A Domain Analytic Method in Modular-Designed Dialogue System: Application to a System for Japanese

Motoki Yatsu, Rafal Rzepka, and Kenji Araki[1]

Abstract. In this paper, we propose implicit and explicit utterance generation models and a dialogue system in which such models are implemented. Modularization of classifiers enables the agent to annotate input utterance with tags of multiple features including types of sentences and mood expressions. In the implicit model, the features extracted from the input sentences define an agent's internal state. A relativity vector to each domain is sustainably computed based on similarity in Japanese WordNet ontology and the system's internal state. The explicit answers are generated if the input text is classified as Question-Answering domain based on the tags given by classifier modules. In other cases of classification, the system generates open-domain utterances. We will discuss the result of experiments intended to show characteristics of both domain detection methods.

1 Introduction

Task-oriented interfaces for mobile devices is widely accepted as information providing agents. However, a preliminary survey showed that in several domains of utterance users have good command if an utterance is recognized not only in an 'explicit' manner but also 'implicit'. This result has motivated us to reconsider what the true design of reflexive agent means in the research area of Artificial General Intelligence (AGI).

As an introduction, this section will provide a view on the current conceptual structure of a dialogue recognition and utterance recognition methods widely used across Natural Language Processing and Human Language Technology areas.

A preliminary survey conducted on the web[2] in Japan showed that some internet users do not accept the intention of being explicit in several domains a part of daily conversations belong. This result shows that humans do not, or cannot converse making their demands or need for information only explicit in some dialogue domains.

In the first half of this paper, we review current development in utterance recognition, and propose a model which explains an important and feasible portion of present capabilities. In the latter part of the paper, we will discuss the results of a dialogue system that handles the both non-task-oriented and task-oriented utterances, based on the proposed model.

2 Current Defining of Key Concepts

Here we revise concepts developed in research on dialogic interaction, in both the linguistic observation of conventional human-human dialogues and human-computer interaction.

2.1 Models in Human-Computer Interaction: Task-Oriented vs. Non-Task-Oriented

Various research has been conducted on methods to simulate the capability to generate utterances and perform a dialogue with human, i.e. to make correct responses which satisfy intention of human user, through generated utterances. A dialogue consists of several pairs of utterances, traditionally considered as performed alternately between the two participants of a dialogue.

In such attempts, many researchers have proposed a model that distinguishes dialogues or utterances which intend to complete a specified task shared by dialogue participants. Many have utilized a task-oriented dialogue model [1]. Task-oriented domain has a relevance to knowledge based on the Kintsch and van Dijk model [14]. In the view of this model, an utterance which belongs to a task-oriented domain has also an orientation within a domain of conversational topics (topic domain).

Many systems that show some performance in resolving tasks that we can regard as relevant to a specific domain, we can also be regarded as restricting the coverage of the domain of a dialogue performed between a user and the system. This is because any utterance such a system classifies out of the specified topic domain(s) is rejected telling the user that the utterance is unrecognizable or irrelevant to the desired task.

If dialogue system design involves this classification, dialogues or utterances that the system does not classify in any task domain are marked as non-task-oriented. Non-task-oriented utterances form a domain equal to a set of utterances with the ones from the task-oriented domain excluded.

Directly following this observation, we considered a dialogue system as a state machine using non-task-oriented and task-oriented states in our previous work [16]. The system proposed in this work initiates dialogue in the non-task-oriented state where chat modules work to produce a non-task-oriented dialogue. The system is given several task domains which need to satisfy the condition of relevance, which we discuss and compare later, to generate an utterance based on the domain. The relevance condition is a threshold of similarity between the input utterance and keywords specific to the task domain.

2.2 Relevance Theory and Human-Computer Interaction

According to the Relevance Theory [15] (RT), intentions that we infer from a dialogue include the intention of the transmitter to show

[1] Graduate School of Information Science and Technology, Hokkaido University, email: {my,kabura,araki}@media.eng.hokudai.ac.jp

[2] Conducted on http://q.hatena.ne.jp/1338376413, directed to internet users older than 20. 130 people took the survey.

some information about a fact, and the intention of the same person to share the information with the hearer. The theory rejects observation in which syntactical representation of an utterance involves its meaning, but accepts the assertion that the communicator intends to elicit among the participants two intentions: that there is information to be elicited (as an informative intention; which we refer as **II**), and the fact that he/she has such an intention (as a communicative intention; **CI**). Each participant of the dialogue preforms ostensive-inferential communication holding the two kinds of intentions.

