Learning the Structure of Human-Computer and Human-Human Dialogs

David Griol[1], *Giuseppe Riccardi*[2], *Emilio Sanchis*[3]

[1]Dept. of Computer Science, Universidad Carlos III de Madrid, Leganés (Spain)
[2]Dept. of Information Engineering and Computer Science, University of Trento, Povo (Italy)
[3]Dept. de Sistemes Informatics i Computació, Universitat Politècnica de València, València (Spain)
dgriol@inf.uc3m.es, riccardi@disi.unitn.it, esanchis@dsic.upv.es

Abstract

We are interested in the problem of understanding human conversation structure in the context of human-machine and human-human interaction. We present a statistical methodology for detecting the structure of spoken dialogs based on a generative model learned using decision trees. To evaluate our approach we have used the LUNA corpora, collected from real users engaged in problem solving tasks. The results of the evaluation show that automatic segmentation of spoken dialogs is very effective not only with models built using separately human-machine dialogs or human-human dialogs, but it is also possible to infer the task-related structure of human-human dialogs with a model learned using only human-machine dialogs.
Index Terms: Domain Knowledge Acquisition, Spoken Dialog Systems, Dialog Structure Annotation.

1. Introduction

In the past decade, the computational linguistics community has focused on developing language processing algorithms that can leverage the vast quantities of corpus data that are available. The same idea can also be applied to the problem of acquiring domain-specific information. In this field, a machine learning technique could potentially reduce human effort in the knowledge engineering process and facilitate the development of a new dialog system.

Research on data-driven approaches to dialog structure modeling is relatively new and focuses mainly on recognizing a structure of a dialog as it progresses. Unsupervised clustering and segmentation techniques are used in [1] to identify concepts and subtasks in task-oriented dialogs. A vector-based approach is used in the Amitis system [2] to train both a task identification agent and a dialog act classifier in order to identify the customer's desired transaction and the corresponding dialog act of each utterance. Data-driven techniques are also applied in [3] to build task structures for dialog corpora. The resulting dialogs models include the combination of manually annotated information about dialog acts and task/subtask detection. Their research covers users dialog acts classification, user task/subtask classification, task/subtask system utterance prediction and system utterance dialog act prediction.

There is also a wide range of natural language processing applications for which discourse segmentation assists in. For instance, Angheluta, Busser and Moens adapted a three-step segmentation algorithm for automatic text summarization [4]. Different works apply discourse segmentation to segment text into different fragments in a preprocessing phase of a information retrieval system to improve its operation [5]. Walker applies this kind of techniques for anaphora resolution [6]. Finally, different studies show the benefits of using discourse segmentation for question answering tasks in order to take into account the context for the interpretation and answer questions [7], and also for dialog acts segmentation and classification [8].

In this paper, we present a machine learning approach for the automatic segmentation of spoken dialogs. The objective is to detect sequences of turns that accomplish a specific objective (tasks and subtasks) inside the dialog flow. These parts can be necessary for obtaining the final goal of the dialog, and also general parts not strictly related to the domain of the dialog system (greetings, error recovery, etc.). The detection of the dialog structure is useful to develop dynamic and adaptable dialog systems, using the information related to the dialog segment to adapt the behavior of the speech recognition, dialog manager, natural language understanding, and response generation modules to the specific characteristics of each dialog segment. Modeling subdialog structures is also useful to extend dialog systems to deal with more complex tasks.

In the literature, there are different studies for comparing human-machine (HM) and human-human (HH) interactions. Most of them compare specific features of both kind of conversations [9]. In our work, we try to infer the dialog structure of HH corpora by means of an active learning approach based on training an initial dialog model using the HM corpus and then use this model to learn the structure of HH conversations. To achieve this goal, two problems need to be addressed: i) creating a dialog representation that is suitable for representing the required domain-specific information, and ii) developing a machine learning approach that uses this representation to capture information from a corpus of in-domain conversations. This field presents as a main challenge the need of detecting the dialog segment using different information sources that are provided by both user and system entities during the course of the dialog (semantic information, confidence scores, task dependent and independent information, etc.). Our methodology is based on the use of decision trees classification to integrate all these different characteristics. This technique has been applied for the development of a discourse segmentation module for the LUNA project, using HM and HH interactions acquired for this project.

