Master's Research Project Report

Survey for spoken language understanding in dialogue system

1710225 Sixia LI

Supervisor	Jianwu Dang
Main Examiner	Jianwu Dang
Examiners	Masato Akagi
	Masashi Unoki
	Atsuo Yoshitaka

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

February 2019

Abstract

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 information, 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 information that are not flexible to use, not only rely on training of neural networks, but also make explainable hypothesis and verify them.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Background

Spoken language is an indispensable thing in our daily life. People communicate with each other in speech, dialogue and other ways, in these ways the spoken language is playing an important role [1]. According to the speech act theory, communication by speech is communication with the intention or action that people want to convey, for this reason, understanding to spoken language can be described as understanding to speech act, which contains locutionary act, illocutionary act and perlocutionary act [2,3]. The locutionary act can mainly be the real meaning of spoken utterance, the illocutionary act can is an intended significance as a socially valid verbal action to the spoken utterance, the perlocutionary act is the actual effects of the spoken utterance. From these definitions, it can be considered as the locutionary act and illocutionary act are cues to understand the action of speaker, and the perlocutionary is the cue to understand the influence of spoken contents. For this reason, the understanding to spoken language can be considered as both the understanding to locutionary act and the understanding to illocutionary act.

Modeling spoken language understanding (SLU) for computer is to make computer be able to understand human's spoken language, this kind of task is necessary and important for many applications, such as text classification, emotion recognition, and for the importance of interaction between human and machine, the dialogue system is an important application and part of spoken language understanding. To make a dialogue system that can understand both human's contents and intentions 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.

Like the way human understand each other, in such studies, not only linguistic information, but the paralinguistic and nonlinguistic information are also used to made approaches to such task.

In this report, to make an overview and handle the direction of spoken language understanding, we survey the studies in such area by the order of development. Then, because of the importance of dialogue system, we focus on SLU in dialogue system and survey studies about it. Finally, the achievement and the problems that still remaining is

summarized in the end of report.

1.2 Spoken language understanding and dialogue system

As mentioned above, spoken language understanding (SLU) can be defined as understanding the content and the intention. there are different approaches for these two main goals of SLU.

In the studies of understanding to content, the information extraction a typical task of SLU to content, the goal of such task is to extract or summarize the content of given utterance and to understand it. Then with the information technique development and popularization, the text classification task became important for data classification, for this task, some linguistic representations are proposed, and achieved good performance. In addition, the extraction of relationship between parts in the given utterance is also an important task for the understanding of contents. Because the contents are the basic information in a spoken language utterance, the processing of contents is a basic way to make the spoken language understanding model. In such way, the linguistic information is the main cue, and the natural language processing (NLP) methods have been extensively used.

In the studies of understanding of intention, the emotion recognition and user's attitude identification are two typical tasks. In such tasks, to distinguish different emotions or attitude with same contents, only the linguistic information cannot handle such goals very well. For this reason, in methodology design, more information from spoken language are needed, such as paralinguistic and nonlinguistic information.

Besides the tasks mentioned above, the dialogue system task is a SLU task that need both understanding to content and understanding to intention. A dialogue system is a computer system intended to converse with a human with a coherent structure.

Dialogue system can mainly be classified into two categories, task-oriented system and chatbot system. From the very beginning of dialogue system ELIZA [4] to the newest one made based on neural networks [8], the task-oriented system is more developed than chatbot system, because to make a chatbot system must consider more aspects than task-oriented system. But as the development of newest technology, some chatbot systems are developed in a good performance, such as Xiaobing from Microsoft.

In dialogue system task, the understanding to content is to make computer be able to understand human's dialogue and make dialogue to human. Some models have been developed for this goal. In the early time, this kind of models can only understand the utterance from human in finite key words or domains [4-6]. Recent years, with the development of machine learning and neural networks, some models can generate the

answer in a larger domain and have better performance than before [7-9]. Because the contents are the basic information in a spoken language utterance, the processing of contents is a basic way to make the spoken language understanding model. In such way, the natural language processing (NLP) methods have been extensively used.

The understanding to intention is to make computer understand human's expression of their real will. Because it is hard to describe intentions directly, the recognition or classification of dialogue acts are approaches for this goal. Dialogue act can be considered as the illocutionary act in the speech act theory, which can describe the speaker intended action. For such tasks, some models are proposed and achieved good performance up to now [10, 11].

1.3 The objective of this report

Our final goal is to make computational model for understanding both locutionary act and illocutionary act in dialogue. To approach such goal, an understanding of how the locutionary act and illocutionary act are modeled up to now is necessary. For this reason, the goal of this report is to survey the studies of modeling spoken language understanding. Then, because of our goal is faced to dialogue processing, this report will mainly focus on the SLU algorithm in dialogue system after an overview on SLU tasks.

1.4 The structure of this report

There are 5 chapters in this report. Chapter 1 is the introduction of this survey. Chapter 2 makes a short survey to the locutionary act and illocutionary act understanding in SLU tasks, chapter 3 surveyed the development of SLU algorithm in dialogue system and recent studies that are not used in dialogue system yet. Chapter 4 summarized the studies and the still remaining problems, then propose an approach to the feature research. Chapter 5 is the conclusion of this report.

Chapter 2

Spoken language understanding tasks

As mentioned above, the spoken language understanding can be divided into two kinds of tasks, one is understanding to locutionary act, the other one is understanding to illocutionary act.

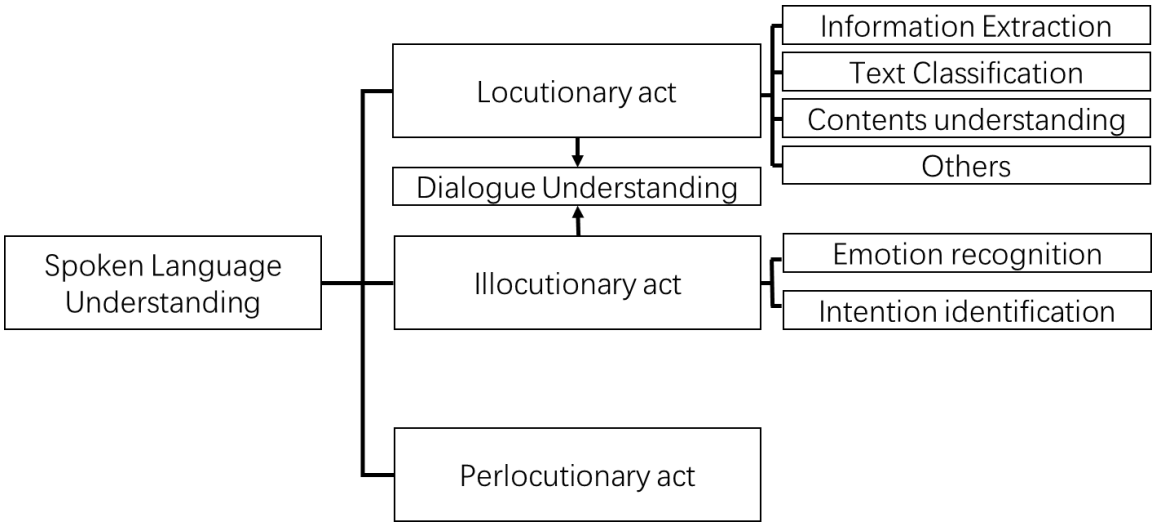

Figure 2.1: Tasks in spoken language understanding

The understanding to locutionary act is considered as understanding to content of given utterance. The main goal of this task is to understand the dominant information that expressed in given utterance, there are some different approaches to make close to such goal, such as information extraction, text classification and semantic relationship detection. This report will not focus on one of them but make a generalized survey for the developments of methods for understanding on content. Because the locutionary act mainly appears by linguistic information, the development of understanding of locutionary act is also a development and changes of linguistic representation for linguistic information.

The understanding to illocutionary act is considered as understanding the real will or intended action of speaker but not expressed on dominant level. It always hides in the utterance not only via linguistic information, but also via paralinguistic information such as prosody or emotion. There are not many approaches to understand illocutionary act by dense representation, such as vector representation for intention. For now, the well-used ways are making labels of emotion or intention to given utterance and build model to

recognize the labels. In such way, only the linguistic information is not enough, multimodal information is necessary. It is showed that paralinguistic information has its specific effect for SLU task [1,13], and nonlinguistic information that are not related with the spoken language utterance itself are also considered be useful for SLU tasks [12].

In next parts, we will survey the developments and changes of understanding of locutionary act and illocutionary act based on applying different information.

2.1 Understanding of locutionary act

To understand the locutionary act is mainly focus on extracting information from given utterance. In this kind of task, the linguistic information is the most useful thing. The problem solved in studies about locutionary act understanding is also an improvement of using linguistic information, or a proposal of new type of linguistic information.

2.1.1. Rule or frame methods

A simple consider of getting information from understand content of given utterance is extracting of useful information from it. Focus on key words only and extract them can be an approach. A more flexible approach can be using rules or frame to match the utterance and extract information by defining necessary part of it can be another approach.

