Predicting Obligation Dialogue Acts from Prosodic and Speaker Information

Sergio R. Coria and Luis A. Pineda

Institute of Research in Applied Mathematics and Systems (IIMAS), National Autonomous University of Mexico (UNAM), Cto. Escolar S/N, Ciudad Universitaria, Coyoacan, 04510 Mexico City, Mexico
coria@turing.iimas.unam.mx, luis@leibniz.iimas.unam.mx

Abstract. In this paper a methodology and preliminary results of a machine learning experiment for correlating intonation patterns and speaker information with dialogue acts are presented. The goal of this work is to assess the extent to which prosodic and speaker data can help to identify obligation dialogue acts within a specific practical-dialogues audio and video corpus in Mexican Spanish. The machine learning method is decision trees. Current results show that the presented methodology is useful to the prediction of dialogue acts for the construction of conversational systems.

Keywords. Dialogue act interpretation, intonation, prosody, practical dialogues

1 Introduction

In this paper a methodology and preliminary results of a machine-learning experiment for correlating intonation patterns and speaker data with dialogue acts are presented. The goal is to assess the extent to which prosodic information can help to identify dialogue acts, along the general lines of [1]. The empirical resource used in this investigation is the DIME Corpus [2], a Mexican Spanish speech and video corpus, collected and tagged within the context of the DIME Project [3]. For the representation of intonation patterns, the INTSINT (International Transcription System for Intonation) [4] tagging scheme is used. Finally, the annotation of dialogue acts is performed with DIME-DAMSL [5], which is a multimodal extension to DAMSL [6]. With these resources, a machine learning experiment focused on the construction of decision trees using a CART-style algorithm [7] and the WEKA software [8] is currently being developed. For the experiment, the predictor data consist of INTSINT tags, utterance duration, modality and kind of speaker, and the target data is the obligation dialogue act tagged with DIME-DAMSL.

2 The DIME Corpus

The DIME corpus consists of a set of 26 task oriented dialogues in the kitchen design domain. The corpus was collected in a Wizard of Oz scenario (although the subjects knew that the Wizard was human). In the first phase of this project the corpus was segmented and transcribed orthographically. In the present phase a time aligned annotation in several layers is being developed; this includes the segmental (i.e. allophones) and suprasegmental (i.e. syllables, words and intonation patterns) layers; the corpus is also being tagged at the level of dialogue acts using the DIME-DAMSL annotation scheme. The most relevant tagging tiers for this experiment are: orthographic transcription, the INTSINT transcription, utterance duration (in milliseconds), modality (surface form), which was automatically predicted by a CART-style tree, kind of speaker (system or user), and dialogue acts transcription. The orthographic transcription of some instances of the corpus are as follows. In these transcriptions, *s* is the system (Wizard) and *u* is the human user.

utt1: s: ¿Quieres que desplace o traiga algún objeto a la cocina? (Do you want me to move or displace some object into the kitchen?)

utt2: u: <ruido> No (<noise> No.)

utt3: ¿Puedes mover la estufa hacia la izquierda? (Can you move the stove to the left?)

utt4: s: <ruido> ¿Hacia dónde? (<noise> where to?)

utt5: u: <ruido> Hacia <sil> hacia la derecha (<noise> to <sil> to the right.)

3 The Prosodic Transcription

Intonation patterns in the DIME Corpus are characterized through the INTSINT annotation scheme; in this scheme, intonation is modeled through a sequence of tags associated to the inflection points of the F0 (fundamental frequency) contour. The tag assigned to each inflection point is relative to its predecessor and its successor along the contour. The the tag set is: M (medium), T (top), B (bottom), H (higher), L (lower), U (up-step), D (down-step) and S (same). Tags are computed automatically by the INSINT tool using the MOMEL algorithm [9], and the MES software tool [10]. MOMEL provides a default stylized F0 contour; then a perceptual verification task is performed by human annotators. In this latter process inflection points are modified, added or deleted, until the stylized intonation matches the original intonation of the utterance. In addition to the prosodic transcription produced by MOMEL and INSINT, and utterance duration, the duration of lower units including syllables (phonetic), pauses, and break indices will be also available for future classification experiments.

