Unsupervised Approach for Dialogue Act Classification

Kiyonori Ohtake

National Institute of Information and Communications Technology (NICT)/
Advanced Telecommunications Research Institute International (ATR)
2-2-2 Hikaridai, Keihanna Science City 619-0288, JAPAN
kiyonori.ohtake (at) nict go.jp

Abstract. This paper presents an unsupervised approach for dialogue act (DA) classification. We used a latent variable model to compress the dimensions of the feature vector. We introduced a paraphraser to reduce the variety of expressions and to solve the pragmatic problem for DA classification. The paraphraser seemed to work well on some DA classifications in the unsupervised approach. The results obtained by the unsupervised approach were compared with the manually annotated labels. A preliminary experiment for semi-supervised tagging was also carried out, and we discuss these results.

Keywords: Unsupervised, Dialogue Act, Paraphrasing, Latent Variable Model.

1. Introduction

Recognizing the intentions of a user in a dialogue system is very important. So far, many methods have been developed to infer a user's intention in a dialog situation. To infer the user's intention in an utterance, the utterance can be categorized into given classes. Therefore, many studies have designed the classes called dialogue act (DA) labels that approximate a speaker's intention. They annotated the labels on a corpus to analyze the phenomena for DA interaction or to develop a DA tagger in order to infer the DA label from a speech segment (e.g. some utterances, an utterance, or a part of an utterance). Most studies on DA taggers were based on a supervised method (e.g., (Stolcke et al., 2000; Tanaka and Yokoo, 1999)). The labels used in a DA tagger have to be predefined, and supervised methods require a corpus that is manually annotated by the labels.

On the other hand, it is difficult to design a tag set (labels) that can be used to annotate a corpus because the design of a tag set depends on the domain and the task. Therefore, we have to redesign the tag set and construct a corpus annotated with a new tag set if we apply our system to different domains or tasks. In addition, designing a tag set that can be used in any domain or task is very difficult. However, we have to annotate DA tags on a corpus, because many applications require predefined DA tags.

This paper discusses an unsupervised approach to infer the user's intention in a situation by using a dialog system. Unsupervised approach may not achieve highly accurate results when compared to the supervised approach. However, in any domain or task, the unsupervised approach can yield human DA annotators with machine judgments of the DA classification that may be useful to keep the consistency of DA annotation results for a corpus.

In addition, annotating a corpus with given labels is very time-consuming. An unsupervised method is independent of annotation and designing the tag set. In order to achieve an unsupervised method, we need an unsupervised clustering method. So far, many clustering methods have been proposed and discussed for applications in natural language processing (NLP), such as works by Zhao and Karypis (Zhao and Karypis, 2005). However, an utterance is very short against a document that is used in a common NLP application. In addition, the feature space that is used to express any natural language expression is extremely large and an utterance is expressed by a very sparse vector in the feature space. Therefore, it is very important to handle a sparse feature vector of an utterance in the huge feature space.

2. DA Annotation

Here, we construct a dialogue system to make an itinerary of one-day sightseeing tour and also develop a dialog corpus for this system. The corpus consists of 100 dialogues between a professional tour guide and a tourist. Each dialog is almost 30-min long. An annotated corpus with DA is needed to construct our dialogue system. Therefore, we have started to design a DA tag set and annotate the DA tags on the corpus.

However, there are several problems that make it difficult for us to maintain consistency in the annotation as follows: (a) segmentation, (b) pragmatics, and (c) multifunctionality.

Sometimes, utterances are fragmental, and it is difficult to recognize an appropriate boundary of an utterance for a DA tag. Hinarejos et al. reported that the correct segmentation for DA is very important for obtaining an accurate result in DA tagging (Hinarejos et al., 2006).

There is a pragmatic problem in the annotation of DA tags. For example, the utterance "Do you know what time it is?" can be recognized as a yes/no **question** from the surface information, but the speaker's intention is a **request** such as "Please tell me the time."

In addition, utterances are generally multifunctional. This problem is closely related to the design of the DA tag set. So far, many DA tag sets have been proposed and used to annotate corpora. Some of them have several layers (e.g., DAMSL (Allen and Core, 1997)) and dimensions.

