Evaluating a Probabilistic Dialogue Model for a Railway Information Task*

Carlos D. Martínez-Hinarejos and Francisco Casacuberta

Institut Tecnològic d'Informàtica, Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, Camí de Vera, s/n, 46022, València, Spain
{cmartine, fcn}@iti.upv.es

Abstract. Dialogue modelling attempts to determine the way in which a dialog is developed. The dialogue strategy (i.e., the system behaviour) of an automatic dialogue system is determined by the dialogue model. Most dialogue systems use rule-based dialogue strategies, but recently, the probabilistic models have become very promising. We present probabilistic models based on the dialogue act concept, which uses user turns, dialogue history and semantic information. These models are evaluated as dialogue act labelers. The evaluation is carried out on a railway information task.

1 Introduction

The Computational Linguistics field covers a lot of natural language applications which have been developed over the past few years. Most of these applications are based on extracting rules from real data and, afterwards, include them in a computer system to develop the task. In contrast, viewpoint, probabilistic modelling attempts to automatically extract these rules from real data by using statistical inference techniques [1].

Dialogue systems are one of the most recent natural language processing applications. In these systems, a machine tries to emulate a human being in a dialog with a person[1] in order to achieve a final objective. The way the system behaves (i.e., the kind of answers and questions the system makes) is known as dialogue strategy [2]. This dialogue strategy is determined by the dialogue model. As we mentioned above, the dialogue model has usually been a rule-based model which is obtained from analyzing real dialogues from the task that the system is developed for [3]. However, recent efforts have been made in probabilistic models for dialogue systems [4]. The main advantage of these models is that they are easy to build, while the rule-based models are more difficult to build. However, the probabilistic models require annotated corpus.

* This work was partially supported by Spanish CICYT under projects TIC98–0423-C06 and TIC2000-1703-C03-01.

[1] We will use *dialogue* when referring to the general fact (*the dialogue*) and *dialog* when referring to a specific realization (*a dialog*)

This corpus annotation is based on the fact that dialogue models should only take into account essential information to determine the dialogue strategy. Therefore, the corpus should be annotated with labels which determine the essential information for each turn. One of the most popular options in dialogue labelling is the use of dialogue acts [5] to annotate the corpus. A dialogue act is a semantic label which takes into account the user's intention and the basic data given in a segment (a segment is a sub-utterance in the turn with isolated semantic meaning). This definition could be easily extended to system turns. Within this framework, the dialogue model should determine the next dialogue act(s) the system should perform.

In this work, we present two probabilistic dialogue models which are based on dialogue acts. Section 2 describes the basic model which is only based on the user's words. Section 3 extends this model by using semantic information provided by a semantic module. Section 4 describes the corpus and the experiments carried out with both models and the results obtained. Finally, Section 5 presents some conclusions about the results.

2 The initial probabilistic dialogue model

The general problem of dialogue could be viewed as a process of searching for the most correct (or most likely) action that the system could perform. This decision could be based on several factors, but the most common ones are last user turn and the previous history of the dialog. Using these two factors, our model should be able to determine the next dialogue act(s) to be performed.

From this perspective, the dialogue problem can be formulated as a search for the optimal system dialogue act that can be carried out (according to the last user turn and the dialog history). Therefore, given the last user-turn word sequence ω and the dialog history d (dialogue acts sequence), the system dialogue act $\hat{\mathcal{D}}$ that the model should determine is:

$$\hat{\mathcal{D}} = \operatorname*{argmax}_{\mathcal{D}} \Pr(\mathcal{D}|\omega, d) = \operatorname*{argmax}_{\mathcal{D}} \Pr(\mathcal{D}, \omega, d) \tag{1}$$

