Survey for spoken language understanding in dialogue system

1710225 Sixia Li

Spoken language is an indispensable thing in our daily life. People communicate with each other mainly through spoken language. The spoken language is playing an important role in our daily communication. According to the speech act theory, people communicate with each other is to convey intended actions to each other. Based on this theory, the understanding to spoken language can be described by understanding to locutionary act, illocutionary act and perlocutionary act. For this reason, the understanding to spoken language can also be described by these speech acts.

The understanding to locutionary act can be considered as understanding to spoken content. For now, this understanding contains information extraction, text recognition and semantic relationship recognition. These tasks mainly represent the understanding to the content itself and the basic actual information that speaker wants to give.

The understanding to illocutionary act can be considered as understanding to intention, emotion or any intended willing but not appeared in the spoken content. This understanding contains dialogue act recognition, emotion recognition, intention modeling. These tasks mainly make an approach to get hidden information in given utterance and find out the real will that speaker wants to convey.

These two acts are focus on identifying speaker, the perlocutionary act is focus on listener, in our research we want to focus on analyzing speaker, for this reason, in this report we do not survey the understanding to perlocutionary act.

Modeling spoken language understanding (SLU) for computer is to make computer be able to understand human's spoken language, in other words, is to understand human's speech acts. This goal can be considered as a way to make an advanced AI. In this processing, the mechanism of human conveys their intentions during dialogue conversation can also be researched by analyzing and modeling the speech acts into mathematical representation.

Our final goal is to make computational model for understanding illocutionary act in dialogue. In this report, we firstly make a short survey of tasks in SLU, and then focus on development of SLU in dialogue systems.

In the short survey of tasks in SLU, this report firstly surveyed the understanding of locutionary act, which is mainly extract information from given content. At early time, the key words or phrases are used directly to extract information from given utterance by matching the pre-defined dictionary, but this kind of way can only handle finite domain

information because the pre-definition can only be done in finite conditions. For this reason, a statistical method is proposed. During modeling by statistical method, the probability representation of syntactic or semantic information is also proposed, this kind of representation are trained by statistic model and big dataset based on statistical results, the advantage of this method is the linguistic information can be represented in a mathematical space and can capture the appearance relationships between words, this makes model can understand not on a word level, but can understand a structure in given utterance. However, it also has disadvantage, that is the performance of model is good or bad is totally rely on training dataset, the generalization to other datasets may be not good and it cannot handle the recognition of semantic relationships, without understanding of semantic relationship, the model cannot be considered as understanding the real meaning of given utterance. To solve the processing of understanding semantic relationship, with the development of neural networks (DNN), an advanced representation Word2Vec has been proposed, this representation is trained by neural networks and can represent linguistic information in a vector space, with this representation the performance of many locutionary act understanding tasks have been improved. This representation method has generation capability, but the performance of this representation is still relied on training dataset. Recently, BERT representation has been proposed, it is showed with this representation method, many understanding tasks have been improved, it can be expected that this representation can be a good usage for SLU task.

In illocutionary act understanding, unlike the locutionary act, the linguistic information is not enough for the understanding. The application of paralinguistic information is also necessary. However, the relationship between paralinguistic information and illocutionary act is still not clear, for this reason, there are some studies to find out how to model illocutionary act understanding with paralinguistic information. At early time, the selected features are showed that they are useful, such as F0 and energy and their derivations. Then, for some specific tasks, some feature sets have been shown that be useful, such as IS09E for emotion recognition and 57 new features for dialogue act recognition. These feature sets contain more features and can make model have better performance on some tasks than selected ones. Recently, the modeling of illocutionary act also use spectrogram directly. This is a way to use original information directly by some neural network structure, such as CNN. Other paralinguistic information that are showed be useful are prosody and gestures, but these features are not well-used in SLU tasks.

After surveying the features is survey the development of SLU in dialogue system, this report focuses on applications of SLU algorithm in dialogue system for early studies

and focuses the recent studies on themselves for they are not well-applied in dialogue systems for now.

The ELIZA system is the very beginning of modern dialogue system, it uses key-words directly and use pre-defined rules to understand the given utterance, but for the property of rules this system can only understand part of locutionary act and cannot understand the illocutionary act of speaker. To improve this, a frame-driven dialogue understanding algorithm was proposed, it use a pre-defined framework to match given utterance and understand which pattern the given utterance is, based on this understanding system will process for the next step. GUS system is a basic frame-driven dialogue system for travel management, but still, the frame can handle only several tasks in limited domain for it needs pre-definition. To make dialogue system have generalization capability, the statistical algorithm was developed, in such algorithm, the understanding model is trained by statistical method such as N-grams and HMM, this make the system can capture statistical relationship in sentence level and in the level between given utterance and illocutionary act. In this way, the system can understand both locutionary act and illocutionary act, but it has similar disadvantages with the representation of statistical linguistic features, which is very relied on training data. To make system have more generalization capability, based on the development of neural network, many end-to-end models have been proposed, these models can understand utterance and locutionary act well by using sequence to sequence structure, and can understand illocutionary act by recognizing or classifying dialogue acts. These models can also understand given dialogue through turns.

From the survey, it is showed that the neural network based models have very good performance on SLU tasks, but there are still some problems. One is the representation of speech act is still on label level, such as dialogue acts. For this representation, modeling the speech acts relies on the end-to-end training, but this kind of end-to-end training is a black box and cannot be explained very well, this makes the study of find out mechanism of human convey their intentions into an unexplainable condition. Another problem is even the performance is good based on neural networks, but the neural networks is still a statistical method essentially, for this reason its generalization capability is limited.

For these problems, the future work can be modeling and explanation the illocutionary act based on using more features that are not flexible to use, not only rely on training of neural networks, but also make explainable hypothesis and verify them.