In some early dialogue systems [4] or order system [14], treating words as key words was a useful approach to understand the content and trigger next process of system. It can

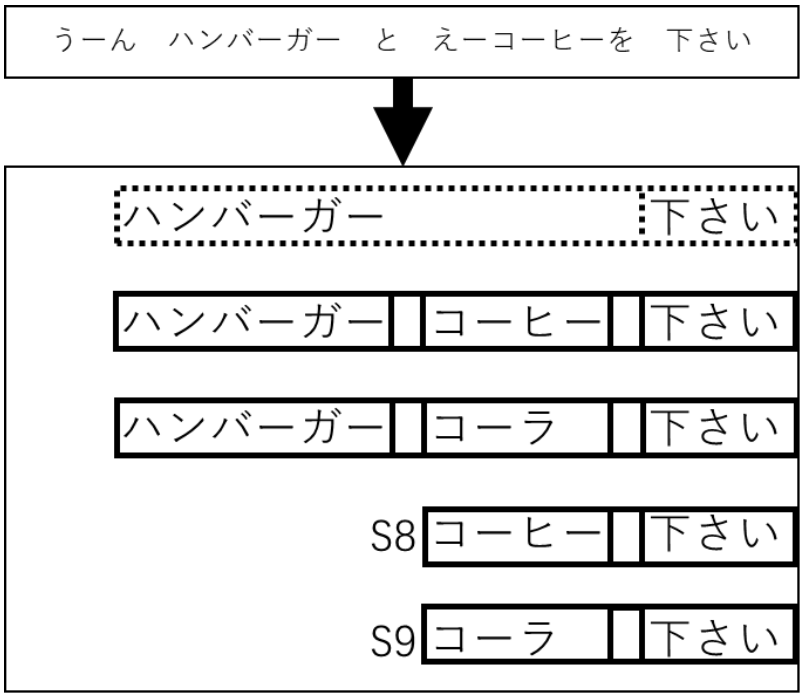

Figure 2.2: Example of using key words detection method (from [14])

also be a step of generating give-back answer for given utterance. This kind of way makes its system can understand the useful information only and get rid of influence from other words. Here is an example of this method in figure 2.2.

However, this way did not consider the meaning and semantic relationship between words and is weak for out-of-vocabulary (OOV) words, for these reasons, recent studies do not use such design to extract information so much, but in application of some question-answer or task-oriented dialogue system, the using of key words and frames is still very generally for its low cost and ease to make model.

Another way is to treat key words into its corresponding category or framework [5, 15], in this way it is also necessary to pre-define a rule or frame to match the given utterance and get the useful parts. This kind of way makes system can understand what category the input utterance belongs to and can understand the meaning of given utterance in a structured level, and then according to framework, the system can generate answer and give back to user. But still, this kind of way cannot be used to understand semantic relationship through utterance, and also weak for understanding OOV words.

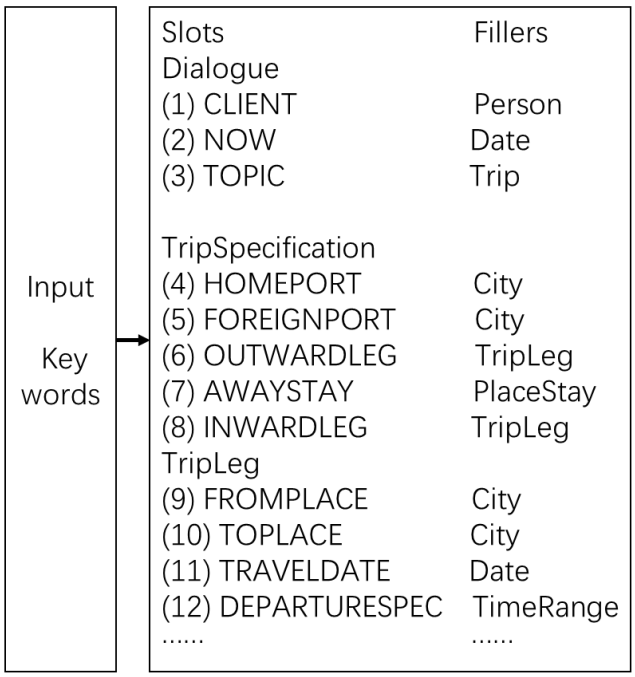

Figure 2.3: Example of frame method (from [5])

2.1.2. Statistical methods

As mentioned above, based on key words or rules cannot make computer understand the

content in many cases, such as understanding OOV words or semantic relationship in utterance. Another disadvantage of such methods is the key words dictionary or rules must be pre-defined, this makes the understanding cannot perform in large domain and cannot be generalized well.

It is necessary to use a kind of method that can represent semantic feature and have generalization for large amount of words for understanding content, with such methods the computer can understand not only the information of what content is but can also understand the relationship between contents.

With such needs, the statistical method has been proposed, with such method, the linguistic information can be treated as a probability representation. At beginning, this kind of representation is usually trained by data in a statistic way, with the development and application of well-performance deep neural networks (DNN), some DNN based trained representation have been proposed and developed, and they have a better performance on many tasks than traditional statistic-based representation.

Normal statistical methods

To use statistical methods for SLU task, a statistical representation of linguistic information is needed to be trained firstly. Relies on language model [16, 17], it is trained by labeled data and can be used for information extraction from utterance with given words or phrases. It can also be used to predict the coming utterance for given utterance. For these properties, it is considerable to use such linguistic feature in dialogue system to generate answer.

The statistical language model can be trained in many algorithms, such as by n-grams [18], by maximum-likelihood [19], by conditional random fields (CRF) [20] and by support vector machines [21]. The choice of training method is based on methods of the statistical methods used in understanding to locutionary act.

Basic thought of this kind of representation is to capture the probability relationship between words in a big data by applying statistic processing. For example, using n-grams to train the language model, the probability relationship of words from given dataset can be describe as:

$$\mathrm{P}(W_n|W_{n-N+1}^{n-1}) = \frac{C(W_{n-N+1}^{n-1}W_n)}{C(W_{n-N+1}^{n-1})}$$

in which W_n means the word n. When n=2, this representation becomes bigram, it can represent the probability of appearance of a word after given word based on frequency. Here is an example of bigram statistic result in the Berkeley Restaurant Project corpus of 9332 sentences:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Figure 2.4: Example of bigram result (from [22])

from such statistic result, we can compute the probability relationship between words, such as $\mathrm{P}(\mathrm{food}|\mathrm{chinese}) = 0.0065$. Certainly, only use bigram or n-grams method is a rough way to get more information from given data, for this reason, there are many smoothing methods to make the statistical representation better, such as treat with OOV words.

For now, there are some toolkits that can be used for building language models, two commonly used toolkits are SRILM [23] and KenLM [24,25]. SRILM offers multiple options and types of discounting, while KenLM is optimized for computation speed and memory size [18].

With these kinds of representations, the understanding model can be built. For example, with a representation trained with corresponding labels of linguistic information, an statistical model can be built to process the part-of-speech, which can tag the parts into different categories, and the system can understand the content with these categories [67]. This kind of statistical methods can also be used for text segmentation and named entity recognition.

Although statistical methods and representation of linguistic information make the SLU tasks not only relying on rules or pre-defined dictionary , this kind of method is totally based on dataset and statistics, it is hard to extend the representation to sentence-level and does not get the meaning relationship between words. For this reason, this kind of representation is weak to distinguish the different meaning of words, and the system that rely on this representation is able to have good performance on understand by appearance probability but cannot represent and understand meaning simultaneously.

Word vector

To make the model can understand semantic meaning of words, the statistical methods is not able to handle it. With a representation that can represent both appearance probability and semantic relationship, this task can be handled. Based on this consideration, word vector representation was proposed. At early time, the vector model of words meaning are generally based on a co-occurrence matrix, this is a close way to the statistic representation, but focus on meaning distribution. Based on this thought, the term-document matrix was proposed as part of the vector space model of information retrieval [26], in this study, documents are classified by number of some picked words and map the result to a vector space, such that the difference of documents' vector can be computed and can be retrieved.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

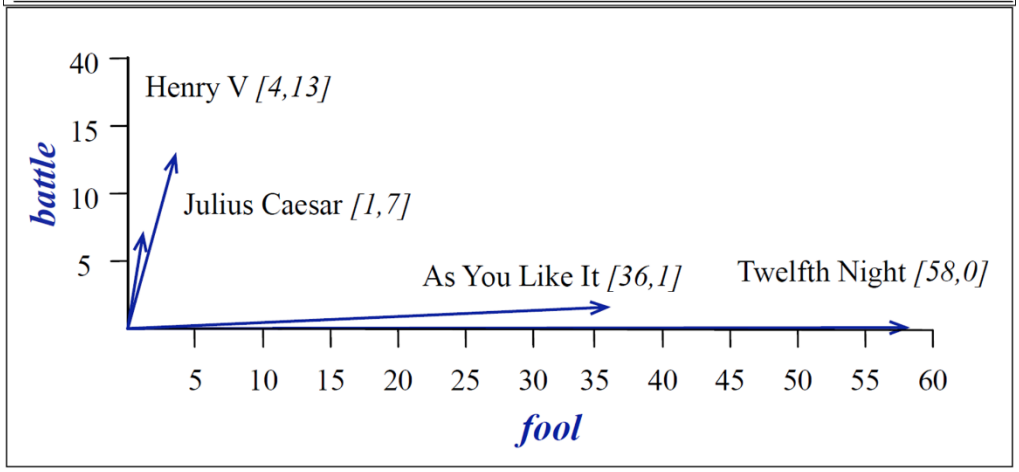

Figure 2.5: Example of term-document matrix and
its mapping on vector space (from [26])

Like treating documents in this way, it can be considered to treat words in the same way to make a semantic representation. A word-word matrix was proposed and based on such kind of matrix, semantic information can be mapped into a vector space. Then, by the cosine distance, the difference of semantic meaning can be computed by this kind of representation.