For the sake of simplicity, in the presentation of the model, and without a loss of generality, we assume that ω is divided into *segments*. Therefore, we can rewrite the probability term as:

$$\Pr(\mathcal{D}, \omega, d) = \sum_{D} \Pr(\mathcal{D}, D, \Omega, d)$$

where $D = D_1 D_2 \dots D_l$ is the sequence of dialogue acts of the last user turn (each of which corresponds to a segment) and where $\Omega = \Omega_1 \Omega_2 \dots \Omega_l$ is the sequence of segments (each of which are a word sequence). The sum can be approached with a max operator, which is:

$$\Pr(\mathcal{D}, \omega, d) \approx \max_{D} \Pr(\mathcal{D}, D, \Omega, d)$$

Now, we can make the following breakdown:

$$\Pr(\mathcal{D}, D, \Omega, d) = \Pr(d)\Pr(\mathcal{D}, D, \Omega|d)$$

$$\Pr(\mathcal{D}, D, \Omega|d) = \Pr(D, \Omega|d)\Pr(\mathcal{D}|d, D, \Omega) \tag{2}$$

Now, the first factor on the righthand side of (2) can be broken down into:

$$\Pr(D, \Omega|d) = \prod_{i=1}^{l} \Pr(D_i, \Omega_i|D_1^{i-1}, \Omega_1^{i-1}, d) \tag{3}$$

Each term of the product of (3) can be rewritten as:

$$\Pr(D_i, \Omega_i|D_1^{i-1}, \Omega_1^{i-1}, d) = \Pr(D_i|D_1^{i-1}, \Omega_1^{i-1}, d)\Pr(\Omega_i|D_1^{i}, \Omega_1^{i-1}, d)$$

Where $D_k^m = D_k D_{k+1} \dots D_m$ and $\Omega_k^m = \Omega_k \Omega_{k+1} \dots \Omega_m$, with $k \leq m$.

Several approximations can be adopted to reduce the model complexity. For this term, we make the following assumption:

$$\Pr(D_i, \Omega_i|D_1^{i-1}, \Omega_1^{i-1}, d) \approx \Pr(D_i|D_{i+1-n}^{i-1})\Pr(\Omega_i|D_i) \qquad n \geq 2 \tag{4}$$

That is, we assume that a dialogue act only depends on the last n dialogue acts (not on the whole dialogue history), and that the words of a segment depend only on the corresponding dialogue act of the segment. The first term can be easily approached by a n-gram (i.e., a $n-1$ length history) and the second term can be approached using Hidden Markov Models (HMM).

Now we deal with the second term on the righthand side of (2). It can be assumed that:

$$\Pr(\mathcal{D}|d, D, \Omega) \approx \Pr(\mathcal{D}|{d'}_{s+2-m}^{s}) \qquad m \geq 2 \tag{5}$$

Where s is the number of dialogue acts in the history d', which is equal to concatenating D to d. This assignation of the new dialogue act can be performed using a m-gram language model (i.e., a $m-1$ length history). Obviously, other (more realistic and more expensive) assumptions can be adopted.

Eventually, the simplified and approximated model obtained is:

$$\underset{\mathcal{D}}{\operatorname{argmax}} \left[\max_{D} \Pr(\mathcal{D}|{d'}_{s+2-m}^{s}) \prod_{i=1}^{l} \Pr(D_i|{D'}_{i+1-n}^{i-1})\Pr(\Omega_i|D_i) \right] \tag{6}$$

where D' is equal to $d \cdot D_1 \cdots D_{i-1}$. In other words, it is necessary to extend the current user history to the previous dialogue acts in the dialog in order to get enough history for the used n-gram. Therefore, the argument $\mathcal{D}$ which maximizes the previous formula is the next dialogue act that the system should perform. This model was previously presented by the authors in [6].

The meaning of the basic parts of the model are:

- $\Pr(D_i|D_{i+1-n}^{i-1})$: this is the N-gram used for assigning the user's dialogue acts; note that the used history is not only limited to the dialogue acts of the current turn (it can be extended to previous dialogue acts and even to system dialogue acts).
- $\Pr(\Omega_i|D_i)$: this is the model that assigns a dialogue act based on the words of the segment, i.e., an emitting model (such as a HMM) [2].
- $\Pr(\mathcal{D}|d_{s+2-m}^{s})$: this is the N-gram which assigns the most likely dialogue act based on the dialogue history (which is limited to $m-1$ previous dialogue acts).