2. The LUNA corpora

The main objective defined in the LUNA project is to advance the state of the art in understanding conversational speech in spoken dialog systems [10]. Three aspects of SLU are of particular concern in LUNA: generation of semantic concept tags, semantic composition into conceptual structures, and context sensitive validation using information provided by the dialog manager. In order to train and evaluate SLU models, differ-

ent corpora of spoken dialogs in multiple domains and multiple languages have been acquired.

The LUNA corpus is currently being annotated, with a target to collect 8,100 human-machine dialogs (HM) and 1,000 human-human dialogs (HH) in Polish, Italian and French. The dialogs are collected in the following application domains: stock exchange, hotel reservation and tourism inquiries, customer support service/help-desk and public transportation. The Italian LUNA corpus will contain 1,000 equally partitioned HH and HM dialogs. These are recorded by CSI, an Italian customer care and technical support center. HH dialogs refer to real user conversations engaged in a problem solving task in the domain of software/hardware repairing. HM dialogs are acquired with a Wizard of Oz approach (WoZ). Ten different dialog scenarios inspired from the services provided by CSI were designed for the WoZ acquisition. Two corpora of 200 dialogs already labeled (200 HH and 200 HM dialogs) have been used for the experiments shown in this paper.

2.1. Features defined for the annotation of the LUNA corpora

The labeling defined for the LUNA corpora contains different types of information, that have been annotated using a multi-level approach [10]. The first levels are necessary to prepare the corpus for subsequent semantic annotation, and include segmentation of the corpus in dialog turns, transcription of the speech signal, and syntactic preprocessing with POS-tagging and shallow parsing. The next level consists of the annotation of main information using attribute-value pairs. The other levels of the annotation show contextual aspects of the semantic interpretation. These levels include the predicate structure, the relations between referring expressions, and the annotation of dialog acts.

The attribute-value annotation uses a predefined domain ontology to specify concepts and their relations. The attributes defined for the task include *Concept*, *Computer-Hardware*, *Action*, *Person-Name*, *Location*, *Code*, *TelephoneNumber*, *Problem*, etc.

Dialog act (DA) annotation was performed manually by one annotator on speech transcriptions previously segmented into turns as mentioned above. The annotation unit for DAs was the utterance; however, utterances are complex semantic entities that do not necessarily correspond to turns. Hence, a segmentation of the dialog transcription into utterances was performed by the annotator before DA labeling. The DAs defined to label the corpus are the following: i) Core DAs: *Action-request*, *Yes-answer*, *No-answer*, *Answer*, *Offer*, *ReportOnAction*, *Inform*; ii) Conventional DAs: *Greet*, *Quit*, *Apology*, *Thank*; iii) Feedback-Turn management DAs: *ClarificationRequest*, *Ack*, *Filler*; iv) Non interpretable DAs: *Other*.

For the predicate-argument structure annotation, we adopted the original FrameNet description of frames and frame elements, introducing new frames and roles only in case of gaps in the FrameNet ontology. In particular, we introduced 20 new frames out of the 174 taken from FrameNet because the original definition of frames related to hardware/software, data-handling and customer assistance was too coarse-grained. In this model, the meaning of predicates (or lexical units, usually verbs, nouns, or adjectives) is conveyed by frames, conceptual structures describing prototypical situations or events and the involved participants. Some of the frames included in this representation are *Telling*, *Greeting*, *Contacting*, *Statement*, *Recording*, *Communication*, *Being operational*, *Change operational state*, etc.

An example of the attribute-value, dialog-act and predicate structure annotations of a user utterance is shown below:

Good morning, I have a problem with my mouse.
Attributes-values: *Concept*:problem; *Hardware*:mouse;
Dialog acts: *Answer*;
Predicate structure: (*Greeting*)(*Problem_description*) *Device Problem*

2.2. Task structure

We consider a task-oriented dialog to be the result of incremental creation of a shared plan by the participants [3]. The basic structure of the dialogs included in the LUNA corpora is usually composed by the sequence of the following tasks: *Opening*, *Problem-statement*, *User-identification*, *Problem-clarification*, *Problem-resolution*, *Ticket-retrieval*, and *Closing*. The shared plan is represented as data register that encapsulates the task structure, dialog act structure, attribute-values and predicate-argument structure of utterances. Figure 1 shows an example of the incremental evolution of dialog structure with the complete set of tasks and substasks. It can be observed the difficulty of correctly detecting the complete structure of the dialog.