Referring to the observation of RT, the hearer makes a deduction of such two intentions, from an utterance in any domain. The present domain shared by participants limits the resulting meaning of deduction made after the ostensive-inferential communication.

To help make an II into mutualized knowledge, which can fail due to noise interference and knowledge deficiency of the hearer, the communicator may generate stimulus which consists of encoded CI. The hearer notices the mediacy of the II by the communicator by decoding the message.

2.3 Application of Both Models to Dialogue Agent

In the field of Natural Language Processing, we can estimate that presumption of information by decoding the encoded CI is more precise than deduction from the surface of the utterance. The hearer's lack of background knowledge and any ambiguity in surface features of utterance can lead to a failure of deduction.

Many ways to show informative or communicative intention in everyday conversation people use acoustic methods are usually limited to textual input to each application software. Therefore, if using a natural language as medium, text usually spoken dialogue is a rather restrictive method.

There are several works about discourse analysis [10] [12] from the point of view of computational linguistics focusing on the Relevance Theory. Stone [12] claims that intention is a mental representation with a complex structure. In this paper, we rethink his position and define intention as the main objective of the system given by the developer of the system.

One feasible example of limiting the requirement of the source intention, which comes from the mind of the participant, is limiting the intention into not completely humanlike motives that occur to sustain one's life, but an intention for the agent to only serve the user. We can consider that this limitation is successful for the following reason. The system should have an initial primitive motivator (agenda) [3] that always motivates the agent to act upon the environment. The primitive motivation in this case to filter and classify user's input to known domains grippes the system's attention, so that the system assists user's decision to select a task and let the system do the needed task. This classification is achievable in a cognition model [4], which is thus compatible with a design that separates and modularizes filtering and reactive functions to the perceptual input.

In this model, the primitive motivator perpetually motivates and fires the attention of the dialogue agent. However, the motivation must have an essential account of the agent's capabilities (tasks broadly explained) and their objectives.

In this paper spoken dialogue interface which can run on electrical devices like smartphones. The problems the users are likely to have are that they cannot make full use of functionality of the device due to lack of procedural knowledge of manipulation, which can form a strong aspiration for an interface which is more easy-to-use.

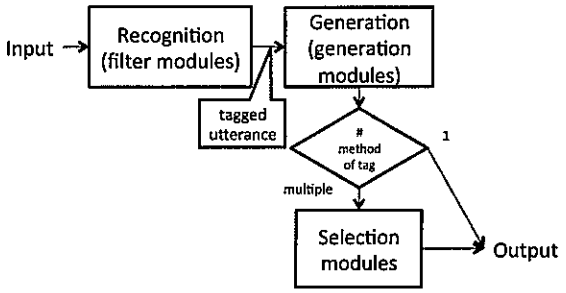

Figure 1. Schematic representation of the system.

3 Definition of Terms

In this paper, we are using a vocabulary that points to the current consensus of keywords which describe participant's behavior during a dialogue:

- Domain decision
  Attaching a domain tag to an utterance or an entire dialogue with methods discussed below.
- Open-domain (*or* non-task-oriented domain)
  A domain that does not belong to any specified domain and does not have a textual label.
- Task-oriented domain
  A domain which relates to a specific task that is expected to be achieved.

4 Methodology

In the previous sections, we have proposed an ostensive-informational model of utterance recognition and generation, which we consider to also be suitable for explaining the language division of the human cognitive system. In our view, it is possible with an all-by-simulation approach to investigate the model in question, where humans classify received utterances into non-task-domain and task-domains related to the knowledge and initial intention of the system. Fig. 1 shows a schematic representation of the system.

4.1 Implementation of Dialogue System

We have created an experimental dialogue agent. The system breaks down into two capabilities: capabilities of recognition and generation of utterance, each separated, obtaining a modular design with which the system is constructed from submodules. Functionality of each part contains a number of submodules, namely each tagger and utterance generation functionality is modularized separately.

4.2 Utterance Recognition

In this part of the system (filtering part), modules annotate its own tag to the text input. Recognizing an utterance means to annotate a tag to its text input, and tags annotated on an utterance represents a domain. In this paper, we apple a single domain tag which relates to the question answering task. Tags annotated in the filter module influence utterance generation module.