In this paper, we focus on the pragmatic problem in the DA annotation. We try to resolve the pragmatics problem by paraphrasing. If a euphemism is paraphrased into a straightforward expression, the dialogue system can easily understand the expression.

3. Unsupervised method for DA annotation

In this section, we describe an unsupervised approach to classify an utterance. The overview of the unsupervised approach is as follows:

1. Construct a feature vector from an utterance.
2. Reduce the dimensions of the feature space using a latent variable model.
3. Classify the vector whose dimension was reduced using an unsupervised classification method.

After constructing the feature vector, we use a latent variable model to reduce the dimension of the feature space. Then, we use an unsupervised classification method to classify the vector that produced by using a latent variable model. Finally, we find the class to which the utterance belongs.

We also introduce a rule-based paraphraser to reduce the variety of expressions because a different expression is treated to be completely different in a latent variable model.

3.1. Latent variable models

Several unsupervised text modeling methods, such as PLSI (probabilistic latent semantic indexing (Hofmann, 1999)) and LDA (latent Dirichlet allocation (Blei et al., 2003)), are available to model a text based on the features of words and their frequencies. In general, the latent variables indicate the topics of each segment (some sentences for text or some utterances for speech), and we can use the topic information indicated by the latent variables of the model as a compact surrogate expression for a given feature vector of an utterance. In other words, we can use these models to reduce the dimension of the feature space. Once the model parameters are learned from a corpus, we can infer the topic of a given utterance. If we constructed a latent variable model with k latent variables, we get a k-dimensional vector. This vector is called a topic vector.

We used PLSI—a latent variable model— for general co-occurrence data that associates an unobserved topic variable $z \in Z = \{z_1, \cdots, z_k\}$ with each observation, i.e. with each occurrence of word $w \in W = \{w_1, \cdots, w_M\}$ in document $d \in D = \{d_1, \cdots, d_N\}$.

The probability of a topic under the document ($P(z \mid d)$) is approximated by the following formula:

$$P(z)^2 \prod_{w \in d} P(w \mid z) \sum_{w} n(d, w) P(w \mid z), \quad (1)$$

where $n(d, w)$ indicates the frequency of word w in the document d. The details about how to introduce Equation (1) have been previously shown (Ohtake, 2005). In that paper, Ohtake used PLSI and LDA to evaluate whether a paraphrasing pair is contextually independent or not, as well as if there was not a big difference in the performances between them. Therefore, we use PLSI because it is simpler and faster than LDA.

3.2. Unsupervised clustering method

There are several unsupervised clustering methods. We used the K-means clustering algorithm (e.g., (Duda et al., 2000)) that is very simple because, at the moment, a highly sophisticated method in which analyzing the tendency of the results by an unsupervised approach and manually annotated labels is not necessary. In addition, we have to investigate whether a topic vector reasonably expresses a DA before using a sophisticated clustering method.

3.3. Paraphrasing to reduce variety of expressions

The use of a wide variety of expressions that conveys the same information is natural. However, a different expression is treated to be completely different in a feature space. Therefore, paraphrasing techniques seem to be promising approaches to understand the variety of expressions. In particular, in Japanese, the ending of a sentence or utterance has many expressions even though they convey the same meaning. These expressions are related to the Japanese honorific system, and in most cases the difference in the expression does not affect the DA classification.

We construct a rule-based paraphraser that is very similar to the paraphraser proposed by Ohtake and Yamamoto (Ohtake and Yamamoto, 2001), and most of the rules in the honorific system were derived from their paraphraser. The paraphraser was carefully designed to be free from errors and developed to paraphrase a variety of expressions that convey the same meaning into a standard expression.

The rules of the paraphraser are based on a morphological analysis. We can use regular expressions for pattern matching in a rule and we can conjugate any morphemes that have conjugation to fit in its context. Therefore, a small number of rules cover a large number of targets that need to be paraphrased.

4. Experiments

In this section, we describe our experiments and introduce the data set. We also mention the features that were used in the construction of the PLSI models.