This kind of representation can be used on document classification based on topic and compute the similarity of two words. This makes the understanding can understand not only the contents that given, but also can understand some related knowledge based on vector representation.

	aardvark	...	computer	data	pinch	result	sugar	...
apricot	0	...	0	0	1	0	1	
pineapple	0	...	0	0	1	0	1	
digital	0	...	2	1	0	1	0	
information	0	...	1	6	0	4	0	

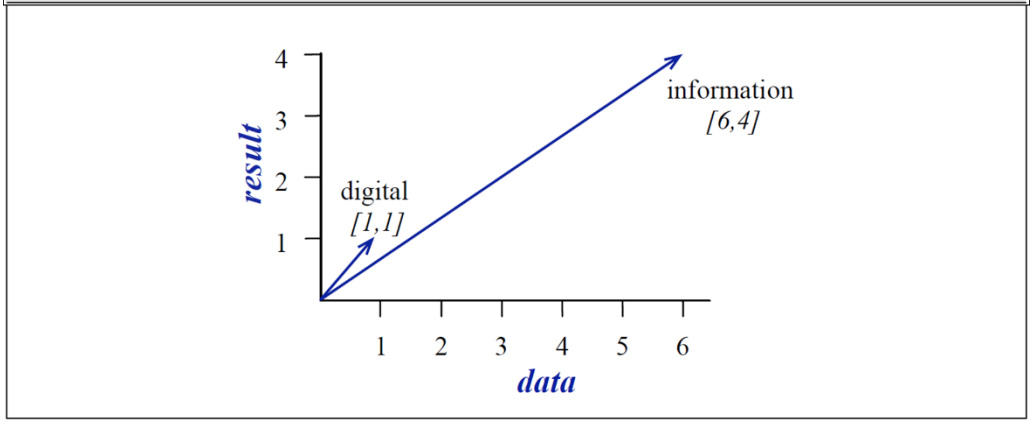

Figure 2.6: Example of word-word matrix and
its mapping on vector space (from [18])

But this representation is made for documents retrieving and not well-used for dialogue understanding because it is based on co-occurrence in one document, but in dialogue, different people have different customers for word-using and speech structure, this made this kind of word vector representation hard to use in understanding content in dialogue processing and article classification.

With the development of neural network, some new representation of linguistic information has been proposed, the Word2Vec model [27,28] is the most famous and useful one to represent semantic feature for words. The application of this kind of representation makes a better performance on semantic and syntactic relationship recognition than before [27], this means the model that applied this representation can understand such relationships better.

Word2Vec can be trained in 2 ways, one is by skip-gram, the other one is by CBOW. In skip-gram training, the model can predict contextual information based on one word in an utterance. In CBOW training, the model can predict contextual information based on given word.

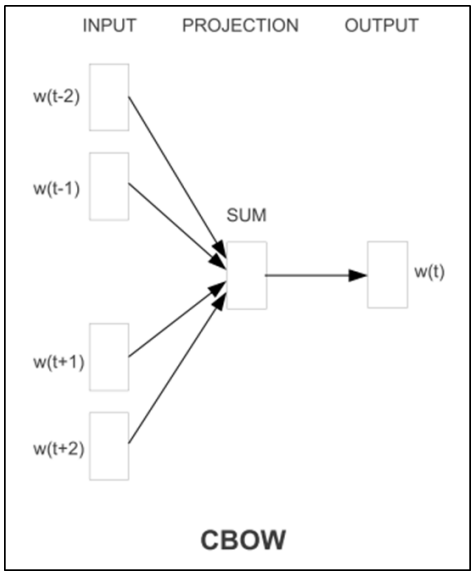

Figure 2.7: Training model by CBOW (from [27])

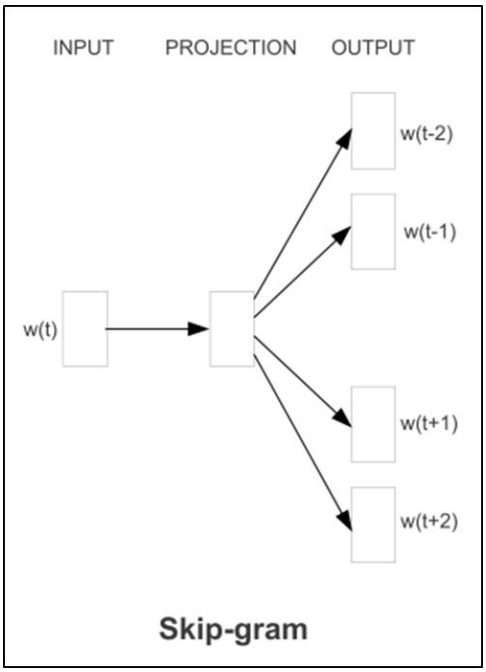

Figure 2.8: Training model by skip-gram (from [27])

The Word2Vec representation has been widely used in many NLP and SLU tasks as linguistic feature representation, for its performance on capturing semantic relationship and relationship between words.

Based on consideration of Word2Vec, some other representations for different level of linguistic feature have been proposed. An interesting one is making Chinese character (Kanji) parts into vectors representation and it achieved good performance in some Chinese SLU tasks, such as text classification and named entity recognition tasks [29].

New representation

Since Word2Vec is a good representation for linguistic feature, a better-performance representation BERT (Bidirectional Encoder Representation from Transformers) [30] has been proposed, this representation trained by bidirectional encoder and the application of BERT get better performance on many NLP tasks. Even if this kind of representation has not been used widely, but it can be expected to use in SLU tasks.

2.2 Understanding of illocutionary act

The understanding of illocutionary act is to understand the intended action of speaker, such actions contain intention, emotion and other aspects that can express speaker's real will. The illocutionary act sometimes is not obvious from the spoken contents, it can be hidden and express in dialogue by speaker. For this reason, to understand the illocutionary act, only the linguistic information is not enough, using other information is also necessary.

Paralinguistic feature contains many aspects, acoustic feature, prosody, emotion and gesture [31,32], facial expression [33,34,35] and other information in acoustic data, such as cough and pause.

The paralinguistic information is considered as cues that reveal a speaker's intention in dialogue processing [36] or helps the transmission of verbal information [37]. For this reason, applying for paralinguistic information in SLU tasks could be a useful approach to recognize speaker's illocutionary act.

However, the relationship between understanding of illocutionary act and this paralinguistic information and the way to use them in building model are still in researching. There are some attempts of applying different paralinguistic information in modeling illocutionary act understanding and achieved good performance.

2.2.1. Illocutionary act understanding with acoustic features

Acoustic feature mainly plays an assistant role in SLU task. At early time, SLU task relied on understanding of locutionary act only and not considered illocutionary act well because it is difficult use linguistic information to understand illocutionary act. Recently, some studies begin to use acoustic features as an assistant parameter to the SLU of illocutionary act.

At beginning, researchers always chose feature by experience, such as choosing some sub-features of F0, energy and duration feature [38,39], these features are showed that they are very useful cues for illocutionary act understanding in these studies. For the needs of applying such features and with the acoustic feature extraction and analysis

studies developing, some toolkits have been proposed, Praat [40] and OpenSmile [41] are well-used toolkits, with these toolkits, the extraction of acoustic feature becomes easier.

For some specific illocutionary act understanding, such as emotion recognition, some feature sets have been applied as paralinguistic information in SLU task, such as Interspeech 2009 Emotion Challenge (IS09E) [42] and Interspeech 2013 Computational Paralinguistic Challenge [43].

The IS09E feature set contains totally 384 features, they 16 categories of features of ZCR, RMS Energy, F0, HNR and MFCC 1-12, and their first order difference, in each category, there are 12 function to describe the property of that category, detailed introduction in table 2.1. With these features, the accuracy of emotion recognition has been improved [42].

Feature	Functionals
ZCR	mean
RMS Energy	standard deviation
F0	kurtosis, skewness
HNR	extremes: value, real position, range
MFCC 1-12	linear regression: offset, slope, MSE

Table 2.1: Caption of the table (from [42])

Feature Name	Description
Pitch and Voicing ($\mathcal{P}$)	
1. prcF0_extremes*	percentage of F0 values that are less than 0.75*(F0 floor) or more than 1.5*(F0 floor)
2. loc_σF0_gt_XprcΣF0	normalized location in utterance ($\in [0,1]$) where the cumulative F0 sum surpasses X% of total F0
3. loc_maxF0	normalized location in utterance ($\in [0,1]$) where F0 attains its maximum value
4. avg_vSegDur*	average duration of continuous voicing
5. num_vSegs*	number of continuously-voiced segments
6. min_dct123_F0Segs	F0 contour → continuously-voiced segments → [for each segment, percentage energy in the first three DCT coefficients of F0] → minimum over all segments
7. prcNcorr_gt_0.9	percentage of voiced frames having a normalized cross-correlation greater than 0.9
8. min_range_F0Segs†	F0 contour → continuously-voiced segments → [for each segment, (max F0 − min F0) normalized by F0 floor] → minimum over all segments
Duration and Pausing ($\mathcal{D}$)	
1. loc_firstPause	normalized location in utterance ($\in [0,1]$) where the first pause occurs
2. dur_by_dur{P/N}	ratio: duration of the given utterance to duration of the previous/next utterance
3. num_spSegs	number of segments having continuous speech activity
4. pause_dur{P/N}	duration of silence between the given utterance and the previous/next utterance
5. prc_vSpeech	percentage of speech frames that are voiced
Intensity ($\mathcal{I}$)	
1. 98prcInts_spRegs*†	98th percentile of intensity (excluding non-speech frames), normalized by intensity floor
2. avg_dct123_intSegs	intensity contour → 300 ms chunks → [for each chunk, percentage energy in the first three DCT coefficients of intensity] → average over all chunks
3. stdev_intPks	standard deviation of intensity peaks that are higher than 0.5 times the mean peak value
4. stdevInts_spRegs*†	standard deviation of intensity (excluding non-speech frames), normalized by intensity floor
Speaking Rate and Rhythm ($\mathcal{S}$)	
1. 95prcEntr_utt*	given utterance → 200 ms chunks → [for each chunk, spectral entropy of Mel filter-bank output] → 95th percentile over all chunks
2. modn_dct2to10_beg	first one second of the given utterance → Mel filter-bank output → [for each filter-bank channel, percentage energy in DCT coefficients 2–10] → average over filter-bank channels
3. avg_modn_dct2to10	given utterance → 1 sec. chunks → [modn_dct2to10_beg for each chunk] → average over chunks
4. rangeF1_utt†	(max F1 − min F1) normalized by median value of F1
5. min_stdevF1_200	F1 contour → 200 ms chunks → [for each chunk, standard deviation of F1] → minimum over chunks

Figure 2.9: Part of 57 new features from [44]

Recent years, some artificial feature sets have been proposed for different tasks. A useful one for dialogue act recognition is from [44], in this study, 57 new features with normalization or utterance-related variance have been proposed, they are extracted from pitch, duration, intensity and speaking rate. With these features, the performance for dialogue act recognition have been improved. Some of these features are showed above [57]. Unlike using acoustic features directly, this kind of feature set pre-process the features to capture a fixed representation of pitch, duration, intensity and speaking rate. It makes an approach to explore the relationship between illocutionary act with acoustic features in an analysis way.