All these probability distributions could be automatically estimated using a labeled corpus of dialogs. However, more information sources can be added to this simple model. In the following section, we explain how to add semantic information to this basic model in order to obtain a more powerful model.

3 Including semantic information: the extended dialogue model

The model presented in (6) is quite simple, but this simplicity makes it weak in certain situations. More information sources can be added to strengthen this model. Most dialogue systems use an understanding module which assigns a sequence of semantic units to a word sequence. This feature can be easily incorporated into our dialogue model, which can assign dialogue acts using only the semantic unit sequence or combining it with the word sequence.

Therefore, the initial optimization problem can now be formulated by:

$$\hat{\mathcal{D}} = \underset{\mathcal{D}}{\operatorname{argmax}} \Pr(\mathcal{D}|u, \omega, d) \tag{7}$$

where u is the semantic units sequence which is given by the understanding module.

If we develop (7) in the same way as we did (1), the following simplified model (which was also presented in [6]) can be obtained:

$$\underset{\mathcal{D}}{\operatorname{argmax}} \left[\max_{D} \Pr(\mathcal{D}|{d'}_{s+2-m}^{s}) \prod_{i=1}^{l} \Pr(D_i|{D'}_{i+1-n}^{i-1}) \Pr(U_i|D_i) \Pr(\Omega_i|D_i) \right] \tag{8}$$

where U_i represents the semantic unit sequence of the segment i (as Ω_i represents the word sequence of the segment i). In this model, $\Pr(U_i|D_i)$ can be modeled as $\Pr(\Omega_i|D_i)$, i.e, using an emitting model such as a HMM.

The usefulness of the models presented in (6) and (8) when implementing a complete dialogue system was presented in [6]. However, a specific evaluation of the models to compare their quality with other models was not carried out in that work. In the following section, we describe the experiments carried out in order to make the evaluation of the models as labellers [4].

[2] Note that in practice the segments are not given.

4 Evaluation experiments and results

In this section, we will describe the dialog corpus used in the evaluation, the implementation results and the results obtained with the presented models.

4.1 Corpus details

The corpus used in our evaluation is known as *Basurde* [7]. *Basurde* is a project about building a telephone dialogue system for spontaneous speech in Spanish for a railway information task. In this task, the user can ask about timetables and fares for the nation-wide trains. This corpus contains a total of 226 spoken dialogs in Spanish which were obtained using the Wizard of Oz technique [8]. These dialogs ask for typical information about railways, such as departure and arrival times, cost of the trip, train types, extra services, etc.

These 226 dialogs were transcribed and semantically annotated. The semantical annotations were made at two levels: at the *frame* level, which provides the information of the current turn, and at the *understanding* level, which provides the adequate understanding labels for the subsegments of the current turn. Only 194 dialogs were fully semantically annotated; 19 were partially annotated due to their complexity and the rest were not annotated due to special complications (they were mainly nonsense or out-of-task dialogs).

The entire 226 dialogs were also annotated at dialogue level using the set of labels defined in [9]. These labels are composed by three levels; the first level expresses the *speech act*, i.e., the intention of the user in the segment; the second level provides the *frames* (collections of data) used in the segment; the third level provides the *cases*, i.e., the specific data given in the segment. A total of 565 dialogue acts (labels) were defined (391 for the user turns and 174 for the system turns).

This corpus was divided into training and test subsets. The training subsets were different for each model (because of the different availability of data for each model). However, the test subset was formed by the same 75 dialogs. These test dialogs were also annotated, and this annotation is the reference for the comparision of the transcriptions that the model provided.

