Dialogue act classification using a Bayesian approach[*]

Sergio Grau (1), Emilio Sanchis (1), María José Castro (1), and David Vilar (2)

(1) Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camino de Vera s/n. 46022. València (Spain)
{sgrau,esanchis,mcastro}@dsic.upv.es

(2) Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen University
D-52056 Aachen (Germany)
vilar@informatik.rwth-aachen.de

Abstract

In this work, we make a contribution to natural speech dialogue act detection. We focus our attention on the dialogue act classification using a Bayesian approach. Our classifier is tested on two corpora, the Switchboard and the Basurde tasks. A combination of a naive Bayes classifier and n-grams is used. The impact of different smoothing methods (Laplace and Witten Bell) and n-grams in classification are studied.

With respect to the Switchboard corpus, an accuracy of 66% is achieved using a uniform naive Bayes classifier, 3-grams and Laplace smoothing to avoid zero probabilities. For the Basurde corpus, our system achieves performances similar to other methodologies we have previously tested. Through a combination of a naive Bayes classifier with 2-grams and Witten Bell smoothing we achieve the best accuracy of 89%. These results show that a Bayesian approach is well suited for these tasks.

1 Introduction

Dialogue systems constitute an outstanding objective in the field of language technologies. The structure of systems of this kind is usually the following: a signal goes through a speech recognizer; the recognized text is passed through a natural language understanding module which gives a semantic interpretation of the utterances; a dialogue manager takes a decision and the user receives the output of the answer generator through a synthesizer. Due to the fact that there are many error sources in these modules (recognition errors, misinterpretations, unexpected answers, etc), it could be useful to have a method to reliably detect which type of sentence has been uttered. This problem is called *dialogue act classification*. To carry out this classification, a combination of a naive Bayes classifier [1, 2] and n-grams. The impact of different smoothing methods (Laplace and Witten Bell) and n-grams on classification has been studied.

We focus on the process of the *user* dialogue act classification with regard to the type of dialogue act; i.e., the identification of the dialogue act type of the utterance pronounced by the user. Automatic dialogue act classification is useful in building specific models depending on the user dialogue act type or predicting the next user dialogue act. This information combined with confidence measures in each module can help the dialogue manager to take a more correct decision and make the human-computer communication more natural and fluent.

We have conducted all the experimentation using two corpora: the Switchboard corpus and the Basurde corpus. The Switchboard task [4] consists of human-to-human spontaneous telephone conversations. These conversations are about a general topic of interest between two people who are selected randomly and who do not know each other from before. An example appears in Figure 1.

The Basurde task [5] consists of information retrieval by telephone for nation wide Spanish trains. Queries are restricted to timetables, prices and services for long distance trains. An example of a dialogue is shown in Figure 2.

2 Dialogue structure

One of the most common ways to represent the dialogue structure is by using dialogue acts [6, 7], which represent the successive states of a dialogue. The labels must be specific enough to take all the different intentions of each turn into account, but, at the same time, they have to be general enough to be able to be adapted to different tasks.

Turn	Dialogue Act	Utterance
A	Statement (utt1)	The question was kind of interesting to me because I was just trying to put together a long term financial plan and monthly budget.
	Statement (utt2)	The only thing I do now is, put the data into Quicken.
	Yes_No_Question (utt3)	I don't know if you are familiar with that.
B	Yes Answers (utt1)	Yeah.
	Statement (utt2)	I have some friends of mine who use Quicken.
	Statement (utt3)	I've considered using it once myself.
	Statement (utt4)	But I decided that the amount of information that would have to go in would be a lot of time keeping that up to date.

Figure 1: A labeled dialogue from the Switchboard corpus.

Turn	Dialogue Act	Utterance
M	Opening	Bienvenido al sistema automático de información de trenes regionales y de largo recorrido, ¿qué desea? *(Welcome to the automatic system of regional and nation wide trains. How can I help?)*
U	Query	Hola buenos días, quería información de trenes a Lleida el día seis de noviembre. *(Hello, good morning, I would like information about trains going to Lleida on the 6th November.)*
M	Validation	A Lleida, ¿quiere viajar desde Zaragoza? *(To Lleida, do you want to travel from Zaragoza?)*
U	Affirmation	Sí, desde Zaragoza perdón. *(Yes, from Zaragoza, I'm sorry.)*

Figure 2: A labeled dialogue from the Basurde corpus. (The English translation is also given.)

The set of labels of the dialogue acts defined for the Switchboard corpus are defined in [7]. The corpus is labeled using the DAMSL standard [8], modified for the Switchboard task (42 labels correspond to the dialogue acts). This new labelling is called SWBD-DAMSL and has been defined in [9]. Each turn is composed of one or more utterances. In order to have a one-to-one correspondence between dialogue act labels and utterances, a manual process to segment the turns was performed (see, for example, that user **A**'s turn of the dialogue shown in Figure 1 is composed of three utterances). The Switchboard task is human-to-human conversation, and we need to classify the dialogue acts of the two people. The Switchboard data is composed of 1,155 labeled dialogues and of 173,153 utterances.