The advantage of such feature set is making the difference of features more outstanding for specific tasks, on the other hand, the disadvantage is that these processing for features are done by experience and may weak for generalization to other tasks.

Because of the disadvantage of human-selection features, there is also a way to use paralinguistic information directly. In this second way the spectrogram is treated as an image to be processed. With the ability of space feature capturing, the CNN structure neural network is widely-used in image identification, in the using of spectrogram as paralinguistic feature in SLU tasks, CNN is usually used for capturing acoustic feature [45]. This way of using paralinguistic information can cover the disadvantage of human-

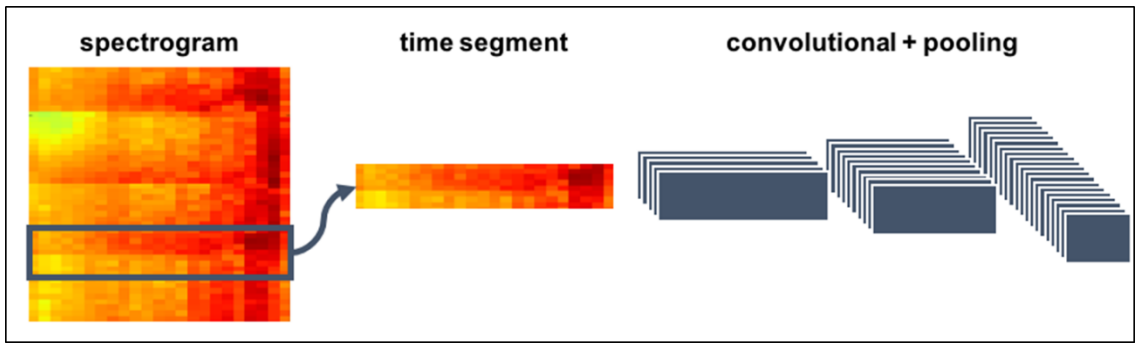

Figure 2.10: Example of using CNN on spectrogram

selection features, but it becomes into a black box for the property of neural network structure, with this way the accuracy of emotion has been improved [45] but it is not explainable. And it is hard to analyze the relationship between paralinguistic information with illocutionary act.

2.2.2. Illocutionary act understanding with other features

For now, other paralinguistic information such as prosody and gestures are not well-used in SLU task to be an assistant parameter for illocutionary act understanding. But there are some studies showed that these features are helpful cues for recognizing speaker's

intentions.

Prosody conveys additional information that goes beyond the linguistic content [46], it contains some acoustic feature such as performance of pitch, loudness on given utterance [47] and some prosodic features such as prosody structure, stress and intonation. In studies about prosody in Mandarin language, it is showed that stress is related with focus [48] means the content or part of utterance that speaker want to emphasize [49].

T-test value	Gesture feature	Acoustic feature
Less than 5%	Mean of shoulder location on Y axis	Dialogue duration
	Mean of shoulder location on Z axis	Pitch maximum
	Mean of elbow location on Y axis (side)	Energy minimum
	Mean of norm of shoulder (side)	Mean of MFCC
Less than 2%		Pitch minimum
		Mean of pitch

Table 2.2: Gesture relationship with attitude (from [51])

Gesture is also a related feature to the intention recognition. There are some studies use gesture as paralinguistic information in dialogue system. For example, the head gesture can be used for a robot dialogue system to recognize whether speaker's attitude is positive or negative [50], this system combines prosody feature with head gestures with different definitions of movement and use HMM model to make the recognition model. Another study of interview robot development showed shoulder and elbow gesture during dialogue has a relationship with user's attitude to current topic [51] by applying T-test, this kind of attitude is also user's illocutionary act.

Although it is showed that these features can be useful cues in SLU task in dialogue, it is still not widely used in such task. One reason for prosody can be considered as the accuracy of automatically extraction of these features is not high, and reason for gesture feature can be considered as high cost and not easy to get annotated data for training or make system.

Although it is still not very clear for relationship between illocutionary act and paralinguistic information, but for the linguistic information cannot handle some intention expression, and the extraction of paralinguistic information becoming easier and easier, the multimodal model can be considered as the direction of SLU studies.

Chapter 3

SLU in dialogue system

In this chapter, this report will focus on SLU in dialogue system based on different algorithm model of dialogue systems. And in a flow of time and development.

3.1 SLU in Rule based model

ELIZA [4] is a typical rule-based dialogue system with using words directly as linguistic feature. The processing of ELIZA can be summarized into 3 steps. Firstly, recognize key words that in pre-defined rules, then change words and word order of given utterance to match the generation rules. Finally, generate answer based on rules.

Here is an example of processing given utterance 'You are very helpful':

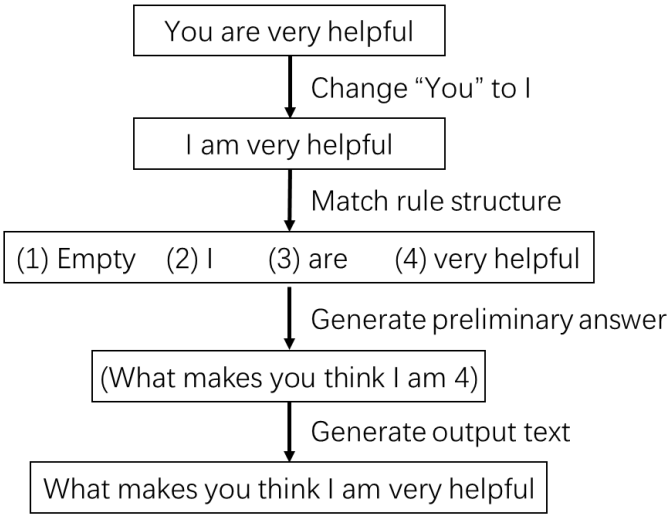

Figure 3.1: Example of processing by ELIZA (from [4])

From the processing we can get an output text 'What makes you think I am very helpful' to the given utterance. By this kind of rules, ELIZA can make dialogue continuously but cannot give back new contents of dialogue.

In the view of spoken language understanding, ELIZA is a totally text-based dialogue agent and has no capability to treat with original oral data because the speech recognition was not much developed in 1960s as its situation recently. Moreover, ELIZA cannot understand contents or intentions of speaker well because the rule only process with how to change words to give back answer. For this reason, ELIZA can only understand what the input words is and their relationship with some other words based on rules but cannot understand whole meaning of input contents and intentions.

But still, ELIZA is an early exploration of dialogue system in this area, even if it has some disadvantages.

3.2 SLU in Frame-driven model

To make a dialogue system that can understand not only the locutionary act such as contents, but also the illocutionary act such as intentions, an approach of this goal is to understand the speech act of speaker. As mentioned before, speech act contains locutionary, illocutionary and perlocutionary acts, to understand content is to understand locutionary act, for this reason, to make utterance into categories can be considered as a generalization way than only treat with words and use rules.

A frame-driven dialogue system can be used to make words in given utterance into different categories and understand the contents by matching frames [53] for specific tasks.

The GUS system [5] is an early frame-driven dialogue system faced to a travel arrangement task. It recognizes key words as linguistic feature to deal with speaker's travel needs. In the processing of GUS, firstly it recognizes the category of input text, based on different categories of inputs, system will make different processing.

Input Structures	Permanent Knowledge Structures	Processes	Outpu Structures
1. Text String (input)	Stem dictionary: Morphological rules	Dictionary lookup: Morphological analysis	Chart of word data structures
2. Query context (6): Chart (1)	Transition net grammar	Syntactic analysis	Parsing of a sentence
3. Parsing of a sentence (2)	Case-frame dictionary	Case-frame analysis	Case-frame structure
4. Case-frame structure (3)	Speech patterns: Domain specific frame forms	Domain dependent translation	Frame change description
5. Frame change descriptions (4, 5): Current frame instances (5)	Prototype frames and attached procedures	Frame reasoning	Frame change descriptions Output response descriptions: Current frame instances
6. Output response description (5)	Dialog query map: Flight description template	Response generation	English text: Query context

Figure 3.2: GUS knowledge structures (from [5])

Then the system divides input utterance into parts and use pre-defined framework to match them, after this, system will generate knowledge data and can find out what parts of framework are lacked, then question speaker for the lacked information for circulation. Finally, after the framework are filled, system will treat the arrangement task as completed.

When generate the knowledge data, GUS uses slots as categories for necessary data and guide the servants of system.

Slots	Fillers	Servants	Demons
Dialogue			
(1) CLIENT	Person	Create	Link to TRAVELLER
(2) NOW	Date	GetDate	
(3) TOPIC	Trip	Create	
TripSpecification			
(4) HOMEPORT	City	Default-Palo Alto	
(5) FOREIGNPORT	City		Link to OUTWARDLEG
(6) OUTWARDLEG	TripLeg	Create	
(7) AWAYSTAY	PlaceStay		
(8) INWARDLEG	TripLeg	Create	
TripLeg			
(9) FROMPLACE	City	FindFrom	
(10) TOPLACE	City	AskClient	
(11) TRAVELDATE	Date	AskClient	
(12) DEPARTURESPEC	TimeRange	AskClient	Propose-Flight-By-Departure
.....	.....	.....	.....

