A Labelling Proposal to Annotate Dialogues

Carlos D. Martínez-Hinarejos*, Emilio Sanchis†, Fernando García-Granada†, Pablo Aibar‡

*Instituto Tecnológico de Informática
U. Politécnica de Valencia, Cno. de Vera, s/n, 46071, Valencia, Spain
cmartine@iti.upv.es

†Departamento de Sistemas Informáticos y Computación
U. Politécnica de Valencia, Cno. de Vera, s/n, 46071, Valencia, Spain
{esanchis,fgarcia}@dsic.upv.es

‡Departamento de Informática
U. Jaume I, Campus Riu Sec, 12071, Castellón, Spain
aibar@inf.uji.es

Abstract

Stochastic models are widely used in some fields of Language Technology. Dialogue systems are one interesting application in Language Technology. In recent years, the stochastic modelling approach of dialogue systems has gained interest. These stochastic models are estimated from a set of annotated dialogues. The definition of the set of labels to annotate dialogues is therefore an important issue in the development of stochastic dialogue models. We propose a set of labels, which is composed of three levels, and a set of rules for using them. The application of this labelling to a specific set of dialogues is reported. The adequacy of the set of labels for stochastic modelling is also demonstrated.

1. Introduction

Stochastical models are widely used in the field of Language Technology, such as acoustic-phonetic modelling or language models. One of the most interesting applications in the Language Technology field is the development of dialogue systems. The dialogue management of many of these systems is based on rules obtained from the knowledge about the general behaviour of dialogues and the observation of some training dialogues. However, in recent years the stochastic modelling approach for dialogue systems has gained interest (Levin et al., 2000; Stolcke et al., 2000).

Although the specific characteristics of the dialogue structure and strategies seem to make the use of stochastic models alone difficult, they can be useful in some parts of dialogue management. To obtain the models (n-grams, HMM), a set of dialogue-act labels must be defined and the parameters of the models must be estimated from a set of dialogues which are annotated using these labels. The models can be used to classify the dialogue acts which are associated to each user turn, in order to give a prediction of the expected user dialogue act and to generate the system turns.

An important issue in the development of stochastic models is the definition of the set of labels to annotate dialogues (Allen and Core, 1996; Klein, 1999). In order to define a good set of labels we have to take into account the number of labels (the number must be enough to show the different intentions of the turns and to obtain good estimations of the stochastic models). We also have to consider that the set of labels should be general enough to be used in any task and be precise enough to deal with a specific task.

In this work, we propose a set of dialogue act labels which is divided into three different levels. This allows us to annotate an entire dialogue corpus and to estimate stochastic models which are used in some dialogue tasks.

2. The Three-level Labelling

The definition of dialogue acts is an important issue because they represent the successive states of the dialogue. The labels must be specific enough to show the different intentions of the turns and thus cover all the situations and they must be general enough to be easily adapted to several tasks. If the number of labels is too high the models will be underestimated because of the sparseness of the training samples. On the other hand, if we define a set of just a few labels only general purposes of the turn can be modelled.

The main feature of the proposed labelling is the division into three levels which is based on the idea presented in (Fukada et al., 1998). The first level, called *speech act*, is general for all the possible tasks. The second and third level, called *frames* and *cases*, respectively, are specific to the working task and give the semantic representation (Fillmore, 1968). With this structure, the labelling is general enough to be applied to other tasks and specific enough to cover all the possible situations in the dialogue.

A label is associated to a segment. A segment is a basic understanding unit inside a turn (i.e., a segment by itself has significant information at dialogue level). Thus, a turn contains one or more segments and each label takes the semantics of the associated segment.

2.1. First Level: Speech Act

The first level takes into account the intention of the segment (i.e., the dialogue behaviour) and has a unique value. For this level, we define the following values, which are common to every task:

- Opening: greetings at the begining of the dialogue.

- Closing: final dialogue segments.
- Undefined: filling words.
- Not understood: the previous turn was not understood and it is necessary to repeat it again.
- Waiting: the system is consulting an external data source.
- Consult: the system requires the user to make a new query.
- Acceptance: accepts data from the previous turn.
- Rejection: rejects data from the previous turn.
- Question: a question about data not given in previous turns.
- Confirmation: a question to confirm data given in previous turns.
- Answer: any answer which cannot be considered as acceptance or rejection.

2.2. Second Level: Frames

The second level is specific to each task. It is assumed that any dialogue has a repository to store the data given by the user or retrieved from the external databases. This repository of information is divided into *frames*. Each frame organizes the data in sets of necessary and optional values used to query the database or to give the response to the user. Therefore, this level indicates the frames which are used in the associated segment. Some segments do not use any frame, so a metalabel *Nil* is defined and used to indicate segments of this kind.