The original F0 of the utterance *Eh... ¿me puedes mostrar los tipos de muebles que tengo?* (*Mmm... can you show me the kinds of furniture that I have?*) is shown in Figure 1.

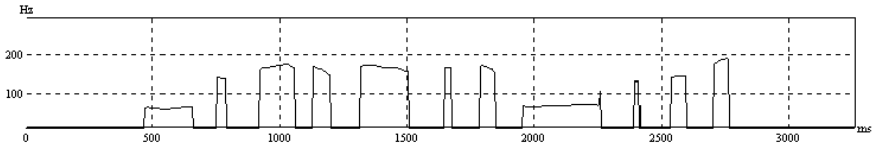

Fig. 1. Original F0

The prosodic transcription is performed in four major stages using M.E.S. The first is to extract the original F0 contour using AMDF (Average Magnitude Difference Function), autocorrelation or comb function algorithms; the second step is to produce the stylized contour using the MOMEL algorithm, which does not guarantee a perfect stylization and might produce a contour different from the original F0, as can be seen in Figure 2 (i.e. in the regions marked with 1, 2, 3 and 4); in the third stage, a human annotator develops a perceptual verification task in which inflection points could be relocated, eliminated or inserted until the stylized contour is perceived as the original F0 curve as shown in Figure 3; finally, the fourth step consists in to produce INTSINT tags automatically, as can be seen in Figure 4; for our example these are BSSUHSLHBSUTS. In addition to these four stages, and for the particular purpose of this experiment, INTSINT strings were cleansed by deleting *S* (*same*) tags because these are redundant. This transformation produces simpler strings without reducing the reliability of the representation. The final string for our example is BUHLHBUT.

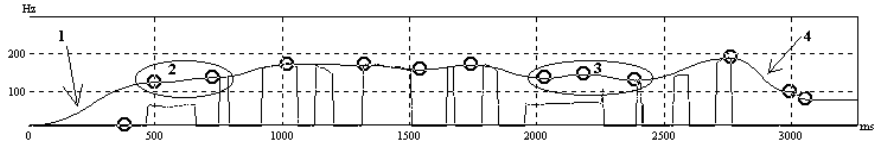

Fig. 2. Stylized F0 (*dark contour*) with its inflection points (*circles*)

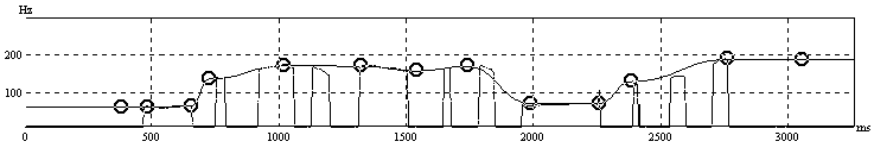

Fig. 3. Stylized F0 (*dark contour*) after perceptual verification

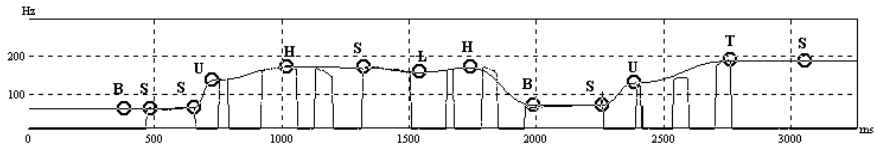

Fig. 4. INTSINT annotation of the inflection points

4 The Dialogue Act Transcription

The dialogue act transcription has been developed by using DIME-DAMSL scheme, which is a multimodal extension of DAMSL (Dialogue Act Markup in Several Layers). DAMSL is a dialogue acts annotation scheme structured in four dimensions: communicative status, information level, forward and backward looking functions. DIME-DAMSL extends DAMSL with the annotation of the graphical modality; this involves, for instance, pointing to, moving or adding a piece of furniture, or showing a catalogue. DIME-DAMSL considers two planes of annotation: obligations and common ground; this latter plane is divided into agreement and understanding levels. Dialogue acts annotation is a relatively subjective process, and a high enough agreement among taggers is required to produce significant conclusions; in our experiment the inter-annotators agreement was measured with the Kappa statistic [11].