4.1. Data set

We used the ATR Dialogue Database (Morimoto et al., 1994). This database consists of 1,983 dialogues (83,052 utterances) in traveling situations. We used manually transcribed Japanese texts in the database. In the transcribed texts, fillers and disfluencies are tagged with a marker. In order to use precisely analyzed results, we eliminated the fillers and disfluencies in the transcribed texts by a morphological analyzer that was used to obtain morphemes as units like words.

We annotated 13 dialogues (489 utterances) with DA tags used in the paper by Tanaka and Yokoo (Tanaka and Yokoo, 1999) to evaluate the unsupervised classification. The remainder of the data, namely 1,970 dialogues (82,563 utterances), were used to estimate the parameters of the PLSI model.

The original DA tag set that consisted of 26 tags was designed to annotate the dialogue segments that were shorter than an utterance. Therefore, there were multi-labeled utterances in our annotation results because in some cases, a person utters several things in a single utterance. For example, when a person is asked a YES or NO question (YN-QUESTION), the person who answers might say "Yes, I will…(YES, INFORM)." In this case, we treated the last DA tag as the labeled tag of the utterance. In the annotated dialogues, 16 tags were actually used.

4.2. Features for PLSI

We used uni-gram and bi-gram word frequencies. In this paper, a word is considered as a morpheme[1] in Japanese. An element of the feature consists of a pair of morpheme's basic form and POS (part of speech). However, numbers and proper names are generalized by eliminating this basic form. In other words, the features of the numbers and proper names are recognized by only by their POS.

In general, PLSI requires words and their frequencies in order to construct a model from a corpus. However, Serafin et al. showed that adding extra features works well with latent semantic analysis in the DA classification (Serafin et al., 2004). The PLSI model can be regarded as a probabilistic version of a latent semantic analysis. Therefore, we can expect the same effect on PLSI, and we introduced the uni-gram and bi-gram features.

The segment for a unit of a document consists of the utterance and its previous utterance. The dialogues in the database are conversations between two people such as a customer and a clerk.

4.3. Number of variables and performance on differentiation

We constructed PLSI models[2] on the number of latent variables, namely 10, 50, 100, 200, and 300, in order to determine the number of latent variables. The parameter for tempered EM (TEM)—a technique used to ease the over-fitting problem—was set to 0.9 (we use this value in all of the experiments in this study) because this value exhibited the best performance in the preliminary experiments.

We formulated topic vectors from the evaluation dialogue set, and we prepared the average vectors for each DA label from these topic vectors. Finally, we compared each average vector with the others according to their cosine values, and we averaged the cosine values. Therefore, these numbers indicate the distinguishing ability of topic vectors, where a smaller number is better. The average values for each number of latent variables (10, 50, 100, 200, and 300) with all the DA labels are as follows: 0.607, 0.334, 0.288, 0.290 and 0.275, respectively.

4.4. Impact of paraphrasing

We applied the rule-based paraphraser to the data set (83,052 utterances), and all of the 56,027 utterances were paraphrased.

First, we show the result of an unsupervised clustering result with manually annotated labels using a non-paraphrased corpus. We constructed the PLSI model with 100 latent variables from the learning corpus that was not paraphrased. The test set was fed to the PLSI model, yielding the topic vectors. Then, we used the K-means clustering method with 16 clusters because the size of the tag set that was used to annotate the test set is 16. The result is shown in the "without paraphraser" column of Table 1.

Second, we show the result of an unsupervised clustering result with manually annotated labels using a paraphrased corpus. The result is shown in the "with paraphraser" column of

[1] We used a morphological analyzer available at `http://mecab.sourceforge.net/`

[2] We used the package available at `http://chasen.org/~taku/software/plsi/`

Table 1. Note that the cluster IDs found in the columns, “with paraphraser” and “without paraphraser” are independent of each other.