A filter module may annotate multiple tags. At the end of utterance processing, the sum of scores for all the tags is calculated and ranked.

The tag which obtains the highest score is selected and the system chooses an utterance generation method based on the selected tag.

We have also considered of another design in which a combination of tags works as data to decide the final domain of the utterance, where the machine-learning engine of the system uses the combination as learning data.

Tags, domains and functionality of those modules are listed in Table 1.

Table 1. The tags and generation (G), filter (F) modules used in the system.

Type	Name	Tag	Function/Generated Utterance
G	*Maru*	NT	Open-domain ut. based on N-gram [13]
G	*Moda*	NT	Open-domain ut. with modal expression [5]
G	*Eliza*	NT	Open-domain ut. based on scenario [8]
G	*QAC4*	QA	Answer to open-domain question [7]
G	*Recom*	RU	Domain selection inquiry
G	*Task n*	Dn	Task-oriented response of keyword a_n
Type	Score	Tag	Function
F	1.5	NT	Choose a non-task-oriented utterance
F	2.0	QA	Detect question-answering utterance
F	2.0	RU,Dn	Find a task relevance of user's utterances

Table 2. The utterance selection modules used in the system.

Type	Name	Scoring method
S	*ChatLog*	Maximum reciprocal edit distance among sentences in IRC chatlog
S	*ChatLogAbst*	*ChatLog* using POS-abstracted chatlog
S	*NGramHitnum*	Relevance of words based on frequency of coocurrence in the Web

We can name a tag with a string of ASCII characters: the name does not depend on a given structure. Though the diversity of classification of utterance types is essentially broader than the single task, in this paper we will discuss the effect of a single utterance-type classification.

5 Functionality

Here we describe the range of utterances produced by sub-modules in the generation module.

5.1 Question Answering

The task requires [7] an agent to detect a question type from utterances intended to be in other domains, and retrieve information needed to answer the question, which needs a filter's support based on WWW knowledge. Though we can classify target of this system module into a simple open-domain utterances, this method of implementation is suitable for testing the general dialogue capability (response to utterances with an explicit communication intention).

5.2 Generation of Open-Domain Utterance

The system currently uses one of three methods to generate an open-domain utterance we above listed above as type 'G' in Table 1. When creating an open-domain response, one of of the outputs of the 3 methods is selected by selection modules in Table 2.

5.3 Domain Selection Inquiry

As we showed using a survey result in ??, the communicative or informative intention should be communicated with a form of implication. Moreover, a user's goals are included in both non-task-oriented and task-oriented domains. We vectorize averages of similarity expressed with a distance and a graphed thesaurus (Japanese WordNet [2]) between content morphemes and keywords. The filter module uses the similarity vector to select the domain keyword set most relevant to the current history of dialogue from the vector norm. Table 8 shows the sampled dialogues between the system and the user, which mainly consist of QA utterances.

5.3.1 Target Domains and Keyword Set

We chose 5 target domains, listed below. A keyword is a Japanese general noun that is bound to a domain and a task-oriented utterance generation module. Keywords are treated in a set and can express a centroid of meanings in multiple words.

5.3.2 Aim Vector $\boldsymbol{a}$

In the aim vector $\boldsymbol{a} = (a_1, a_2.a_3, ...)$ the current *aim* of the user's dialogue is calculated. Here, each element $a_1, a_2, ...$ represents average similarity between all of the content morphemes from the user's utterance and a keyword which exists as a concept in Japanese WordNet.

5.3.3 Similarity in WordNet

$\boldsymbol{a}$ is the cumulated sum of $\Delta\boldsymbol{a}$, which represents semantic similarity measured by Leacock-Chodorow [9]:

$$\Delta a_i = \frac{1}{N_U N_{K_i}} \sum_{u \in U} \sum_{k \in K_i} sim(u, k) \tag{1}$$

$$where\ sim(c_1, c_2) = max\left(-\log\frac{N_p}{2D}\right), \tag{2}$$

and N_U stands for the number of content morphemes in user utterance, N_{K_i} number of words in domain keyword set a_i , N_p the graph distance between c_1, c_2 in an ontology, with D as taxonomic depth in the ontology.