4.2 Experimental issues

A categorization was defined for the model defined by (6) in order to reduce data sparseness. This categorization included names of cities, hours, days, fares, services and train types among others. The total number of dialogue acts included in the training dialogs was 432 (137 for the system and 295 for the user), and the total number of training segments (for the user) was 1197.

The model defined by (8) was implemented using one more assumption. In this implementation, $\Pr(U_i|D_i) \cdot \Pr(\Omega_i|D_i)$ was implemented by $\Pr(\Omega_i, U_i|D_i)$. Ω_i, U_i is considered as a unique sequence. Therefore, the input for this model is different from the input for the previous model; this input must contain the words and the understanding labels. In our implementation, the understanding

Original user turn: *I want to go to Madrid on the Alaris.*
Semantically annotated turn: *I want*:consult *to go*:<dept_hour> *to*:dest_marker *Madrid*:dest_town *on the Alaris*:train_type .
Semantically annotated and categorized turn: *I want*:consult *to go*:<dept_hour> *to*:dest_marker **INSTANCE**:dest_town *on the* **INSTANCE**:train_type .

Fig. 1. User turn and its semantic final form

labels were obtained from an understanding module which is fully described in [10]. The obtained sequence can also be categorized in a way similar to the one for the previous model. An example of this process is shown in Fig. 1.

The total number of dialogue acts included in the training dialogs for this model was 394 (137 for the system and 257 for the user), and the total number of training segments (for the user) was 1060.

For both models, HMM with a two-state, left-to-right topology with loops were used as emitting model, using the available data for each model to estimate their parameters. A non-smoothed 3-gram, which was estimated from the 151 training dialogs, was used as a dialog history model. The evaluation was carried out assuming the correct history previous to the current user turn and using the model to obtain these user turn dialogue acts (i.e., only the user assignation models, $\Pr(D_i|D_{i+1-n}^{i-1})$ and $\Pr(\Omega_i|D_i)$ or $\Pr(\Omega_i, U_i|D_i)$ were used).

4.3 Results

Each model was applied to the 75 test dialogs in order to obtain a labelling of the user turns. This labelling was compared to the reference labels in the original annotated dialog using the *accuracy* measure [4], which in our case was computed as $acc = 100 \cdot \frac{corr}{corr+sub+ins+del}$, where *corr* is the number of matching dialogue acts, *sub* the incorrect substitutions, *ins* the number of insertions and *del* the number of deletions. In order to obtain results with a more reasonable number of labels, we can reduce our evaluation to the first and second levels only, or to a more reduced set of labels using the similarities among these labels.

We also have to take into consideration the number of test dialogue acts that are not present in the training set, i.e., the segments which are labelled with a dialogue act which is not present in the training set. We call the proportion of these segments in the test set the *forced error rate*. Therefore, these segments are disregarded when computing the real accuracy, which is defined as $acc_{real} = 100 \cdot \frac{acc}{100-forced}$, where *acc* is the accuracy calculated with the previous formula and $forced$ is the forced error rate.

All the evaluation results (for the different sets of labels) are presented in Table 1. These results indicate a very low accuracy rate for the whole set of labels. The rate becomes higher when the set of labels is reduced. Accuracy also improves when semantic information is used. A brief comparation with other models is shown in Table 2.

Table 1. Accuracy results for dialogue models defined by (6) and (8)

Model	Number of labels	Accuracy	Forced errors	Real accuracy
Defined by (6)	391	17.9 %	16.1 %	21.3 %
Defined by (8)	391	23.6 %	17.5 %	28.6 %
Defined by (6)	101	33.6 %	3.4 %	34.8 %
Defined by (8)	101	38.8 %	6.7 %	41.6 %
Defined by (6)	35	54.9 %	0.8 %	55.3 %
Defined by (8)	35	55.3 %	3.8 %	57.4 %

Table 2. Accuracy results for different models, with different number of labels and different size in the training sets