The example in Figure 1 shows first the user who is speaking (A or B), second, the type of dialogue act and the number of the utterances in the turn, and third, the utterance pronounced by the user. When averaged, a typical conversation has 144 turns, 271 utterances and took 28 minutes to label [7].

The Basurde task is composed of computer-to-human dialogues, and only the user dialogue acts need to be classified. The Basurde data is composed of 226 dialogues. The corpus is labeled by hand. The example of a dialogue in Figure 2 shows first, the type of turn (Machine or User), second, the dialogue act and third, an example of utterance. This corpus has 226 dialogues and 867 utterances. There are 15 labels for the dialogue acts, but only 10 classes of dialogue acts were frequent enough to

be considered for analysis. Each utterance has only one label. Table 1 shows the defined labels for the Basurde corpus.

The Basurde task has a vocabulary of 190 words after categorization and lemmatization. The categorization consisted of grouping classes of words into categories like CITY-NAME, TRAIN-TYPE, TRAIN-STATION, etc. The lemmatization consisted of putting the verbs in infinitive, transforming plurals to singular, etc.

3 Naive Bayes classifier

3.1 Dialogue act classification

Dialogue act classification is a special case of text classification where the text to be classified is the user utterance.

Text classification is done to find a function $f^*(\cdot)$ that maps a document d_i into a class c_i. The range is defined by the number of classes C.

The training of this function is done by learning from a set of samples with the form

$$\{(d_j, c_j)\}_{j=1}^{J}, d_j \in D, c_j \in C \qquad (1)$$

where d_j is the j-th sample, J the number of samples, and c_j its corresponding class label.

In our classification task, the samples are the utterances, and the attributes are the words of the utterances.

Table 1: Dialogue act types defined for the Basurde corpus and their frequencies.

Dialogue act	Example	#	%
Close	No, gracias. *(No, thanks.)*	208	23.99
Departure time	Quería saber el horario del jueves por la mañana. *(I would like to know the timetable for Thursday morning.)*	195	22.49
Affirmation	Sí. *(Yes.)*	146	16.84
Void	Jueves dos de enero. *(Thursday January the second.)*	135	15.57
Price	¿Me puede decir el precio? *(Could you tell me the price?)*	99	11.42
Arrival time	¿A qué hora llega a Zaragoza? *(At what time does it arrive to Zaragoza?)*	21	2.42
Departure time return	ara la vuelta, Sevilla-Zaragoza en Talgo por la mañana. *(For the return, Sevilla-Zaragoza on a Talgo in the morning.)*	19	2.19
Confirmation	¿Y cuál es el precio de una plaza en ese tren? *(And what is the price of a seat on this train?)*	15	1.73
Negation	No. *(No.)*	15	1.73
Train type	Quiero saber el tipo de tren. *(I would like to know the train type.)*	14	1.61

In our work, the function $f^*(\cdot)$ maps each user utterance to one of the $|C| = 42$ dialogue acts in the Switchboard task or to the $|C| = 10$ dialogue acts of the Basurde task. The decision of which class is assigned to a user utterance is made by the Bayes decision rule for minimizing the probability of error. The Bayesian classifier assigns the class with maximum a posteriori probability to the sample d:

$$f^*(d) = \underset{c \in C}{\operatorname{argmax}} \, Pr(c|d) \qquad (2)$$

3.2 Naive Bayes

In this work, we use the naive Bayes classifier in its multimodal event model. The representation of the utterances is a vector of word counts, which is usually called "bag of words". Some of these simple Bayesian classifiers have grown in popularity lately and it has been proven that, despite their simplicity, they give good results [2]. They use a set of labeled examples for training to estimate the parameters of the generative model. Classification of new examples is carried out by the Bayes decision rule through a selection of the class that has produced the largest probability.

The naive Bayes classifier assumes that all the attributes are independent of each other. This is what is called the "Naive Bayes assumption". Although this assumption is false in most real tasks, the naive Bayes classifier performs well in text classification tasks.

Text classification is a field with a large number of attributes. The examples attributes to be classified are the words, and the number of different words is really large. While some classification tasks can be solved with a vocabulary of a few hundred words, like the Basurde task, other tasks, like the Switchboard, are more complex with a vocabulary of thousands of words.