Obligation dialogue acts chosen for the experiment were action directive (*action-dir*), information request (*info-request*), and commitment (*commit*) which belong to the forward looking function of DAMSL, and other (*other*). *Action-dir* requires the listener to perform an action or, if it is impossible or if he is not able to perform it, to inform this to the interlocutor. In the DIME Corpus, this dialogue act is frequently uttered by the user to give the Wizard a command. *Info-request* label is used if the speaker asks the listener some information; in the corpus, these utterances are frequently uttered by both the user and the Wizard. *Commit* is the obligation on the speaker himself to perform an action. We constrast these three dialogue acts with the *other* label, which is used to tag any other dialogue act. Table 1 shows examples of utterances representing the four dialogue acts considered in the experiment.

Table 1. Dialogue act taggings

UTTERANCE	DIALOGUE ACT TAG
utt3: u: Can you move the stove to the left?	action-dir
utt53: s: Where do you want me to put it?	info-request
utt26: s: okay	commit
utt63: s: These are the four types of cabinets that we have.	other

5 The Classification Task and Results

The present investigation has the purpose to assess the extent to which obligation dialogue acts can be predicted from their intonation pattern and speaker information by machine learning techniques, regardless their surface form (declarative, interrogative or imperative); for this experiment, utterances were classified into four different categories, as mentioned above. Table 2 shows a sample of some utterances of the DIME Corpus tagged for the experiment.

Utterance modality was first annotated manually according to the surface form of the utterance in Spanish: declarative (*dec*), interrogative (*int*) or imperative (*imp*). Table 3 presents some instances of the three modalities tagged for the experiment.

Table 2. A sample of the DIME Corpus annotations

UTT_ID	INTSINT TAGGING	DUR	MOD	KIND OF SPEAKER	DIALOGUE ACT
D12_utt1	BTLUTDLULUDUT	2564	int	s	other
D12_utt2	BT	837	dec	u	other
D12_utt3	MLHLTDBHLT	2671	int	u	action-dir
D12_utt4	MBT	1016	int	s	info-request
D12_utt5	BHDHTDB	3500	dec	u	other
D12_utt6	MTB	1276	dec	s	other
D12_utt7	BTDB	983	dec	s	other
D12_utt8	MTDDLHB	1725	int	s	info-request
D12_utt9	MUTDDLHDB	3103	dec	u	other
D12_utt10	MTB	582	dec	s	other

Table 3. Modality taggings

UTTERANCE	MODALITY TAG
utt35: s: <no-vocal> This is the catalogue of sinks and dish washing machines.	dec
utt59: u: <no-vocal> Can you show me the catalogue of stuff I have, again?	int
utt91: u: to me... show me... mm... the furniture <sil> all of them.	imp

5.1 Utterance Modality Prediction

Utterance modality is useful to predict dialogue acts in the DIME Corpus, as shown in previous experiments in [12]; however, modality tagging would not be available in a real world application, so it should be predicted from the real world data. Table 4 reproduces 3 out of the 19 rules from the tree presented in [12] to predict modality in the same annotated dialogue, where the numbers in parentheses are the number of cases complying/not comply each rule. The 19 rules use the data of the last 2 INTSINT labels of the INTSINT tagging. The tree accuracy is 85.1%, and Kappa (comparing against the manually tagged modality) is 0.70390. Recalls, precisions and F-Measures of the tree to predict modalities are reproduced in Table 5. The same tree is used to predict modality in the present experiment, using this as one of the predictor data.

Predicted modality is not as good as tagged modality to predict dialogue acts; however, the latter would not be available in a real-world system, and we should use the former. A pilot experiment showed that using predicted modality is better than using no modality at all.