Model (Authors)	Task	Nr. of labels	Training size	Accuracy
Our best model (361 lab.)	Basurde,2002	391	1060	28.6 %
Our best model (101 lab.)	Basurde,2002	101	1060	40.4 %
Our best model (35 lab.)	Basurde,2002	35	1060	57.4 %
Samuel	Verbmobil,1998	18	2701	75.1 %
Wright	MapTask,1998	12	3276	64.0 %
Fukada	Jap. C-Star, 1998	26	3584	81.2 %
Nagata	ATR, 1994	15	2450	39.7 %
Stolcke	SwitchBoard, 2000	42	198000	65.0 %

As Table 2 indicates, our results using all the labels are worse than the results obtained using other models. However, when restricting the comparison, our results improve dramatically and are closer to other systems' results. Nevertheless, these results are not really comparable, because of the variability of the tasks involved, the different label sets used and the different training set size.

5 Conclusions and future work

In this work, we have presented a new probabilistic dialogue model. This dialogue model basically uses a Hidden Markov Model based on words to assign the dialogue act(s) to the user turn. It also uses a N-gram as language model when assigning these dialogue acts. This model can be easily extended to use semantic sequences. The evaluation of the model is carried out by using it as labeller for the user turns.

However, the results obtained are not comparable to other tasks, and no conclusions about performance can be obtained. Furthermore, details on the evaluation of the other models are not available (accuracy definition, test set details, evaluation process, etc.) and, therefore, it is not clear whether the evaluation process is adequate to compare results.

In our opinion, this independent evaluation of the model is not precise enough to evaluate a complete dialogue system, which depends on more factors. A more complete evaluation of the dialogue system (using the EAGLES metrics [11], for example) is necessary, although it would be more costly. In spite of this, we also conclude that the proposed model might be a good starting point for developing more complicated and correct models which may improve the results obtained.

Future work on the proposed model is channeled in two main directions. The first is to improve the basic models on which the model is based; for example, the use of a 4-gram or 5-gram instead of the current 3-gram may improve the accuracy of the model. The second is to add more information sources to the current model; for example, the frame state (the data provided the past dialog turns) could be an appropriate source of information. It would be interesting to model the influence of the frame state on the probabilistic model we have proposed.

References

1. C. Manning, H. Schütze 2001. *Foundations of statistical natural language processing.* MIT Press
2. A. Zampolli G.B. Varile. 1996. *Survey of the state of the art in human language technology.* Cambridge University Press, Giardini Editori.
3. M. Araki and S. Doshita. 1998. A robust dialogue model for spoken dialogue processing. In *Proc. Int. Conf. on Spoken Language Processing*, volume 4, pages 1171–1174.
4. A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. 2000. Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):1–34.
5. T. Fukada, D. Koll, A. Waibel, and K. Tanigaki. 1998. Probabilistic dialogue act extraction for concept based multilingual translation systems. In *Proc. Int. Conf. on Spoken Language Processing*, volume 6, pages 2771–2774.
6. C. D. Martínez-Hinarejos, F. Casacuberta. 2002. Probabilistic dialogue modelling Submitted to *40th. Aniversary Meeting of Association for Computational Linguistics.*
7. A. Bonafonte, P. Aibar, N. Castell, E. Lleida, J. B. Mariño, E. Sanchis, M. I. Torres 2000. Desarrollo de un sistema de diálogo oral en dominios restringidos. I Meeting on Language Engineering, Sevilla, Spain, 6th-10th November 2000
8. M. Fraser and G. Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, 5:81–99.
9. C. Martínez, E. Sanchis, F. García, P. Aibar 2002. A labelling proposal to annotate dialogues Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, 29-31 May, 2002, to appear
10. E. Segarra, E. Sanchis, M. Galiano, F. García, and L. Hurtado. to appear. Extracting semantic information through automatic learning techniques. *Pattern Recognition Letters.*
11. N. Fraser, 1997. *Assessment of interactive systems*, pages 564–614. Mouton de Gruyter.