During the *Problem-statement* task (P1), the caller explains the problem the reasons why he/she calls the help-desk. In the *User-identification* task (P2), the operator asks for additional information regarding the identity of the caller. Once the caller has described the problem, the operator can ask for additional information to clarify it during the *Problem-clarification* task (P3).

During the *Problem-resolution* task, the operator asks the user to perform specific tests. We have defined nine different subtasks inside this generic segment, given that our goal is to detect not only that the dialog is in this segment, but also what are the specific problem that has to be resolved: *Printer* (P4), *Network connection* (P5), *PC going slow* (P6), *Monitor* (P7), *Keyboard* (P8), *Mouse* (P9), *CD-DVD player* (P10), *Power supply* (P11), and *Virus* (P12).

In the *Ticket-retrieval* phase (P13), the operator assigns a ticket number for the current call. The user must take note of this number and inform about this to the operator. Finally, the dialog participants close the dialog. The dialog participants ends the dialog during the *Closing* phase (P14).

The complete set of dialogs were manually labeled including this task/subtask information. This information was incorporated for each user and system turn in the dialogs. Figure 2 shows the distribution of dialog segments annotated in both corpora related to the tasks. As can be seen, the tasks distribution is very different in each corpus, presenting HH dialogs an additional 27.31% percentage of situations that has been labeled as *Out of the Task* (P15), as it is described in Section 4.

3. Our methodology for task/subtask prediction

As discussed previously, the dialogs in our task consist of several tasks and subtasks. The goal of subtask segmentation is to predict if the current utterance in the dialog is part of the current subtask or it starts a new subtask. We model this prediction problem as a classification task as the following equation shows:

$$\hat{S}_i = \underset{S_i \in \mathcal{S}}{\operatorname{argmax}} P(S_i | U_1 \cdots U_{i-1}, S_1 \cdots S_{i-1})$$

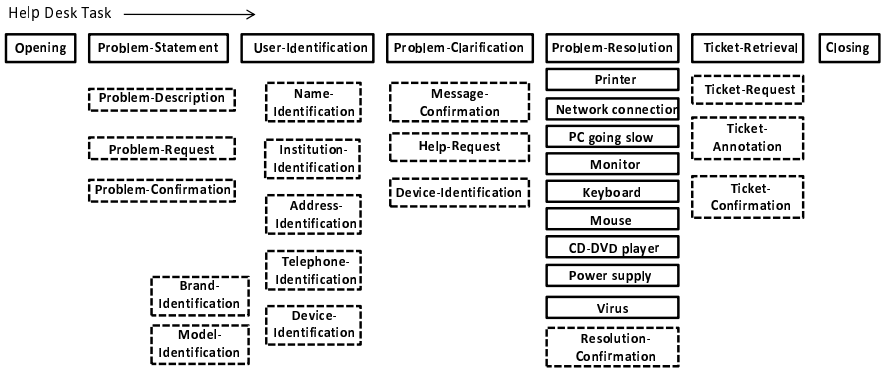

Figure 1: Incremental evolution of the dialog structure

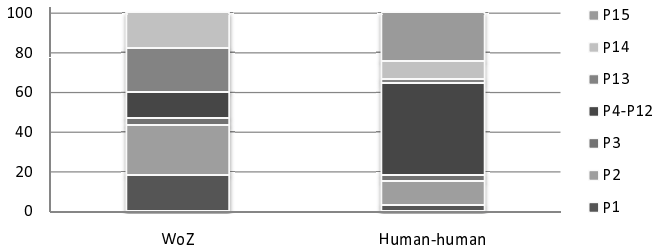

Figure 2: Distribution of dialog segments annotated in the HM and HH corpora

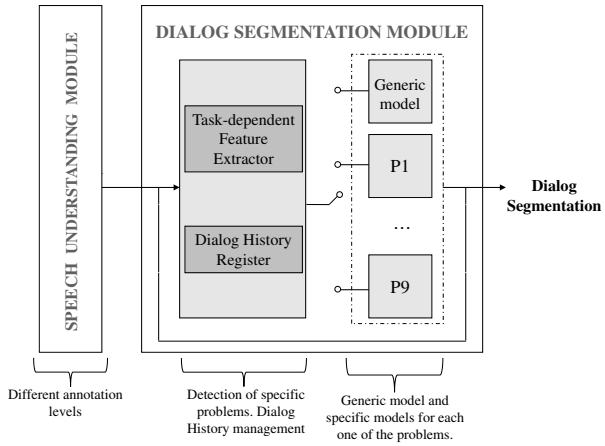

Figure 3: Graphical scheme of the proposed architecture for task and subtask detection

where set $\mathcal{S}$ contains all the possible kinds of tasks/subtasks defined for the dialog segmentation and U_n is the semantic representation of the user utterance at time n in terms of the list of features provided by the SLU module.