Table 1: Unsupervised clustering result with/without paraphrasing and manual labels

	without paraphraser	with paraphraser
manual labels (freq.)	(cluster ID: its frequency)	(cluster ID: its frequency)
ACK (68)	B:7, G:24, H:2, I:1, M:3, N:7, O:8, P:16	a:12, c:14, f:2, g:4, k:3, n:8, p:25
ACT-REQ (44)	B:4, C:3, D:1, F:1, H:6, I:1, J:1, K:1, M:11, N:10, O:5	b:1, c:1, d:1, f:2, g:2, i:7, j:25, k:4, l:1
ALERT (1)	N:1	o:1
APOLOGY (2)	H:1, N:1	n:1, o:1
CONF-Q(29)	B:4, C:1, D:1, H:1, I:3, J:4, K:1, L:1, M:3, N:8, O:2	c:2, d:1, e:6, f:1, g:2, i:2, n:5, o:10
FAREWELL (16)	K:8, M:3, N:5	g:1, i:2, k:2, n:1, o:10
G-WISHES (1)	K:1	o:1
GREET (8)	B:3, M:1, N:4	g:3, o:5
INFORM (198)	B:27, C:11, D:10, E:6, F:2, G:2, H:35, I:7, J:12, K:7, L:14, M:26, N:27, O:12	a:11, b:2, c:1, d:4, e:6, f:16, g:28, h:13, i:10, j:3, k:16, m:18, n:5, o:63, p:2
PERM-REQ (1)	E:1	g:1
SUGGEST (6)	F:3, H:2, M:1	c:2, f:2, m:1, o:1
THANK (20)	A:16, I:2, K:1, N:1	a:1, g:2, i:2, l:10, n:4, o:1
THANK-RES (2)	K:2	o:2
WH-Q (40)	C:1, D:2, F:16, H:4, I:2, K:7, L:2,	a:1, c:19, f:6, g:1, i:10, k:2, n:1
YES (18)	B:6, G:2, H:1, I:5, O:4	a:6, g:9, o:1, p:2
YN-Q (35)	B:2, C:4, F:4, H:6, I:1, M:8, N:7, O:3	a:2, b:4, c:12, f:4, g:2, h:1, i:4, j:1, n:1, o:4

4.5. Semi-supervised approach—preliminary experiment

We carried out a very small experiment for the semi-supervised approach. The experiment is small because the amount of annotated data is very small. We have only 13 annotated dialogues. We used 12 dialogues to construct the average vectors for each label, where a withheld dialogue (32 utterances) was used as the test data.

The method to classify an utterance is very simple. From a learning set, we construct the average vectors for each label. Then, an utterance is given to construct a topic vector using PLSI with 100 latent variables and the average vector closest to the topic vector is calculated. Finally, the label of the average vector is inferred from the utterance’s classification. The accuracies of the results are 37.5% (12/32) without paraphrasing and 21.9% (7/32) with paraphrasing.

5. Discussion

When using the latent variable model, the number of latent variables is an issue. In our experiment, there was not a considerable difference between the result using 100 latent variables and the results using more than 100 latent variables; therefore, 100 latent variables seem sufficient for our experiment.

We compared the unsupervised approach and manually annotated labels. It is difficult to conclude whether the unsupervised approach works well or not. There were some cases in which the label and cluster have a strong correlation. For example, the label “THANK” and cluster ID **A** and the label “WH-Q” and ID **F** in the “without paraphraser” column of Table 1 indicate very good cases. On the other hand, the original label “INFORM” indicated a miscellaneous category, and there were so many utterances labeled “INFORM.” Thus, utterances labeled “INFORM” were classified into many clusters.

We paraphrased our data set to reduce the variety of expressions. From Table 1, we find a very clear tendency in the result of the label “ACT-REQ (ACTION-REQUEST)” that was used to label utterances asking someone to perform a certain task. In Japanese, there is a large variety of expressions to this end. Without paraphrasing, these expressions are treated differently. On the contrary, we treated them as the same expression when they were paraphrased into a single expression. Therefore, paraphrasing works quite well on utterances labeled “ACT-REQ.”

We have to consider the number of variables in a latent variable model and the number of clusters in an unsupervised clustering method. In our experiment, the number of cluster used was the same as the number of labels that were used in the learning corpus. However, if we used more clusters, we might be able to classify a large cluster into proper sub-clusters.