Table 4. Some rules to predict modality (reproduced from [12])

RULES
if last_2 = UT, then int (20/1)
if last_2 = DB, then dec (20)
if last_2 = HB, then imp (3/1)

Table 5. Evaluation of the modality prediction from [12]

MOD	RECALL	PRECISION	F-MEASURE
dec	0.881	0.912	0.897
int	0.850	0.791	0.819
imp	0	0	0

5.2 Statistical Description of the Dialogue Data

Next, we present a general statistical description of the dialogue data set, which is useful to assess the results of the current experiment. Regarding intonation, the last 1 INTSINT tag of most of utterances is B (47%) or T (39%). In addition, considering the last 2 INTSINT tags, most of them (aprox. 81%) finishes in one of the following 6 tone pairs: UT, DB, BT, TB, MB or LT.

Average duration of utterances is 2,228.1 milliseconds, with a maximum of 13,339.2 and a minimum of 211.9. Range is 13,127.3 and standard deviation is 2,654.1. Most of utterances durations (80%) are less than or equal to 3,000 milliseconds. Almost all of utterances (97%) present *dec* or *int* predicted modalities; the remaining 3% is *imp*. The scarcity of *imps* might be a feature of kind of domain and setting; in preliminary analyses of several dialogues from our corpus, speakers use

interrogative or declarative modalities instead of imperative to express action directives.

Speaker information consists of a very simple data, which describes who uttered the utterance: the system (*s*) or the user (*u*). 60.4% were uttered by the system, and 39.6% by the user; the difference between these figures could be produced by the ratio of information requests uttered by the system to the user to confirm action directives and to specify their corresponding parameters; initiative in our dialogues depended on the system most of the time. The statistical relation between kind of speaker and dialogue act is depicted in Table 6, which shows that some dialogue acts are typically uttered by a specific kind of speaker: *commits* were always uttered by the system; *action-dirs* were always expressed by the user; *info-requests* were uttered by the system most of times (76.4%); and *others* are almost the same number for every speaker. Kind of speaker could be a useful data to predict dialogue act type.

Table 6. Relation between kind of speaker and dialogue act

	System	User	TOTAL	%
other	19	18	37	36.6
info-request	26	8	34	33.7
commit	16	0	16	15.8
action-dir	0	14	14	13.9
TOTAL	61	40	101	100.0
%	60.4	39.6		

Table 7. Dialogue acts Pareto

DIAL. ACT	FREQ.	%	ACCUM. %
other	37	36.6	36.6
info-request	34	33.7	70.3
commit	16	15.8	86.1
action-dir	14	13.9	100.0
TOTAL	101		

Table 7 presents a statistical analysis of dialogue acts. The utterances annotated as *other*, *info-request* and *commit* are 86.1% of the data set.

Regarding the relation of dialogue acts and predicted modality, most of the time *info-requests* were uttered as interrogatives (88%) as would be expected; *commits* were uttered in most of cases as declaratives (87.5%); *action-dirs* were both interrogatives (50.0%) and declaratives (42.8%) rather than imperatives; the rest (*other*) was declarative mainly (83.7%). This is consistent with what was already depicted in [12] about tagged modality and dialogue act.

5.3 Experiment

For the experiment, J48, a CART-style algorithm, and WEKA software were used to build decision trees. With these tools a dialogue of the DIME Corpus with 117 utterances, fully tagged in the relevant dimensions was used. Only utterances which had all taggings available were kept; this produced 101 useful utterances.