The naive Bayes algorithm is defined in this way:

$$c^\star = \underset{c}{\operatorname{argmax}} \Pr(c|d) \qquad (3)$$

$$= \underset{c}{\operatorname{argmax}} \frac{\Pr(c)\Pr(d|c)}{\Pr(d)} \qquad (4)$$

$$= \underset{c}{\operatorname{argmax}} \Pr(c)\Pr(d|c) \qquad (5)$$

$$= \underset{c}{\operatorname{argmax}} \Pr(c) \prod_{i=1}^{I} \Pr(w_i, w_{i-1}, .., w_{i-n+1}|c) \qquad (6)$$

where c is the class, $\Pr(d|c)$ is the conditional probability given the class c and $\Pr(c)$ is the probability of the class. In practice, we can estimate $\Pr(d|c)$ and $\Pr(c)$ in the previous equation with the training data. It is easy to estimate each $\Pr(c)$ by counting the frequency of each class in the training data. For $\Pr(d|c)$, the naive Bayes classifier is based on the conditional independence given the goal value. We pass from equation (3) to (4) by applying the Bayes rule. Going from equation (4) to (5) is valid because $\Pr(d)$ is independent of c.

In equation (6), $w_i, w_{i-1}, .., w_{i-n+1}$ is an n-gram. We have a naive Bayes classifier in its multinomial event model where its features are the n-grams. With this approach, the "Naive Bayes assumption" is false because we are considering that there are a relationship between the words. I is the vocabulary size of the task. In the case of 1-grams, the vocabulary size is the number of different words in the task. If $n > 1$ the vocabulary size is the number of different m-grams, where $1 <= m <= n$.

We performed the experimentation with another Bayesian classifier which do not take into account the a priori probability; i.e, the a priori probability for all the classes is uniform. The uniform naive Bayes classifier is formulated as:

$$c^\star = \underset{c}{\operatorname{argmax}} \prod_i \Pr(w_i|c) \qquad (7)$$

4 Experimentation

We present experimental results on different user dialogue act classifications in two corpora: the Switchboard task and the Basurde task. In the first task, we have to classify every turn because it is a human-to-human dialogue. In the second task, we have to classify only the user turns because it is a human-to-computer dialogue.

4.1 Cross-validation

To conduct the experimentation, we split the dataset into five sets. We used the cross-validation technique with 80% of the corpus (4 sets) for the training set and 20% of the corpus (1 set) in the test set. The basic idea is to split the corpus into N sets, distributing the $N - 1$ sets for the training and a unique partition for the test. This set is called hold-out. This proccess is repeated N times to use all the N sets as the hold-out set. The advantage of this method is that the N sets are used for the test, which leads to an efficient exploration of the dataset.

4.2 Naive Bayes classifier training

Naive Bayes classifier training was done with the statistical document classification software package "RAINBOW" [10]. We used the naive Bayes classifier and the uniform naive Bayes classifier. We also studied the influence of the n-grams and the stoplist for classification.

For the Switchboard task we also estimated classification by deleting stopwords from the utterances. The stopword list is the one from the "SMART" [11] information retrieval system. Stopwords are words like prepositions or articles that have little semantic information and that appear frequently in the corpus.

5 Results

Tables 2 and 3, and Figure 3 show the Switchboard result. The result obtained in the Johns Hopkins LVCSR Workshop-97 [7] was 53.9%, using 3-grams models for each of the 42 types of dialogue acts. With our Bayesian approach, we obtained 66% using a naive Bayes uniform classifier, with Laplace smoothing and 3-grams.

The Laplace smoothing technique obtains better results than the Witten Bell smoothing technique because the vocabulary size is large and this technique performs well with a large vocabulary.

The impact of the n-grams on the classification is shown by the fact that classification is better until it reached 3-grams and after that, despite the addition of more context, the naive Bayes classifier has worse results.

Another interesting fact is that the best results with the Laplace smoothing are obtained with a uniform naive Bayes, that do not take into account the a priori probability of the classes. This is because the distribution of
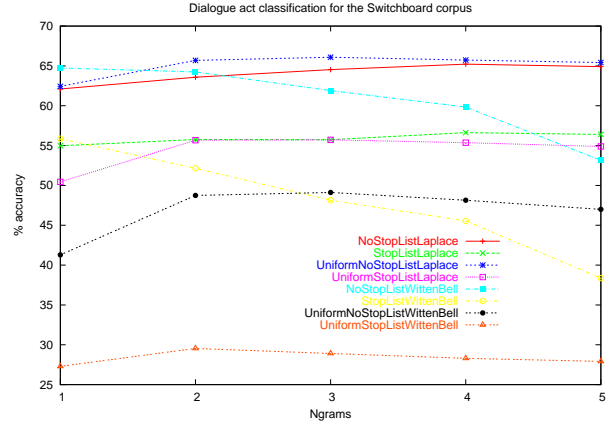

Figure 3: Dialogue act classification with the naive Bayes classifier: Switchboard corpus.