Figure 3.3: GUS slots (from [5])

Here is an example of reasoning from given utterance to fill the slots:

```
CLIENT: I want to go to San Diego on May 28
  CMD: [CLIENTDECLARE                          ... the domain dependent translation
    (FRAME ISA TRIP-LEG
      (TRAVELLER (PATH DIALOG CLIENT PERSON))
      (TO-PLACE (FRAME ISA CITY
                       (NAME SAN-DIEGO)))
      (TRAVEL-DATE (FRAME: ISA DATE
                           (MONTH MAY)
                           (DAY 28]
TO-PLACE = SAN-DIEGO in (TRIP TO ?)     ... filling in the requested information
TRAVEL-DATE = (MAY 28) in (TRIP TO SAN-DIEGO)   ... and the volunteered
                                                    information
dowhen TO-PLACE is put in (TRIP TO SAN-DIEGO)   ... propagating information to
                                                    other slots
  (LINK TRIP-SPECIFICATION FOREIGN-PORT CITY)
```

Figure 3.4: GUS reasoning from utterance (from [5])

From the example, we can see how GUS process with input utterance and it is clear that the system treats the key words into categories and use framework to match them.

There are other frame-driven systems such as VOYAGER [6] and TOSBURG [54], basic consider of these systems are in similar way but the later ones like VOYAGER are using improved frames that can make a better performance on understanding the contents

and generation of knowledge structures.

The frame-driven systems can handle understanding the contents by using categories of words and frameworks, but still it is hard to understand speaker's intentions because there are no reasoning of what speaker wants, the system can only get information from speaker's locutionary acts. In addition, this kind of system can only generate give back answer based on pre-definition and understand contents based on the pre-defined categories, for this reason their generalization capability on other tasks is not good.

3.3 SLU in Statistic based model

One way of applying statistic language model is using it on speech recognition, this way makes the dialogue system become able to handle continuous and open-domain original acoustic input, this make the system be able to interact with speakers better than before [55,56]. In the CU communicator [55], the statistic language model is used for speech recognition task, but the understanding for dialogue contents is still realized by frame-based method.

In cui1's system, it is an information recommendation dialogue system, it has similar condition to the CU system, method of understanding to dialogue contents is not applied with statistical language model but by using frame-based method.

Another way to apply statistic language model is to make an intention understanding algorithm by recognizing dialogue act or speaker's needs. Based on statistic algorithm such as HMM, a corpus-based intention identification method has been proposed [10], in this study, the frequency of sequence and phrase from given utterance is used for making statistic model. For each sequence or phrase, there is a category label, and for each utterance there is an intention label. The algorithm can make recognition for utterance based on sequence and phrase label. Here is an example of processing:

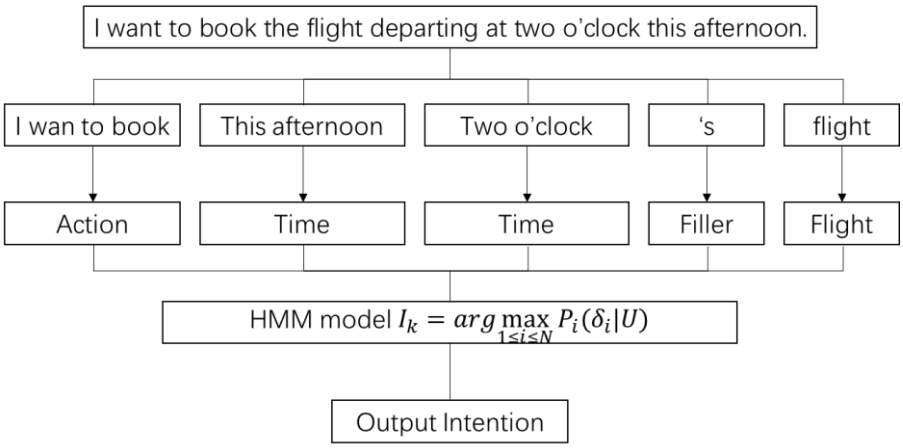

Figure 3.5: Example of processing from [10]

In which $I_k, P_i, \delta_{intention}, U$ means output k-th intention, probability of i-th intention, the given utterance is i-th intention and the given utterance, respectively. This model achieved good performance on intention recognition both on close test data and open test data, in addition, this model also improved response rate for dialogue in a speech recognition rate of 78%

Intention	Booking	Inquiry	Greeting	Ending	Filler
Close test	93.6	92.5	90.9	97.3	82.2
Open test	90.6	84.0	82.6	91.2	80.0

Table 3.1: Intention identification results (from [10])

Some other studies also use this kind of method to recognize speaker's illocutionary act, such as recognition of engagement in an advice-given dialogue system [57] and automatic tagging for dialogue[58].

From the property of statistical model, it can be considered that the method based on statistical method relies on training corpus very much and not robust in many cases. One reason could be the probability relationship in given utterance is totally relying on training data, another reason can be considered as the representation of linguistic feature is weak to generalize to wider domain.

3.4 SLU in Neural network based model

With the development and application of neural network (DNN), the dialogue system can handle more tasks. Based on big data and neural network structure, some end-to-end models have been proposed [11,59,60,61] and achieved good performance for both the understanding of contents and understanding of intentions [18, 62]. It has been showed by comparing two kinds of models [9, 63], on success rate, naturalness and comprehension aspects, the DNN model have better performance than statistic model on end-to-end dialogue system on both textual interaction and speech interaction.

Metric	BASE-DS [9]	E2E-DS [63]
Success	3.34 (1.71)	4.72 (0.75)
Naturalness	3.02 (1.23)	4.16 (0.93)
Comprehension	2.86 (1.40)	4.38 (0.98)
Average # of turns	6.51	4.79
# dialogues	50	50

Table 3.2: Comparison result of textual interaction (from [63])

Metric	BASE-DS [9]	E2E-DS [63]
Success	3.17 (1.73)	4.40 (0.98)
Naturalness	3.18 (1.30)	4.04 (0.97)
Comprehension	2.98 (1.71)	4.10 (1.15)
Average # of turns	6.96	5.25
# dialogues	75	75

Table 3.3: Comparison result of speech interaction (from [63])

In the studies of understanding of contents and generation for answers, a sequence-to-sequence (Seq2Seq) structure is well-used [7]. In such structure, input utterance will be encoded with an encoder which usually be an RNN-based neural network, after input completed, the output sequence can be generate by a decoder which also usually be an RNN-based neural network. During training this end-to-end model, it is not necessary to process input sequence such as analyzing the structure of sequence or semantic meanings of words, the representation for probability in a tensor space will be learned after training. For this reason, it is a very convenient way to make a good-performance dialogue system.

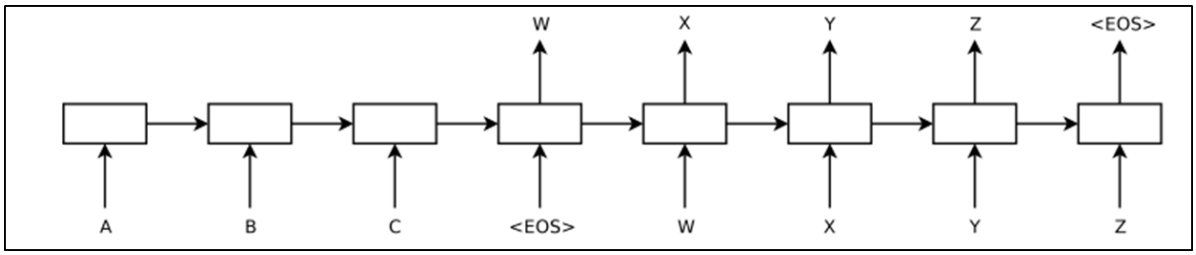

Figure 3.6: Basic thought of sequence to sequence (from [7])

Based on Seq2Seq structure, some dialogue system models have been proposed and been showed be better performance than statistical ones [8].

In one of these studies [64], a hierarchical structure of Seq2Seq dialogue system has been proposed. The processing of such system is similar to the basic Seq2Seq structure. Firstly, the given utterance will be encoded by RNN-based network into an utterance-level representation, then before decoding and generate answer, when input sequence has been encoded, the hidden state of encoder will be used as an input to a context encoder which can capture the temporal context information on whole set of sequences. Finally, the encoded context representation will be an input to the decoder, and the decoding process will be done as the basic Seq2Seq structure.

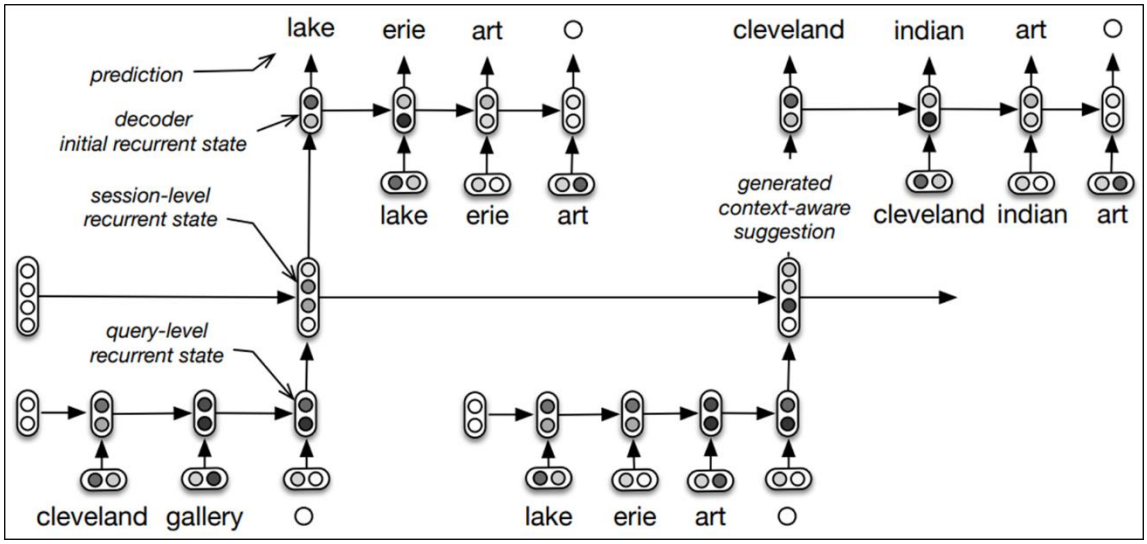

Figure 3.7: Hierarchical Seq2Seq structure (from [8])

By applying Seq2Seq and other DNN structure, the performance of generating answers in dialogue system has been improved, it is a good approach to the understanding of content. However, in the understanding of intention, not the only the information in one sentence, but the in discourse level are important for understanding. As the dialogue succeeding, the information in early sentence cannot be carried well to the later turn and if there are errors of detection of knowledges, it will be propagated to later turn and influent the performance of system.