The predictor data were INTSINT cleansed strings (taking the last 5, 4, 3, 2, and 1 labels from every string), in addition to duration, predicted modality and (in one of the two experimental settings) kind of speaker; the target data was dialogue act. Five attributes were created by using INTSINT cleansed strings. Several trees were created to predict dialogue acts by using different training and testing subsets in order to validate and compare results. Three modes were considered: 1) subsets which are statistically representative (manually stratified) of the dialogue act types were used, where 70% was for training and 30% for testing; 2) subsets which were randomly defined although not strictly representative of the dialogue act classes were used in 10-fold, 5-fold, 3-fold and 2-fold cross validations; 3) finally, 50, 66, 70 and 75 percent of the whole data were splitted for training and the respective remainders were used for testing; these splits were randomly created and also they were not strictly representative of the dialogue act types. The combination of different attributes and training/testing modes permited the creation of forty-five decision trees where the kind of speaker was one of the predictor data, and other forty-five trees where it was not.

5.4 Results

As a result, the general average accuracy to predict dialogue act when using the speaker kind data was 66.1830%, with Kappa equal to 0.5153; the best results were obtained with the last 3 INTSINT labels datasets (68.7182% and 0.5538, averages); from the last 3 INTSINT labels datasets, the best tree had 74.1935% and 0.6265, obtained in mode 3 with 70% split. This could be considered the most useful tree and it is presented in Table 8.

Modality, kind of speaker and duration (on that order) were useful to predict dialogue act, while INTSINT tags were not necessary at this stage, although they were used for predicting modality, which is consistent with the results observed in [12]. The precisions, recalls and F-Measures of the predicted dialogue act types are presented in Table 9, where *info-request* is the best predicted, then *other*, then *commit*, and finally *action-dir*. *Action-dir* instances are the least frequent in the data as can be seen in Table 7 and the dataset available was too small to assess the result for *action-dir*.

If kind of speaker is not included as one of the predictor data, accuracy and Kappa averages are 59.4291% and 0.3987, respectively, with maxima of 71.4286% and 0.5630. The difference in accuracy averages comparing to kind of speaker dataset is 66.1830% - 59.4291% = 6.7539, and, regarding Kappas, the difference is 0.5153 - 0.3987 = 0.1116; so data of speaker kind is useful to improve the dialogue acts prediction.

Table 8. Decision tree to predict obligation dialogue acts

RULES	DIALOGUE ACT
if (pred_mod=int) and (sp_kind=s), then info-request (29/5)	info-request
if (pred_mod=int) and (sp_kind = u) and (dur > 1568.6875) and (dur <= 4514.875), then info-request (9/3)	
if (pred_mod = imp), then info-request (3/1)	
if (pred_mod=int) and (sp_kind = u) and (dur<= 1568.6875), then other (3)	other
if (pred_mod = dec) and (sp_kind = s) and (dur <= 1209.875) and (dur <= 652.75), then other (3)	
if (pred_mod = dec) and (sp_kind = s) and (dur > 1209.875), then other (9)	
if (pred_mod = dec) and (sp_kind = u) and (dur <= 1158.75), then other (12)	
if (pred_mod=int) and (sp_kind = u) and (dur > 1568.6875) and (dur > 4514.875), then action-dir (4)	action-dir
if (pred_mod = dec) and (sp_kind = u) and (dur > 1158.75), then action-dir(10/4)	
if (pred_mod = dec) and (sp_kind = s) and (dur <= 1209.875) and (dur > 652.75), then commit (19/5)	commit

Table 9. Evaluation of the dialogue acts prediction in Table 8

DIAL. ACT	RECALL	PRECISION	F-MEASURE
other	0.889	0.727	0.800
action-dir	0.200	0.500	0.286
info-request	0.917	0.786	0.846
commit	0.600	0.750	0.667

Althoug few data were available (one dialogue only) we consider that these preliminary results seem to be promising. Results show that identifying modality and using kind of speaker data to identify dialogue act could be useful for a prototype dialogue management system. Other interesting setting to be evaluated in the experiments for the short term is dialogue act tag of the previous utterance, using it as and additional predictor data.

Annotation process continues on other dialogues of the corpus. Completion of annotations is expected for the next months, including utterance intensities, stressed syllable durations, etc. This way, a larger ammount of attributes and data will be available for further experiments.