2.3. Third Level: Cases

The third level is also specific to the task. Each frame has a set of slots which have to be filled to make a query or that are filled by the retrieved data after the query. The specific data which fills the slots is known as *cases*. This level takes into account the slots which are filled by the specific data present in the segment, or the slots being used to generate the segment corresponding to an answer. To complete this level, it is necessary to analyze the words in the turn and to identify the case corresponding to each word. The segment may not reference any specific data, so the metalabel *Nil* is also used to label the third level of this class of segments.

3. Applying the Labelling on a Specific Task

A set of dialogue acts using the structure described in Section 2. has been defined for a specific corpus. This corpus is known as *Basurde*.

The *Basurde* task is about an automatic railway information system which can give the user timetables and fares for Spanish trains. In order to define and to limit the task, an analysis of interactions with a real human-to-human information system was done. A set of 200 person-to-person dialogues corresponding to real calls to the Spanish railway information system was recorded. This corpus was used to limit the task domain and to define dialogue strategies and answer generation.

Afterwards, four types of scenarios were defined (one-way trip, return trip, timetables and prices, and free scenario), and some dialogues were acquired by using the Wizard of Oz (WoZ) technique (Fraser and Gilbert, 1991). Specific instances of these scenarios were given to several volunteers who called the Wizard of Oz service. Each one of these 75 volunteers performed 3 scenarios.

The second and third level labels for the Basurde task were defined using the whole corpus of dialogues obtained.. The set of frames defined (second level) is presented in Table 1. The set of cases defined (third level) is presented in Table 2.

A total of 226 dialogues were selected from the whole corpus. There are a total of 2,329 user turns in this final dialogue corpus, and the vocabulary size is 868 words. This corpus was manually segmented and labelled using the defined set of dialogue acts and a set of rules defined by the labelling team. Some of the rules used are:

1. All the labels whose first level is **Opening**, **Closing**, **Undefined**, **Not_understood**, **Waiting** or **Consult** have a value of *Nil* in the second and third level, because these segments do not require any data management.
2. The segments classified as **Undefined** should be whole turns; in any case, **Undefined** segments should never interrupt other kinds of segments, in order to avoid an unreasonable number of segments in the turn.
3. For **Acceptance**, the second level should be labelled with the frames that are being confirmed. If the specific data is repeated, the third level will include the used cases; otherwise, it will be *Nil*. This rule specifies whether the data is repeated or not and thus provides greater or lesser confidence for the data given in the turn.
4. For **Rejection**, the segment only includes the specific part of the negation (which is followed by an **Answer** segment in most turns). The third-level label is always *Nil* and the second-level label corresponds to the rejected frames (if it is clear which data is rejected). This rule clearly differenciates the rejected data from the correct data (they are given in different segments) and avoids disregarding correct data (second level is only given when the rejection is completely clear).
5. For **Question**, it is not common for the second and third-level labels to be associated (in this case, it should be **Confirmation**). This is because of **Question** is usually about unknown data, and, therefore, no specific data is given in the segment (i.e., the third level is *Nil* or does not provide any data about the current item).
6. For **Confirmation**, the labels usually occur in both the second and the third levels.
7. **Answer** is applied only when **Acceptance** or **Rejection** cannot be used. This rule allows us to distinguish

Frame	Definition
Departure_time	Departure time of a train
Return_departure_time	Departure time of the return
Arrival_time	Arrival time of a train
Return_arrival_time	Arrival time of the return
Fare	Cost of the trip
Origin	Departure town or station
Destination	Arrival town or station
Trip_time	How long the trip takes
Stop_at	Stations or towns where the train stops during the trip
Departure_day	Date of departure
Arrival_day	Date of arrival
Train_type	Type of train used on the trip
Trip_type	One-way or round trip
Service	Class of the seat, services

Table 1: Frames definition for the Basurde task

Case	Definition
Origin	Departure town
Origin_station	Departure station
Destination	Arrival town
Destination_station	Arrival station
Day	Date of the trip
Departure_time	Departure time
Arrival_time	Arrival time
Fare	Fares (including terms such as cheap, expensive ...)
Stop_at	Stops during the trip
Train_type	Type of train (Intercity, Expreso, Talgo, ...)
Trip_type	One-way or round trip
Day_type	Labour day, holiday, weekend
Order_number	The order of the train (first, second, last ...)
Trip_duration	How long the trip takes
Number_trains	Number of trains which make the trip specified
Service	Class of seat (first, second ...) or services (bar, bed ...)

Table 2: Cases definition for the Basurde task

between clearly confirmed or rejected data and new data.

8. For user questions which do not have a clear objetive frame, a default frame is defined (in our case, *Departure_time*). This is an arbitrary decision based on the analysis of the dialogues of the task.

An example of a labelled dialogue [1] is presented in Figure 1.