On the other hand, there were some clusters having many elements that correspond to many manually labeled tags. For example, cluster ID **N** in the “without paraphraser” column of Table 1 was related to many labels. From the observation of the test set, the phrase “*onegai shimasu* (please)” seems to be strongly related to this cluster. This phrase is frequently used in Japanese when requesting someone to perform a particular task. Ten utterances labeled “ACT-REQ” were classified in this cluster. However, this phrase is too common to use as a feature. Meanwhile, cluster ID **o** in the “with paraphraser” column of Table 1 was also related to many labels. In these cases, the expressions of number seem to be related to this cluster. We have to consider what feature is effective for DA classification.

We carried out a very small preliminary experiment using a semi-supervised approach. The size of the learning data for the semi-supervised method was too small to evaluate the method. In addition, the accuracies were quite low—37.5% without paraphrasing and 21.9% with paraphrasing. Contrary to our expectations, the result with paraphrasing was worse than that without paraphrasing. The observation results suggested several points. First, some labels did not match their utterances after paraphrasing. The expressions used in the utterances were drastically changed by the paraphraser and the annotated labels had become inappropriate for the paraphrased utterances. Thus, we should control such paraphrasing. When we re-annotated the paraphrased test set, the accuracy increased from 21.9% to 31.3%. Second, paraphrasing caused a side effect. Reducing the variety of expressions constricted the features used by the PLSI. The paraphraser was not designed for DA classification. Some phrases should not be paraphrased and we should retain the original expressions.

6. Conclusion

This paper discussed an unsupervised approach for DA classification using a rule-based paraphraser and a latent semantic model. In the experiments, a PLSI model with 100 latent variables was found to be efficient with respect to its distinguishing ability. At the moment, on the other hand, we are unsure whether the unsupervised approach is promising when comparing the results obtained by the unsupervised approach with the manually labeled results.

The introduction of a paraphraser that reduces the variety of expressions showed good results. In particular, in Japanese, there are many euphemisms for asking someone to perform a particular task. The paraphraser paraphrased such expressions effectively.

Several points remain for our future work as follows:

- A further analysis of the classification results would be useful. In particular, we have to investigate whether the compressed feature space produced by PLSI is really effective for DA classification or not.
- Introducing other features would be effective. We only used uni-gram and bi-gram morphemes. Introducing tri-gram morphemes or other features such as dependency relationships may be effective.
- Tuning the paraphraser is required. The paraphraser was not tuned for DA classification. The paraphraser was designed to be generic.

In addition, we will apply this unsupervised method to the corpus that is now under development for DA annotation.

References

Allen, James and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers. *Technical Report, Discourse Research Initiative.*

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research,* 3:993-1022.

Duda, Richard O., Peter E. Hart, and David G. Stork. 2000. Pattern Classification. A Wiley-Interscience Publication.

Hinarejos, Carlos D. Martínez, Ramón Granell, and José Miguel Benedí. 2006. Segmented and unsegmented dialogue-act annotation with statistical dialogue models. In *Proceedings of the COLING/ACL 2006*, pp. 563-570.

Hofmann, Thomas. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval,* pp. 50-57.

Morimoto, Tsuyoshi, N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi and Y. Yamazaki. 1994. A speech and language database for speech translation research. In *Proceedings of ICSLP '94*, pp. 1791-1794.

Ohtake, Kiyonori and Kazuhide Yamamoto. 2001. Paraphrasing honorifics. In *Workshop Proceedings of Automatic Paraphrasing: Theories and Applications (NLPRS2001 Post-conference Workshop)*, pp. 13-20.

Ohtake, Kiyonori. 2005. Evaluating contextual dependency of paraphrases using a latent variable model. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005) conjunct with IJCNLP 2005*, pp. 65-72.

Serafin, Riccardo and Brbara Di Eugenio. 2004. FLSA: Extending latent semantic analysis with features for dialogue act classification. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pp. 692-699.

Stolcke, Andreas, K. ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339-373.

Tanaka, Hideki and Akio Yokoo. 1999. An efficient statistical speech act type tagging system for speech translation systems. In *Proceedings of the Thirty Seventh Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pp. 381-388.

Zhao, Ying and George Karypis. 2005. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10: 141-168.