As a practical implementation of this methodology, we propose the use of two modules. The first module deals with the detection of the specific problem described by the user. This detection is based on the specific semantic information regarded to the task that is provided by the SLU module, that is to say, the attribute-value and predicate features for the reference annotations. This module also updates a register that contains the complete list of features provided by the SLU module through the dialog history since the current moment.

Until a specific problem is detected, a generic model learned with all the dialogs in the training partition is used for the selection of the segment of the dialog. Once the problem has been detected, a specific model learned using only the dialogs that deals with such a problem is used for the segment detection. The C4.5 decision tree learning algorithm have been used for the learning of these models, using the Weka machine learning software for classifying the complete list of features contained in the history register. Using these models, the current segment of the dialog is selected by taking into account the previous dialog segment detected for the module and the complete list of features provided by the SLU module. Figure 3 shows a graphical scheme of the proposed methodology.

4. Experiments and results

The developed methodology for the task detection has been evaluated by means of the HM and HH dialogs of the LUNA corpora. As HH dialogs are spontaneous, they present several differences with regard to the HM dialogs. The main one is the great difference in the average number of turns (11.18 turns in the HM corpus and 38.71 for the HH dialogs). This is because HH dialogs present other minor topics (like small talks about other persons, previous problems, holidays, etc), a high frequency of interruptions, cut-off phrases, and overlapped contributions. This makes that the 27.31% of the utterances of the HH corpus have been labeled as *Out of the Task*.

Analyzing the annotation available for the DA level, we measured that in average an HH dialog is composed of 48.9$\pm$17.4 (Std. Dev.) DAs, whereas a HM dialog is composed of 18.9$\pm$4.4. The difference between average lengths shows how HH spontaneous speech can be redundant, while HM dialogs are more limited to an exchange of essential information. The standard deviation of a conversation in terms of DAs is considerably higher in the HH corpus than in the HM ones. This can be explained by the fact that the WoZ follows a unique, previously defined task-solving strategy that does not allow di-

gressions.

From a comparative analysis of the DAs occurring in the HM and HH corpora, we noticed several important differences: i) *info-request* is by far the most common DA in HM, whereas in HH *ack* and *info* share the top ranking position; ii) the most frequently occurring DA in HH, i.e. *ack*, was only ranked 11th in HM; iii) *clarification-request*'s relative frequency (4.7%) is considerably higher in HH than in HM.

The relative frame frequency in HH dialogs is sparser than in the HM ones, meaning that the turns uttered by the machine influence the discourse topic and that the semantics of HH dialogs is more variable. The most frequent frame group comprises frames related to information exchange that is typical of the help-desk activity, including *Telling*, *Greeting*, *Contacting*, *Statement*, *Recording*, *Communication*. Another relevant group encompasses frames related to the operational state of a device, for example *Being operational*, *Change operational state*, *Operational testing*, *Being in operation*.

The evaluation of the statistical dialog segmentation module developed for the LUNA project was carried out *turn by turn* using a five-fold cross validation process. Each one of the two corpora was randomly split into five subsets. Each trial used a different subset taken from the five subsets as the test set, and the remaining 80% of the dialogs was used as the training set. Table 1 shows the results of the application of this methodology for the HM and HH corpora. The results show how the prediction is improved once the different SLU features are incorporated to the model. As can be seen, the proposed methodology successfully adapts to the requirements of the HM dialogs, since a 0.98 F-measure is obtained, measuring the dialog segments provided by the developed module that are equal to the segment annotated in the corpus for the HM dialogs. This value is reduced to 0.79 for the HH dialogs, since the *Out of the Task* class is usually confused with the rest of dialog segments related to the task. Therefore, the methodology adapts to the very different nature that has been described for both kind of dialogs.