4. Applying the Labelling to Obtain Stochastic Models

In this section, we explain some of the applications of the labelling. These applications are directed towards obtaining stochastic models which are used in several tasks in the field of dialogue systems: automatic labelling, synthetic dialogue generation and dialogue modelling.

[1]The original dialogue was in Spanish.

4.1. Automatic Dialogue Labelling

Manual dialogue annotation is a hard task which takes a long time and is sometimes very complex. Therefore, a tool which can automatically annotate the dialogues it would be very useful. This annotation tool should take a dialogue and return the same dialogue with its corresponding dialogue acts assigned for each turn.

This process can be seen as a transduction process: given a sequence of words (the dialogue), a sequence of dialogue acts and a segmentation of the word sequence is obtained. Therefore, stochastic transducers seem appropriate for this task. There is a transducer inference technique, called MGTI (Casacuberta, 2000) which can automatically build this transducer from a set of annotated and categorized dialogues. Categorization is necessary in order to avoid excessive data sparseness; the categorization in this work included towns, times, numbers and train types among other concepts. This technique was applied in the *Basurde* corpus to automatically annotate most of the corpus.

However, this process does not obtain completely cor-

Turn	Segment	Sentence	Level	Label
M0	M0-0	Welcome to the automatic information system about nation-wide trains. How can I help you?	1st 2nd 3rd	Opening Nil Nil
U0	U0-0	I want to make a trip on the third of December from Zaragoza to Castellón.	1st 2nd 3rd	Question Departure_time Destination,Day,Origin
M1	M1-0	I am going to consult times from Zaragoza to Castelló de la Plana on Friday, the third of December.	1st 2nd 3rd	Confi rmation Departure_time Destination,Day,Origin
	M1-1	Wait a moment, please.	1st 2nd 3rd	Waiting Nil Nil
U1		*empty turn*		
M2	M2-0	That day there are ten trains. The fi rst one leaves at 4:17 in the morning and the last one leaves at 19:20.	1st 2nd 3rd	Answer Departure_time Day,Departure_time, Order_number,Number_trains
	M2-1	Do you want more information?	1st 2nd 3rd	Consult Nil Nil
U2	U2-0	Yes, I want to know what train leaves in the middle of the morning.	1st 2nd 3rd	Question Departure_time Departure_time
M3	M3-0	Wait a moment, please.	1st 2nd 3rd	Waiting Nil Nil
U3		*empty turn*		
M4	M4-0	Between 10:00 and 12:00 in the morning, only one train leaves at 10:00 and arrives at 15:03.	1st 2nd 3rd	Answer Departure_time Arrival_time,Departure_time, Number_trains
	M4-1	Do you want more information?	1st 2nd 3rd	Consult Nil Nil
U4	U4-0	No, thanks.	1st 2nd 3rd	Closing Nil Nil
M5	M5-0	Thanks for using this service.	1st 2nd 3rd	Closing Nil Nil

Figure 1: An example of a labelled dialogue

rect annotations of the dialogues. Despite this fact, this tool can be used as a help in obtaining a previous annotation which only has to be revised by a human expert. The results obtained using this tool are presented in Table 3. Details on the process are reported in (Martínez-Hinarejos and Casacuberta, 2000b). As can be seen, this process adequately labels nearly half of the turns, and obtains a label which is almost correct in 24% of the turns.

	Total	Correct	Minor errors	Major errors
Turns	2655	1239	643	773
Prop	-	46.67 %	24.22 %	29.11 %

Table 3: Automatic labelling results

4.2. Synthetic Dialogue Generation

The main drawback of the stochastic modelling is that it needs lots of data. For several tasks (speech recognition, language modelling, etc ...) it is easy to obtain new data. However, obtaining new data for dialogue is much more difficult. For instance, if the WoZ technique is used, a human specialist is required and the transcription of the dialogues is complicated. Therefore, a tool which could generate synthetic dialogues would be very useful.

From the viewpoint of dialogue acts, a dialogue is a sequence of dialogue acts which are hidden within a user or system turn. Therefore, we can generate a dialogue as a

k	Correct	Strange	Incorrect
2	11	12	27
3	24	16	10
4	31	14	5

Table 4: Results for dialogue generation

sequence of dialogue acts. Also, a translation from the dialogue act to a word sequence can be provided to obtain real dialogues (i.e., at word-level between user and system).

We propose a dialogue generation technique which is organized in two steps. The first step is to generate a dialogue act sequence and the second step is to obtain a word sequence for every dialogue act. This results in a complete dialogue. The first step can be performed by a generating model, such as a stochastic finite automata. The second step can be performed by a stochastic transducer, as we did in Section 4.1.

The stochastic automata can be obtained from the dialogue act sequences of the corpus using the k-testable inference algorithm (García et al., 1987) with different values of k. After obtaining sequences of dialogue acts using this automata, a transducer can be obtained using the MGTI technique to transform the sequence of dialogue acts generated. This transducer is very similar to the one employed in automatic dialogue annotation but exchanges input and output alphabets.