Table 2: Results for the Switchboard corpus without using a stoplist.

n	WittenBell	WittenBell Uniform	Laplace	Laplace Uniform
1	64.7	41.2	62.1	62.4
2	64.2	48.7	63.5	65.6
3	61.9	49.1	64.5	**66.0**
4	59.8	48.1	65.2	65.7
5	53.1	47.0	64.9	65.4

the data is not balanced between the classes which make that the more frequent classes in the corpus obtain better results than the less frequent ones.

In the Basurde corpus, we obtained 89.5% of dialogue act classification using a naive Bayes classifier with a Witten Bell smoothing and 2-grams (see Table 4 and Figure 4).

In the Basurde task, the Witten Bell smoothing method obtained the best result but we cannot conclude that the Witten Bell smoothing method performs better than the Laplace method in this task.

Dialogue act classes in the Basurde task were more balanced and for this reason the results of the uniform naive Bayes and the regular naive Bayes were similar.

6 Conclusions and Future Work

The results show that the Bayesian approach is well-suited for this task. The automatic detection of the user dialogue act class in a dialogue system and the detection of the dialogue structure are necessary to make specific models or to predict the next user utterance. Other approaches (Neural Networks, Bernoulli classifiers) have

Table 3: Results for the Switchboard corpus using a stoplist.

n	WittenBell	WittenBell Uniform	Laplace	Laplace Uniform
1	55.8	27.2	54.9	50.4
2	52.1	29.5	55.7	55.6
3	48.1	28.9	55.7	55.7
4	45.5	28.3	56.6	55.3
5	38.3	27.9	56.4	54.8

Table 4: Results for the Basurde corpus.

n	WittenBell	WittenBell Uniform	Laplace	Laplace Uniform
1	88.6	79.7	86.5	86.3
2	**89.5**	81.4	86.9	87.3
3	83.4	81.2	86.2	86.1
4	80.9	78.9	84.6	85.4
5	80.6	78.0	84.2	85.0

been studied [12] and provide comparable results.

An extension to this Bayesian approach will be studied in future work. We think that the potential of the Error Correcting Output Codes in similar tasks (text classification) should give good results in the automatic detection of dialogue acts.

References

[1] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[2] Jerome H. Friedman. On bias, variance, 0/1 - loss, and the curse-of-dimensionality. *Data Mining and Knowdlege Discovery*, 1:55–77, 1997.

[3] Fuchun Peng and Dale Schuurmans. Combining Naive Bayes and N-Gram Language Models for Text Classification. In *Advances in Information Retrieval: Proceedings of The 25th European Conference on Information Retrieval Research (ECIR03)*, volume LNCS 2633, pages 335–350. Springer-Verlag, 2003.

[4] J. Godfrey, E. Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of ICASSP-92*, pages 517–520, San Francisco, 1992.
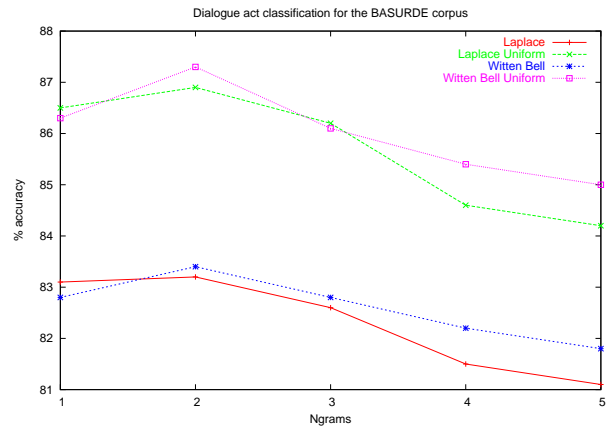
Figure 4: Dialogue act classification with the naive Bayes classifier: Basurde corpus.

[5] A. Bonafonte et al. Desarrollo de un sistema de diálogo oral en dominios restringidos. In *Primeras Jornadas de Tecnología del Habla*, Sevilla (Spain), 2000.

[6] Masaaki Nagata and Tsuyoshi Morimoto. First steps toward statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication*, 15:193–203, 1994.

[7] A. Stolcke et al. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373, 2000.

[8] James Allen and Mark Core. *Draft of DAMSL: Dialog Act Markup in Several Layers*, October 1997.

[9] Liz Shriberg Dan Jurafsky and Debra Biasca. *Switchboard SWBD DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13*. University of Colorado at Boulder and SRI International, August 1 1997.

[10] Andrew Kachites McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[11] G. Salton. The smart information retrieval system. *Prentice Hall, Englewood Cliffs, NY*, 1971.

[12] Emilio Sanchis, María José Castro, and David Vilar. Stochastic Understanding Models Guided by Connectionist Dialogue Acts Detection. In *Proceedings of 2003 IEEE Workshop on Automatic Speech Recognition and Understanding Workshop*, pages 501–506, St. Thomas, U.S. Virgin Islands, December 2003. IEEE.