Some studies tried to make approaches to solve such problem. A recurrent convolutional neural network (RCNN) was proposed for both sentence and discourse level dialogue act tagging [65]. In this study, the sentence level representation is made by a hierarchical CNN (HCNN) with multiple 1-dimension kernels, then the sentence representation is used to make the discourse representation by an RNN model.

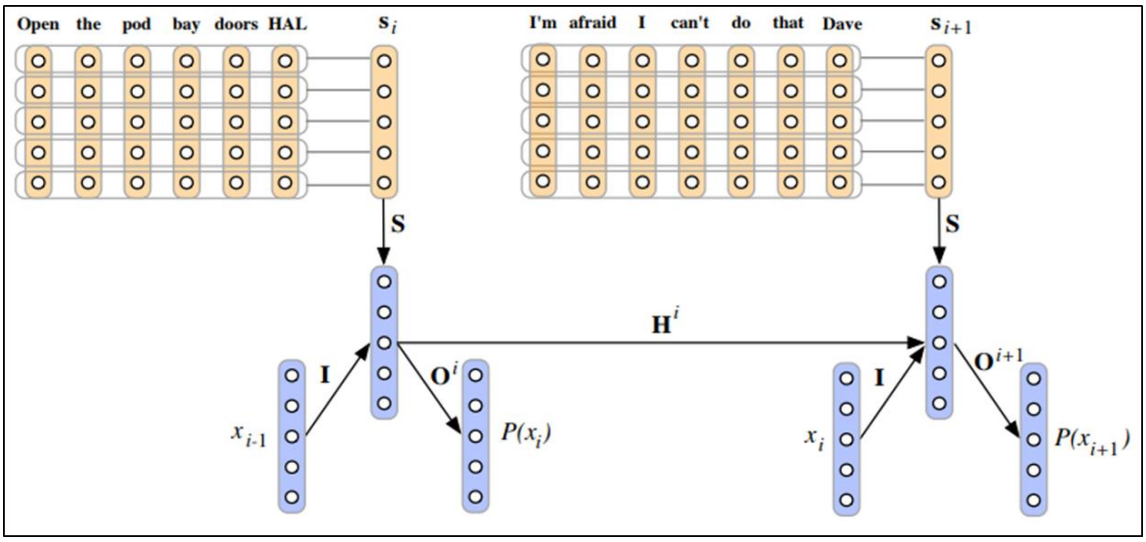

Figure 3.8: RCNN structure (from [65])

The result showed that this model has better performance (accuracy 73.9%) on dialogue act tagging than LH-HMM trigram (accuracy 71.0%). Based on this study, an improvement structure has been proposed which use the long-short-term-memory to process discourse information.

In another study [11], an end-to-end memory network has been proposed for multi-turn spoken language understanding. In the processing of this model, the history utterances are encoded by a contextual sentence encoder, current utterance is processed into two ways, one is encoded into sentence representation and be used to the knowledge attention model as a memory representation, another one is used as a parameter to the RNN tagger at the end of whole network. The contextual representation is used to make the knowledge attention distribution with the current sentence encoded input. Then, combine the current encoded input and weighted memory representation, a knowledge encoding representation can be generalized, finally, using an RNN tagger to tag the knowledge sequence of current utterance by using knowledge encoding representation with on current utterance. This model processes history information and current information respectively and make a weight distribution on history knowledge such that the model can learn whether a kind of knowledge need to be carried or not, this makes the model can capture the essential parts of whole dialogue. While the other way is more common processing than the processing of history information, by encoding the utterance and tagging with the history information representation, the model can finally get the knowledge tagging on current utterance.

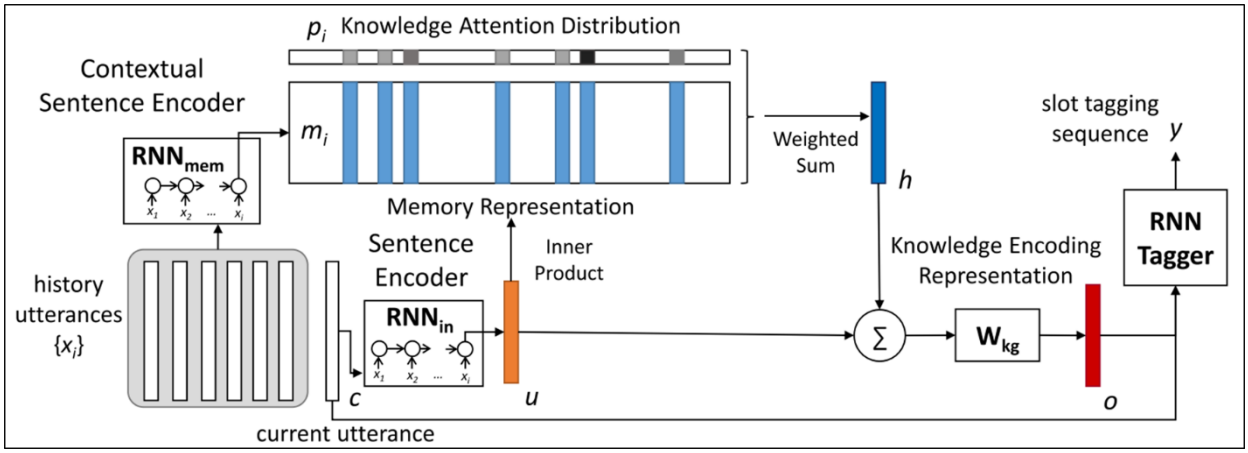

Figure 3.9: Memory network (from [11])

It is showed that this memory model can achieve a better F1 performance of intention recognition and slot on both single-turn and multi-turn level than using RNN-tagger only without history information and normal encoder-tagger with history information.

To make the performance of neural network better, it is considerable to combine DNN with other models, such as statistic model or pre-defined rules. A hierarchical LSTM with CRF model has been proposed for dialogue act recognition [66]. In this study, a bi-directional LSTM (Bi-LSTM) is used for encoding the input utterances into tensor representation, and CRF is used for dialogue act recognition by using Bi-LSTM encoded utterance representation. The result showed this model is better than other models on SwDA and MRDA dataset. This performance can be considered as the capability of generalization of Bi-LSTM for temporal sensitive input, such as input dialogue sentence by sentence, and the stabilization statistic property of CRF can be a good way to use to capture contextual relationship through a dialogue.

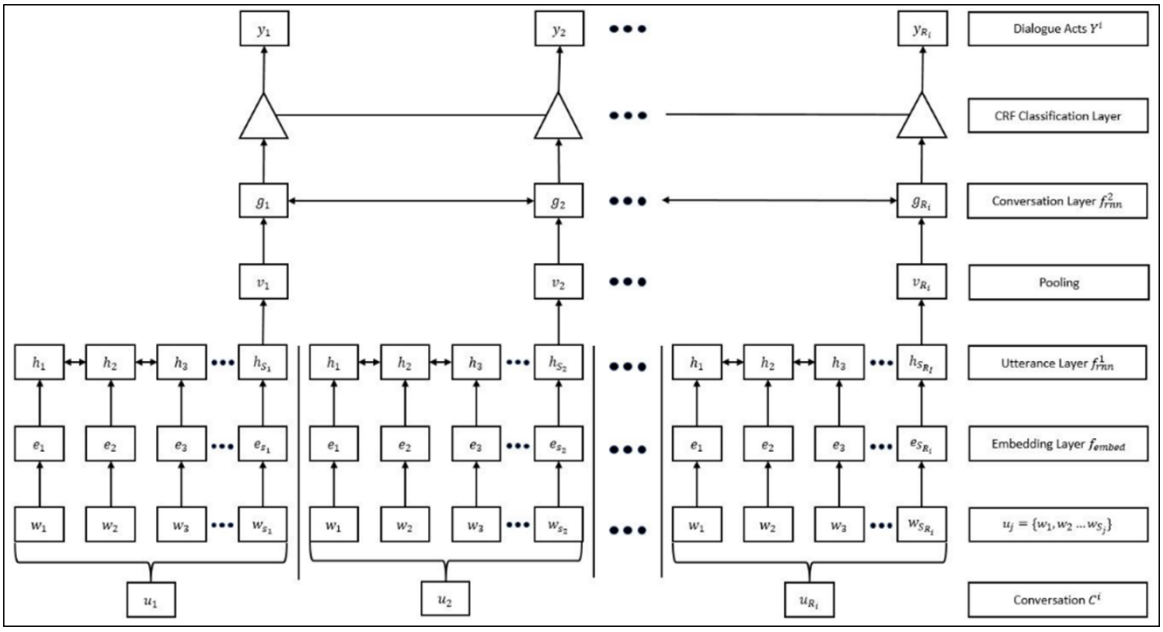

Figure 3.10: Bi-LSTM-CRF structure (from [66])

There is a new way to make understanding of intentions without recognizing the dialogue acts in dialogue but design a DNN structure let the network learn the representation of intentions in the tensor space. For this kind of task, a latent intention dialogue model has been proposed [68].