This process was applied to generate a total of 50 dialogues for each k value used, $k = 2, 3, 4$. The automata were inferred from a total of 210 dialogue act sequences (which were extracted from the corresponding 210 dialogues). The transducer was obtained using the same 210 categorized dialogues (the same categorization as in Section 4.1. was used). The total 150 dialogue act sequences were parsed by this transducer, obtaining 150 categorized dialogues.

These dialogues were evaluated by human experts in order to determine their naturalness and adequacy. The evaluation results are presented in Table 4. We can conclude from the results that the bigger the k value is, the more natural the dialogue is. However, this k value should be limited in order to not reproduce the exact same training dialogue.

4.3. Stochastic Dialogue Modelling

The dialogue model is the core of a dialogue system. The dialogue model determines the so-called dialogue strategy (Varile and Zampolli, 1996), i.e, the system behaviour. Probabilistic dialogue modelling has gained interest in recent years (Levin et al., 2000; Stolcke et al., 2000), and most of the work done on this modelling is based on dialogue act annotation of the corpus.

A stochastic dialogue model based in our annotation scheme was presented in (Martínez-Hinarejos and Casacuberta, 2000a). This dialogue model is based on very simple stochastic models, such as Hidden Markov Models and N-grams. This model was also implemented and integrated in a dialogue system for the *Basurde* task, demonstrating it to be correct to perform the system task. Results with written input (using the EAGLES metrics (Fraser, 1997)) are reported in (Martínez-Hinarejos and Casacuberta, 2002), which demonstrate the adequacy of our labelling scheme.

5. Conclusions and Future Work

In this work, we have proposed a labelling scheme for dialogues. This dialogue scheme is structured in three levels; the first one is task-independent, and the second and third are task-dependent. We have shown the application of this scheme to a specific task, defining the second and third levels and applying the defined labels to an entire dialogue corpus. With this labelling process we demonstrate the appropriateness of the set of labels for the task. We have obtained several results using stochastic models based on this labelling scheme, which have also provided good quality results.

Future work is directed towards demonstrating the applicability of this labelling scheme to other tasks and to obtaining more powerful stochastic models based on this labelling scheme.

6. Acknowledgements

This work was partially supported by Spanish MCT under projects TIC2000-0664-C02-01 and TIC 2000-1599-C02-01. Authors wish to thank the anonymous reviewers for their comments and criticism.

7. References

J. Allen and M. Core. 1996. Draft of damsl: Dialog act markup in several layers. Technical report, University of Rochester, Department of Computer Science, December.

F. Casacuberta. 2000. Inference of finite-state transducers by using regular grammars and morphisms. In A. L. Oliveira, editor, *Grammatical Inference: Algorithms and Applications, volume 1891 of Lecture Notes in Computer Science*, pages 1–14, 5th International Colloquium Grammatical Inference -ICGI2000-. Springer-Verlag.

Ch. J. Fillmore. 1968. The case for case. *Universals in Linguistic Theory*.

M. Fraser and G. Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, (5):81–99.

N. Fraser, 1997. *Assessment of interactive systems*, pages 564–614. Mouton de Gruyter.

T. Fukada, D. Koll, A. Waibel, and K. Tanigaki. 1998. Probabilistic dialogue act extraction for concept based multilingual translation systems. In *Proceedings of the 5th. International Conference in Spoken Language Processing*, volume 6, pages 2771–2774.

P. García, E. Vidal, and F. Casacuberta. 1987. Local languages, the succesor method and a step towards a general methodology for the inference of regular grammars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(6):841–844.

M. Klein. 1999. Standardisation efforts on the level of dialogue acts in the mate project. In *Proceedings of the ACL Workshop: Towards Standards and Tools for Discourse Tagging*, pages 35–41, University of Maryland, May.

E. Levin, R. Pieraccini, and W. Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1):11–23.

C. D. Martínez-Hinarejos and F. Casacuberta. 2000a. Modelado probabilístico de sistemas de diálogo. In *Proceedings of the I Meeting on Language Engineering, to appear*.

C. D. Martínez-Hinarejos and F. Casacuberta. 2000b. A pattern recognition approach to dialog labelling by using finite-state transducers. In *Proceedings of 5th. IberoAmerican Symposium on Pattern Recognition*, pages 669–677.

C. D. Martínez-Hinarejos and F. Casacuberta. 2002. Probabilistic dialogue modelling. Submitted to 40th. Aniversary Meeting of Association for Computational Linguistics.

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 modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):1–34.

G. B. Varile and A. Zampolli. 1996. *Survey of the state of the art in human language technology*. Cambridge University Press, Giardini Editori.