah, R.; Lewis, M. Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 3795–3805.
13. Rastogi, A.; Zang, X.; Sunkara, S.; Gupta, R.; Khaitan, P. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8689–8696.
14. Budzianowski, P.; Wen, T.H.; Tseng, B.H.; Casanueva, I.; Ultes, S.; Ramadan, O.; Gasic, M. MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 5016–5026.
15. Eric, M.; Krishnan, L.; Charette, F.; Manning, C.D. Key-Value Retrieval Networks for Task-Oriented Dialogue. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrucken, Germany, 15–17 August 2017; pp. 37–49.
16. Abro, W.A.; Qi, G.; Ali, Z.; Feng, Y.; Aamir, M. Multi-turn intent determination and slot filling with neural networks and regular expressions. *Knowl.-Based Syst.* **2020**, *208*, 106428. [CrossRef]
17. Yanli, H. Research on Spoken Language Understanding Based on Deep Learning. *Sci. Program.* **2021**. [CrossRef]
18. Park, S.; Moon, J.; Kim, S.; Cho, W.I.; Han, J.; Park, J.; Song, C.; Kim, J.; Song, Y.; Oh, T.; et al. KLUE: Korean Language Understanding Evaluation. *arXiv* **2021**, arXiv:2105.09680.
19. Han, S.; Lim, H. Development of Korean dataset for joint intent classification and slot filling. *J. Korea Converg. Soc.* **2021**, *12*, 57–63.
20. Kim, Y.M.; Lee, T.H.; Na, S.O. Constructing novel datasets for intent detection and ner in a korean healthcare advice system: Guidelines and empirical results. *Appl. Intell.* **2022**, 1–21. [CrossRef]
21. Yu, D.; He, L.; Zhang, Y.; Du, X.; Pasupat, P.; Li, Q. Few-shot intent classification and slot filling with retrieved examples. *arXiv* **2021**, arXiv:2104.05763.
22. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
23. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
24. Firdaus, M.; Golchha, H.; Ekbal, A.; Bhattacharyya, P. A deep multi-task model for dialogue act classification, intent detection and slot filling. *Cogn. Comput.* **2021**, *13*, 626–645. [CrossRef]
25. Liu, B.; Lane, I. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv* **2016**, arXiv:1609.01454.
26. Zhang, X.; Wang, H. A joint model of intent determination and slot filling for spoken language understanding. In Proceedings of the IJCAI International Joint Conferences on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; Volume 16, pp. 2993–2999.
27. Goo, C.W.; Gao, G.; Hsu, Y.K.; Huo, C.L.; Chen, T.C.; Hsu, K.W.; Chen, Y.N. Slot-gated modeling for joint slot filling and intent prediction. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 753–757.
28. Qin, L.; Che, W.; Li, Y.; Wen, H.; Liu, T. A stack-propagation framework with token-level intent detection for spoken language understanding. *arXiv* **2019**, arXiv:1909.02188.
29. Chen, Q.; Zhuo, Z.; Wang, W. Bert for joint intent classification and slot filling. *arXiv* **2019**, arXiv:1902.10909.
30. Jeong, M.S.; Cheong, Y.G. *Comparison of Embedding Methods for Intent Detection Based on Semantic Textual Similarity*; The Korean Institute of Information Scientists and Engineers: Seoul, Republic of Korea, 2020; pp. 753–755.
31. Heo, Y.; Kang, S.; Seo, J. Korean Natural Language Generation Using LSTM-based Language Model for Task-Oriented Spoken Dialogue System. *Korean Inst. Next Gener. Comput.* **2020**, *16*, 35–50.
32. So, A.; Park, K.; Lim, H. A study on building korean dialogue corpus for restaurant reservation and recommendation. In Proceedings of the Annual Conference on Human and Language Technology. Human and Language Technology, Tartu, Estonia, 27–29 September 2018; pp. 630–632.
33. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
34. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
35. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555.
36. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
37. Kudo, T.; Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv* **2018**, arXiv:1808.06226.
38. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
39. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3. [CrossRef]
40. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]

41. Choi, J.; Lee, J. Redefining Korean road name address system to implement the street-based address system. *J. Korean Soc. Surv. Geod. Photogramm. Cartogr.* **2018**, *36*, 381–394.
42. Park, J.H.; Myaeng, S.H. A method for establishing korean multi-word concept boundary harnessing dictionaries and sentence segmentation for constructing concept graph. In Proceedings of the 44th KISS Conference, 2017; Volume 44, pp. 651–653.
43. Hur, Y.; Son, S.; Shim, M.; Lim, J.; Lim, H. K-EPIC: Entity-Perceived Context Representation in Korean Relation Extraction. *Appl. Sci.* **2021**, *11*, 11472. [CrossRef]