This model mainly consisted with three parts, the representation construction part, policy network part and generator part, it is an end-to-end model that with an input utterance the system can give back an answer that considered the intention of speaker. After input an utterance, this model can generate several candidate answers and pick one with the probability representation of intention, this model can be trained in semi-supervision method or reinforcement method. This model is evaluated by BLEU score and success, comprehension, naturalness by human subjects.

The result showed that it has a good performance on each aspects of evaluation and if

training with reinforcement learning, this model can get a better performance than using normal training method.

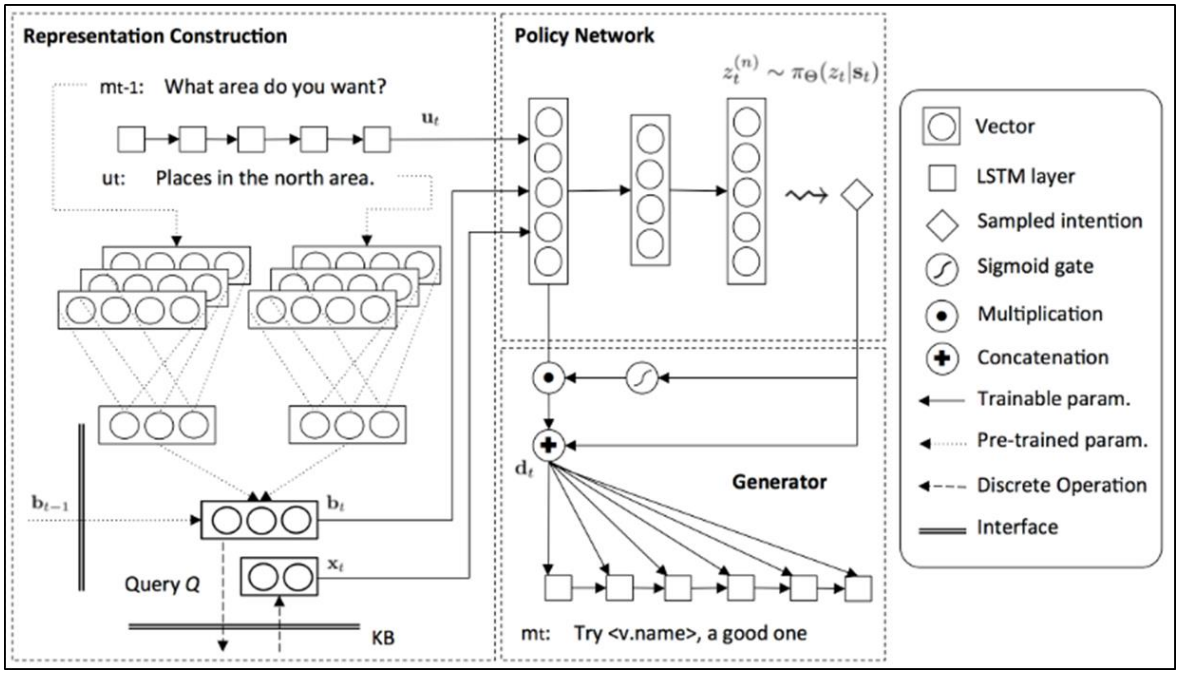

Figure 3.11: Latent intention dialogue model structure (from [68])

Chapter 4

Perception of research

4.1 Remaining problems

From the survey above, we can know that up to now, in SLU task, the computer can understand the contents in a DNN based representation and can recognize dialogue act for a high accuracy for given dataset. This means that the computer can understand human's locutionary act in a mathematical way and illocutionary act in a category-based representation. However, there are still some problems remained in SLU research:

Firstly, because of the black-box property of DNN models, the understanding of locutionary act cannot be explained very well. Moreover, the understanding of illocutionary act can only be done on a discrete level which is by dialogue act. These make it hard to explain how the linguistic or paralinguistic features influent the recognition of locutionary and illocutionary acts.

Secondly, although some paralinguistic features are well-used in speech recognition and synthesis such as prosody cues, they are still not well-used for the SLU task in dialogue. As mentioned above, it is showed that the paralinguistic feature has high relationship with recognizing the intentions of speaker, but how to use prosody on SLU and how to model a SLU model based on prosody are still not clear. In addition, some features in NLP tasks are also used well in SLU, such as semantic roles of words

Then, on the contextual level, some models can carry knowledge through the dialogue and got good performance, but this kind of models are relied on learning and the property of RNN or other DNN structure, the relationship representation based on analytical way of sentences through a dialogue has not been explained, but this kind of representation can be an approach to the understanding of how speech act conveyed and changed through the dialogue, in other words, is to understanding the mechanism of management of human communication with each other.

Finally, although there are some approaches and attempt to model the locutionary without using labels, such as the latent intention dialogue model. But it is still not clear how to use the mathematical representation to describe intentions or illocutionary act. It can be considered that if make a mathematical representation of illocutionary act firstly and then use other method to make the model could be an approach to get close with our final goal, an explainable model of understanding of spoken language.

4.2 Perception of research

To model the illocutionary act to represent intentions, it is necessary to use both linguistic and paralinguistic features, especially make flexible use of features such as prosody and semantic roles. For this reason, it can be considered that finding relationship between some not flexible used features and intentions can be a first step for our goal. The next step can be finding a method to represent the intentions in a mathematical way rather than labels of dialogue act, with such method, how important of each component of sentence for the SLU task can be researched by using these representations.

Chapter 5
Conclusion

In this report, we have surveyed the development of spoken language understanding in dialogue system. The survey is done by two parts, the methods to do SLU tasks and the SLU in dialogue system.

From this survey we can know in recent years, as the development of neural network, the performance of both understanding the speech act such as contents and understanding the intentions are getting better and better than before, it means that the computer can have a better performance on understanding both locutionary acts and illocutionary acts of speaker.

However, there are still some problems remaining, in these problems there are problems about unknown of computation method for useful cues in SLU, such as how to use prosody or emotion cues for understanding the intentions, and problems about modeling the computational model for illocutionary acts into mathematical representation.

In our feature research in doctor course, we would like to try to focus on modeling for illocutionary acts on contextual level, and to find out some relationship of sentences through dialogue based on linguistic and paralinguistic information.

Appendix

To make flexible use for paralinguistic feature, we made an experiment to automatically detect the prosody boundary on Mandarin conversation dataset.

Prosody boundary is defined as the interval between two syllables, it contains four categories. The previous studies used CART or DNN method to do such tasks with selected syntactic features as linguistic feature and acoustic features as paralinguistic feature and got an accuracy for 78.25% on reading dataset.

In Mandarin Chinese, a syllable is a character. We used character vector that can represent the relationship in character level. Other features are IS09E feature set and temporal features of syllables as paralinguistic feature. Then we train a model to detect prosody boundary automatically on conversation dialogue dataset and using a hierarchical DNN structure.

The result showed that our model can get higher mean accuracy on the conversation dataset than the accuracy on reading dataset. But only by such result cannot explain which feature is important to use, next work we will try to find relationship between such features and SLU tasks based on some pre-trained models.

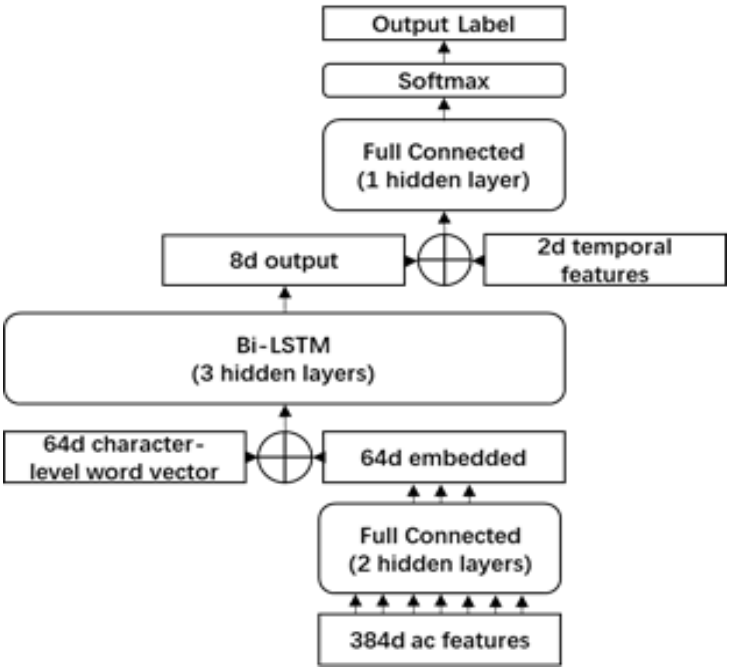

Figure 6.1: DNN structure in our experiment

Model	Ni	Lin	Ours	
Correct Rate	78.25%	77.34%	Non-D	Differ
			77.71%	83.66%

Table 6.1: Results of prosody boundary detection

(Non-D: apply character-level directly, Differ: apply with utilizing difference of adjacent syllables)

Model/Label	0	1	2	3	4
TN-DNN (Ni)	96.1%	21.2%	44.1%	83.6%	75.1%
SY-CART (Lin)	90.9%	48.6%	50.9%	80.8%	61.7%
Ours-D	91.9%	70.5%	66.0%	20.0%	98.0%

Table 6.2: Results of prosody boundary detection in each label

Bibliography

[1] NTS,進化するヒトと機械の音声コミュニケーション, 2015

[2] Searle J R, Searle J R. Speech acts: An essay in the philosophy of language[M]. Cambridge university press, 1969.

[3] Searle J R. Indirect speech acts[J]. 1975.

[4] Weizenbaum, J. (1966). ELIZA---a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36–45. https://doi.org/10.1145/365153.365168

[5] Bobrow, G., Kaplan, R. M., Kay, M., & Winograd, T. (1977). GUS , A Frame-Driven Dialogue System, 155–173.