6 Discussion and further work

The present methodology promises a way to identify dialogue act types for the construction of dialogue managers for practical dialogues; the creation of the predicting model requires a manual tagging stage and a model-training stage; the model can be implemented as a series of *if-then* rules, whose result would be a dialogue-act tagging to feed a real dialogue manager. The rules would be quite simple, similar to those presented in this paper. These classification rules would analyze data obtained from the most recent speaker's utterance. In a real system, information regarding speaker kind (*system* or *user*) would be obtained on an instantaneous basis, because the system itself could distinguish immediately between what it said and what the user said. In a real implementation, manual correction of the stylized *f0* contour is not feasible; this could impact on automatic prosodic tagging and utterance modality prediction, and finally on automatic dialogue act recognition, so a stylization algorithm better than MOMEL is required. Rules to predict utterance modality can be enriched by analyzing more data from the corpus. Utterance duration can be easily extracted from the speech.

The present investigation will be continued with experiments focusing on the identification of other types of dialogue acts, and we will focus on the construction of a complete model including all dialogue act types contemplated in the DIME-DAMSL scheme. For the completion of this experiment we plan to use, in addition, syllable and pause durations, break indices, and some lexical information.

Acknowledgments

We thank useful comments and suggestions from James Allen, at Rochester University, and from Joaquim Llisterri and Monserrat Riera at Universidad Autonoma de Barcelona. We also thank Haydé Castellanos, Varinia Estrada, Fernanda López, Isabel López, Iván Meza, Iván Moreno, Patricia Pérez, Carlos Rodríguez, Javier Cuétara and Ivonne López who participated in the tagging task, and also for useful comments and suggestions. The theory and experiment reported in this paper are being developed within the context of the DIME-II project, with partial support of NSF/CONACyT grant 39380-A.

References

1. Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., Van EssDykema, C.: Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?, Language and Speech 41(3-4), Special Issue on Prosody and Conversation (1998) 439-487
2. Villaseñor, L., Massé, A., Pineda, L.: The DIME Corpus, ENC 01, 3er Encuentro Internacional de Ciencias de la Computación, SMCC-INEGI, Aguascalientes, Mexico (2001)

3. Pineda, L., Massé, A., Meza, I., Salas, M., Schwarz, E., Uraga, E., Villaseñor, L.: The DIME Project, Lecture Notes in Artificial Intelligence, Vol. 2313, Springer, (2002) 166-175
4. Hirst, D., Di Cristo, A., Espesser, R.: Levels of representation and levels of analysis for the description of intonation systems, In M. Horne (ed) Prosody: Theory and Experiment, Kluwer-Dordrecht (2000)
5. Pineda, L., Castellanos, H., Coria, S., Estrada, V., López, F., López, I., Meza, I., Moreno, I., Pérez, P., Rodríguez, C.: Balancing Transactions in Practical Dialogues, Technical report, Department of Computer Science, IIMAS-UNAM, Mexico (2005)
6. Allen, J., Core, M.: Draft of DAMSL: Dialog Act Markup in Several Layers, Technical report, The Multiparty Discourse Group. University of Rochester, Rochester, USA, October (1997)
7. Witten, I., Frank, E.: Data Mining. Practical Machine Learning Tools and Techniques with Java implementations, Morgan-Kauffman Publishers, San Francisco, CA, USA (2000)
8. Frank, E., Hall, M., Trigg, L.: WEKA: Waikato Environment for Knowledge Analysis software (2004). http://www.cs.waikato.ac.nz/~ml/weka
9. Hirst, D., Espesser, R.: Automatic modeling of fundamental frequency using a quadratic spline function, CNRS (URA 261), Institut de Phonétique d'Aix, Université de Provence, France (1993)
10. Espesser, R.: MES: Motif Environment for Speech software (1999). http://www.lpl.univ-aix.fr/ext/projects/mes_signaix.htm
11. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic, Computational Linguistics 22(2), (1996) 249-254
12. Coria, S., Pineda, L.: Predicting obligation dialogue act types from prosodic information, 2nd Midwest Computational Linguistics Colloquium, Ohio State University, U.S.A. (2005)