5.3.4 Inquiry Utterance Generation

The system acquires a norm of aim vector $\boldsymbol{a}$, $\|\boldsymbol{a}\| = \sqrt{\sum_k a_i^2}$. In a dialogue turn when $\|\boldsymbol{a}\|$ exceeds the threshold T, the system understands the user's interest is high in a task domain enough to receive a recommendation utterance. We chose a value of $T = 2.0$ in the experiment discussed later. A filter module outputs a tag 'RU' with score 3.0. The utterance generation module generates an utterance which helps the user decide the task, using k_i as the most relevant keyword which has the maximum value in $\boldsymbol{a}$.

6 Evaluation Experiments

Here we mention the experiments we performed for getting overall ratings by the users, precision of the filtering part to select an utterance generation method, and evaluation of implicit intention detection that helps user's task selection.

6.1 Questionnaire Evaluation Results

We chose Question-Answering agent [7] implemented as a dialogue system, and an ELIZA-type dialogue system [8] as baselines for this measure.

6.1.1 Questionnaire Survey

10 participants (9 male, 1 female) were requested to perform a dialogue with the system which lasts more than 20 turns. The system to evaluate was implemented as a CGI web application[3]. A questionnaire with an answering form was displayted after 20 turns elapsed, in which the participants were asked to evaluate the system using 5-point scales from 1 (I disagree) to 5 (I agree) towards these 6 statements:
(A) I would like to continue the dialogue.
(B) The dialogue is natural in grammar.
(C) The dialogue is natural in its sense.
(D) The system's vocabulary is rich.
(E) The system talked like a human.
(F) The system's recommendation had a strong relevance with my concern and interests.

Table 3. Overall ratings in 5-point scales.

Item	average of ratings
A	1.75
B	1.75
C	1.63
D	2.00
E	2.13
F	1.75
Mean	1.85
Baseline	2.54

6.1.2 Analysis of Impression Using Semantic Differential Method

We conducted an evaluation experiment intended to investigate the orientation of participants' impression toward the experimental system. The same group of participants in 6.1.1 evaluated the system using 35 pairs of adjectives in a 5-point linear scale, which are frequently used [6] in semantic differential method [11] and represent positiveness of subjective impression a participant has held to the system. The adjective pairs are listed in Table 6.

6.2 Measurement of Appropriateness of Utterance Method Selection

To measure the effectiveness of the explicit response generation, we asked another group of participants to score how precisely the system selected an utterance method. Participants choose from one of 4 grades to rank each utterance in a dialogue log (which had 150 utterances combined from the logs taken during the first experiment) the system made as response. Participants were asked to select one value from 0∼3, 3 for acceptable or appropriate utterances, 2 for grammatically correct but unacceptable, 1 for unrecognizable ones, and 0 for error outputs.

[3] http://arakilab.media.eng.hokudai.ac.jp/~my/experiment.html (in Japanese)

6.2.1 Participants

Participants were 4 male graduate students with ages of around 24. The target of this evaluation is the actual dialogue made by the participants in 6.1.1 and 6.1.2. This evaluation was also performed online.

6.2.2 Result

Table 4 shows the result of the evaluation. The mean of all scores was 2.274, with standard deviation of 0.847.

Table 4. Distribution of evaluation scores of the precision experiment.

3	2	1	0 (*errors*)
302	176	104	17
50.3%	29.3%	17.3%	2.83%

7 Discussion

The average score obtained from the experiment was by 10 and fewer participants, a number which the authors do not consider as statistically sufficient for a complete judgement of this system's potential. However, in this paper we use these results to illustrate the system's characteristics.

7.1 User's Impression and Precision of Utterance Selection

The average results of precision of utterance generation shown in Table 5 for each tag exhibit higher precision in utterance generation with tags correctly annotated.

Table 5. Utterance method selection appropriateness of each tag. (* indicates that 1 participant evaluated the item)

Tag	Average value
NT	2.32
QA	2.26
The system	2.28
Question-Answering only*	1.85
Eliza-type system only*	2.45

7.2 Evaluation of Domain Selection

The partial score of the overall rating (F) involves appropriateness of domain selection inquiries made by the module. Looking at the data, we found that when there was no selection of utterance domain in the dialogue, the value decreased.