[6] Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., & Seneff, S. (1990). The VOYAGER Speech Understanding System: Preliminary Development and Evaluation, 0–3.

[7] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks, 1–9. https://doi.org/10.1007/s10107-014-0839-0

[8] Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2015). Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models, 3776–3783. https://doi.org/10.1017/CBO9781107415324.004

[9] Ultes S, Barahona L M R, Su P H, et al. Pydial: A multi-domain statistical dialogue system toolkit[J]. Proceedings of ACL 2017, System Demonstrations, 2017: 73-78.

[10] Wu C H, Yan G L, Lin C L. Spoken dialogue system using corpus-based hidden Markov model[C]//Fifth International Conference on Spoken Language Processing. 1998.

[11] Chen, Y., & Hakkani-t, D. (2016). End-to-End Memory Networks with Knowledge Carryover for Multi-Turn Spoken Language Understanding, 3245–3249.

[12] Lamel, L. F., & Gauvain, J. L. (1995). Phone-based approach to non-linguistic speech feature identification. Computer Speech and Language, 9(1), 87–103. https://doi.org/10.1006/csla.1995.0005

[13] 達也河原, 宏彰川嶋, 高嗣平山, & 隆司松山. (2008). 対話を通じてユーザの意図・興味を探り情報検索・提示する情報コンシェルジュ. 情報処理, 49(8), 912–918.

[14] 橋本秀樹, 坪井宏之, 竹林洋一. 実時間音声対話システム TOSBURG の開発 (2) 音声理解[J]. 全国大会講演論文集, 1992 (人工知能及び認知科学): 155-156.

[15] Ortony A, Clore G L, Collins A. The cognitive structure of emotions[M]. Cambridge university press, 1990.
[16] Manning C D, Manning C D, Schütze H. Foundations of statistical natural language processing[M]. MIT press, 1999.
[17] Hofmann T. Probabilistic latent semantic analysis[C]//Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1999: 289-296.
[18] Jurafsky D, Martin J H. Speech and language processing[M]. London: Pearson, 2014.
[19] Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing[J]. Computational linguistics, 1996, 22(1): 39-71.
[20] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.
[21] Joachims T. Learning to classify text using support vector machines: Methods, theory and algorithms[M]. Norwell: Kluwer Academic Publishers, 2002.
[22] Baker J K. Stochastic modeling for automatic speech understanding[C]//Readings in speech recognition. Morgan Kaufmann Publishers Inc., 1990: 297-307.
[23] Stolcke A. SRILM-an extensible language modeling toolkit[C]//Seventh international conference on spoken language processing. 2002.
[24] Heafield K. KenLM: Faster and smaller language model queries[C]//Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2011: 187-197.
[25] Heafield K, Pouzyrevsky I, Clark J H, et al. Scalable modified Kneser-Ney language model estimation[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2013, 2: 690-696.
[26] Salton G. The SMART retrieval system-experiments in automatic document processing[J]. Englewood Cliffs, 1971.
[27] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[28] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
[29] Cao S, Lu W, Zhou J, et al. cw2vec: Learning Chinese Word Embeddings with Stroke n-gram Information[J]. 2018.
[30] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/arXiv:1811.03600v2

[31] Bucciarelli M, Colle L, Bara B G. How children comprehend speech acts and communicative gestures[J]. Journal of Pragmatics, 2003, 35(2): 207-241.

[32] Enrici I, Adenzato M, Cappa S, et al. Intention processing in communication: a common brain network for language and gestures[J]. Journal of Cognitive Neuroscience, 2011, 23(9): 2415-2431.

[33] Fridlund A J. Human facial expression: An evolutionary view[M]. Academic Press, 2014.

[34] Frith C. Role of facial expressions in social interactions[J]. Philosophical Transactions of the Royal Society B: Biological Sciences, 2009, 364(1535): 3453-3458.

[35] Parkinson B. Do facial movements express emotions or communicate motives?[J]. Personality and Social Psychology Review, 2005, 9(4): 278-311.

[36] Hellbernd N, Sammler D. Prosody conveys speaker's intentions: Acoustic cues for speech act perception[J]. Journal of Memory and Language, 2016, 88: 70-86.

[37] Fujisaki H. Prosody, models, and spontaneous speech[M]//Computing prosody. Springer, New York, NY, 1997: 27-42.

[38] Tamburini F. Prosodic prominence detection in speech[C]//Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on. IEEE, 2003, 1: 385-388.

[39] Ning, Y., Jia, J., Wu, Z., Li, R., An, Y., Wang, Y., & Meng, H. (2017). Multi-Task Deep Learning for User Intention Understanding in Speech Interaction Systems. Proceedings of the 31th Conference on Artificial Intelligence (AAAI 2017), 161–167.

[40] Boersma P, Van Heuven V. Speak and unSpeak with PRAAT[J]. Glot International, 2001, 5(9-10): 341-347.

[41] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor[C]//Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010: 1459-1462.

[42] Schuller B, Steidl S, Batliner A. The interspeech 2009 emotion challenge[C]//Tenth Annual Conference of the International Speech Communication Association. 2009.

[43] Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism[C]//Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France. 2013.

[44] Arsikere H, Sen A, Prathosh A P, et al. Novel acoustic features for automatic dialog-act tagging[C]//Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016: 6105-6109.

[45] Gu, Y., Chen, S., & Marsic, I. (2018). Deep Multimodal learning for emotion recognition in spoken language. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2018–April, 5079–5083.
[46] Bolinger D, Bolinger D L M. Intonation and its parts: Melody in spoken English[M]. Stanford University Press, 1986.
[47] Elections before democracy: the History of Elections in Europe and Latin America[M]. Springer, 2016.
[48] 赵建军, 杨晓虹, 杨玉芳, 等. 汉语中焦点与重音的对应关系——基于语料库的初步研究[J]. 语言研究, 2012, 32(4): 55-59.
[49] 史德明. 论现代汉语中焦点与信息, 话题及重音之间的关系[J]. 现代语文: 下旬. 语言研究, 2017 (6): 62-64.
[50] Fujie, S., Ejiri, Y., Nakajima, K., Matsusaka, Y., & Kobayashi, T. (2004). A conversation robot using head gesture recognition as para-linguistic information. RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759), 159–164. https://doi.org/10.1109/ROMAN.2004.1374748
[51] 長澤史記, 石原卓弥, 岡田将吾, 等. ユーザーの態度推定に基づき適応的なインタビューを行うロボット対話システムの開発[C]//人工知能学会全国大会論文集 2017 年度人工知能学会全国大会 (第 31 回) 論文集. 一般社団法人 人工知能学会, 2017: 2H4OS35b1-2H4OS35b1.
[52] Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings[C]//Advances in Neural Information Processing Systems. 2016: 4349-4357.
[53] Minsky M. A framework for representing knowledge[J]. 1974.
[54] 竹林洋一. 音声自由対話システム TOSBURG< cd02d36. gif>—ユーザ中心のマルチモーダルインタフェースの実現に向けて—[J]. 電子情報通信学会論文誌 D, 1994, 77(8): 1417-1428.
[55] Pellom B, Ward W, Pradhan S. The CU Communicator: an architecture for dialogue systems[C]//Sixth International Conference on Spoken Language Processing. 2000.
[56] 翠輝久, 河原達也, 正司哲朗, 等. 質問応答・情報推薦機能を備えた音声による情報案内システム[J]. 情報処理学会論文誌, 2007, 48(12): 3602-3611.
[57] Novielli, N. (2010). HMM modeling of user engagement in advice-giving dialogues. Journal on Multimodal User Interfaces, 3(1), 131–140.
[58] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., … Meteer, M. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. Computational Linguistics, 26(3), 339–373.
[59] Zhou, J., & Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks. Proceedings of the 53rd Annual Meeting of the

Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 1127–1137.

[60] Liu, B., Tur, G., Hakkani-Tur, D., Shah, P., & Heck, L. (2018). Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems.

[61] Hori C, Hori T. End-to-end conversation modeling track in dstc6[J]. arXiv preprint arXiv:1706.07440, 2017.

[62] 河原達也. 音声対話システムの進化と淘汰: 歴史と最近の技術動向 (< 特集 > 音声対話システムの実用化に向けて)[J]. 人工知能学会誌, 2013, 28(1): 45-51.

[63] Braunschweiler, N., & Papangelis, A. (2018). Comparison of an end-to-end trainable dialogue system with a modular statistical dialogue system. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2018–Septe(September), 576–580.

[64] Sordoni A, Bengio Y, Vahabi H, et al. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015: 553-562.

[65] Kalchbrenner N, Blunsom P. Recurrent convolutional neural networks for discourse compositionality[J]. arXiv preprint arXiv:1306.3584, 2013.

[66] Kumar H, Agarwal A, Dasgupta R, et al. Dialogue act sequence labeling using hierarchical encoder with crf[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[67] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging[C]//Conference on Empirical Methods in Natural Language Processing. 1996.

[68] Wen T H, Miao Y, Blunsom P, et al. Latent intention dialogue models[J]. arXiv preprint arXiv:1705.10229, 2017.

Research achievement list

1. Sixia Li, Jianwu Dang, “Automatic Mandarin Prosody Boundary Detection in dialogue with character-level word vector”, 日本音響学会 2019 春季研究発表会, 東京, 2019.3
2. Yuning Liu, Sixia Li, Jianwu Dang, “MFCC を用いた会話のストレス検出”, 電子情報通信学会 2019 年総合大会, 東京, 2019.3

Acknowledgement

I would like to express my special thanks to teaching and guidance from professor Dang, and my thanks for opinions and suggestions from my lab mates. And appreciate for the teaching from professor Akagi and Unoki at summer campaign and thanks for correct suggestions from professor Dang, professor Akagi, professor Unoki and associate professor Yositaka at my first attempt of examination of entering doctor course. Moreover, not only on academic aspects, but also on correcting the direction and thought of myself, I would like to express my deep appreciation for professor Dang's guidance.