Finally, we learned a model with the total 200 HM dialogs and evaluated it using the total 200 HH dialogs. This experimentation was designed to evaluate if a model learned with HM dialogs can detect the task-related structure of spontaneous HH conversations. The main challenge of this experiment is that only a maximum of 72,69% can be achieved due to the 27,31% *Out of the task* that is only present in the HH corpus. As can be observed, the model succesfully adapts to detect the task-related parts in the HH dialogs, achieving a 0.57 F-measure.

	Precision	Recall	F-measure
HM corpus for learning and evaluating			
Attribute-Values	0.89	0.87	0.88
DAs + Attribute-Values	0.94	0.92	0.93
Complete set	0.97	0.98	0.98
HH corpus for learning and evaluating			
Attribute-Values	0.72	0.60	0.66
DAs + Attribute-Values	0.86	0.71	0.78
Complete set	0.87	0.72	0.79
HM corpus for learning - HH corpus for evaluating			
Attribute-Values	0.52	0.45	0.48
DAs + Attribute-Values	0.59	0.51	0.55
Complete set	0.61	0.53	0.57

Table 1: Average results of the evaluation of the dialog segmentation module designed for the LUNA project

5. Conclusions

We have presented in this paper a statistical approach for automatically dialog segmentation in spoken dialog systems. This approach uses feature selection to collect a set of informative features into a model that includes both the information provided by the user and the system prompts. This model can be used to predict where boundaries occur in the dialog, helping the dialog manager in the selection of the next system prompt. The results of the evaluation of this methodology to develop a dialog segmentation module for the LUNA project show that the statistical approach successfully adapts to the requirements of the task, not only separately for human-machine and human-human dialogs acquired for this project, but also it is possible to successfully detect the task-related information that is present in spontaneous human-human dialogs by learning a model only with human-machine dialogs. As a future work, we want to perform a more detailed analysis of the situations that have been labeled as *Out of the Task*, studying if the task detector module is able to differentiate them. The experiments reported in this paper have been performed on transcribed speech. We want also to assess the performance of dialog structure prediction on recognized speech.

6. Acknowledgements

Work partially funded by the European commission LUNA project contract 33549 and the Spanish MEC and FEDER under contract TIN2008-06856-C05-02.

7. References

[1] A. Chotimongkol, "Learning the structure of task-oriented conversations from the corpus of in-domain dialogs," Ph.D. dissertation, Carnegie Mellon Univesity, Pittsburgh (USA), 2008.

[2] H. Hardy, A. Biermann, R. Inouye, A. Mckenzie, T. Strzalkowski, C. Ursu, N. Webb, and M. Wu, "Data-Driven Strategies for an Automated Dialogue System," in *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, 2004.

[3] S. Bangalore, G. D. Fabbrizio, and A. Stent, "Learning the Structure of Task-driven Human-Human Dialogs," in *IEEE Transactions on Audio, Speech, and Language Processing (Special Issue on New Approaches to Statistical Speech and Text Processing))*, vol. 16(7), 2008, pp. 1249–1259.

[4] R. Angheluta, R. D. Busser, and M. Moens, "The use of topic segmentation for automatic summarization," in *Proc. of the ACL-2002 Post-Conference Workshop on Automatic Summarization*, 2002, pp. 66–70.

[5] M. Kaszkiel and J. Zobel, "Passage retrieval revisited," in *Proc. of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, 1997, pp. 178–185.

[6] M. A. Walker, *Centering, anaphora resolution, and discourse structure*. Oxford University Press, 1998, pp. 401–435.

[7] J. Chai and R. Jin, "Discourse structure for context question answering," in *Proc. of the HLT-NAACL Workshop on Pragmatics of Question Answering*, 2004, pp. 23–30.

[8] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. of ICASSP'05*, 2005, pp. 1061–1064.

[9] C. Doran, J. Aberdeen, L. Damianos, and L. Hirschman, "Comparing several aspects of human-computer and human-human dialogues," in *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001, pp. 1–10.

[10] M. Dinarelli, S. Tonelli, A. Moschitti, and G.Riccardi, "Annotating Spoken Dialogs: from Speech Segments to Dialog Acts and Frame Semantics," in *Proc. of EACL 2009 Workshop on Semantic Representation of Spoken Language*, 2009, pp. 34–41.