8 Analysis of SD Evaluation Scores

We conducted factor analysis on the data obtained from the experiment (6.1.2) using GNU R environment[4]. The number of factors was 4, with their cumulated contribution rate 0.700. The factors gained from this are shown in Table 7. Further, Fig. 8 shows a dendrogram resulting from the application of cluster analysis in the furthest neighbor method to the acquired data, regarding it as a vector with 35 dimensions. When cut at score 6.0, four general clusters appear to divide the vector data.

Table 6. Adjective pairs of with contrast of positiveness. (Positive adjectives in the right column.)

1	dark	bright
2	cold	warm
3	weak	strong
4	dismal	cheerful
5	light	heavy
6	hate	love
7	hard	soft
8	passive	aggressive
9	noisy	quiet
10	inactive	active
11	bad	good
12	unkind	kind
13	violent	peaceful
14	painful	fun
15	sober	flashy
16	boring	interesting
17	dull	sharp
18	bad feeling	good feeling
19	unreliable	reliable
20	feeble	robust
21	small	large
22	flippant	serious
23	slow	fast
24	unpleasant	pleasant
25	unstable	stable
26	taciturn	talkative
27	dirty	clean
28	sloppy	neat
29	simple	complex
30	static	dynamic
31	stubborn	frank
32	irresponsible	responsible
33	sad	happy

Table 7. Result of factor analysis applied to the SD method evaluation.

Cumulated CR	Adjective pairs ID	+/-
0.231	6,9,17,19,22,24,28,31	+
0.444	2,10,13,20,26,30	-
0.580	11,12,23,32	-
0.700	7,33	+

Thus we extracted 4 factors from the factor analysis. Among these factors, the third group including "bad-good", "unkind-kind", "slow-fast", and "irresponsible-responsible" also appear in the same the dendrogram of cluster analysis. These observations represent a factor of functionality of the system. This factor explains the cause of lower

[4] http://www.r-project.org/

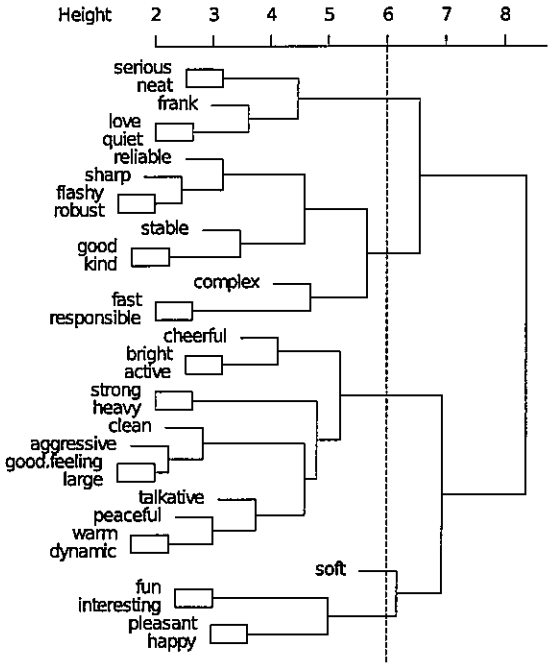

Figure 2. Dendrogram gained by cluster analysis on the SD data.

score of participant's overall ratings of impression compared to the baseline systems, as being the working speed of the system. The average of time the system took to give a response in all the dialogue made in the experiments was 33.1 seconds, against 0.163 seconds by the baseline (ELIZA-type) system.

9 Conclusion

In this paper, we discussed a method for a general-purpose modularized dialogue agent to make a suggestion based on implicit comprehension from the user's content words, as well as utterance method selection in response to explicit requests. We showed that the system that has limited primitive motivator could perform a general-purpose dialogue with minimum difference in human-rating score compared to non-modular baseline systems.

The domain keywords which were used for domain selection inquiry were limited in number and were chosen in an arbitrary manner, which resulted in high rates of incorrect domain decisions. The system should decide a domain dynamically and recommendation of task must be done with general knowledge. Combining Web-based knowledge and commonsense reasoning would suffice for the system to approach the user's need, which may need more computation resources to perform faster.

Finally, as the results of the questionnaire (see 6.1.1) suggest we can view the system's user-friendliness and grammatical, contextual acceptance as substandard (< 2.00) and yet to be improved. In order to make the utterance method selection more precise and appropriate, we are designing a method, with each filter module annotating multiple tags used as cues to form a decision process in general domains.