

Predicting obligation dialogue act types from prosodic information

Sergio R. Coria

coria@turing.iimas.unam.mx

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS)

Universidad Nacional Autónoma de México (UNAM)

Cto. Escolar S/N, Cd. Universitaria, Coyoacan, Ciudad de México, Mexico

Luis A. Pineda

luis@leibniz.iimas.unam.mx

Abstract

In this paper a methodology and preliminary results of a machine learning experiment for correlating intonation patterns with dialogue acts are presented. The goal of this work is to assess the extent to which prosodic information can help to identify dialogue act types.

1 Introduction

In this paper a methodology and preliminary results of a machine-learning experiment for correlating intonation patterns 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 (Shriberg et al., 1998). The empirical resource used in this investigation is the DIME Corpus (Villaseñor et al., 2001), a Mexican Spanish speech and video corpus, collected and tagged within the context of the DIME Project (Pineda et al., 2002). For the representation of intonation patterns, the INTSINT (International Transcription System for Intonation) (Hirst et al., 2000) tagging scheme is used. Finally, the annotation of dialogue acts is performed with DIME-DAMSL (Pineda et al., 2005), which is a multimodal extension to DAMSL (Allen and Core, 1997). With these resources, a machine learning experiment focused on the construction of decision trees using a CART-style algorithm (Witten and Frank, 2000) and the WEKA software (Frank et al., 2004) is currently being developed. For the experiment, the predictor data consists of INTSINT tags, utterance duration and modality, 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 tagged manually, 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 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 (Hirst and Espesser, 1993), and the MES software tool (Espesser, 1999). 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 are also available for the future classification task.

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. 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 BSSUHSBHBSUTS. 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 BUHBHBUT.

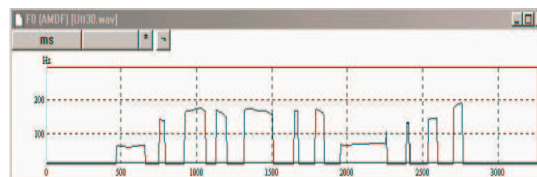


Figure 1: Original F0.

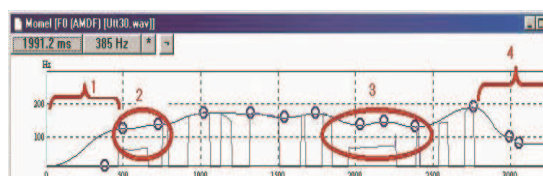


Figure 2: Stylized F0 (dark contour) with its inflection points (small circles).

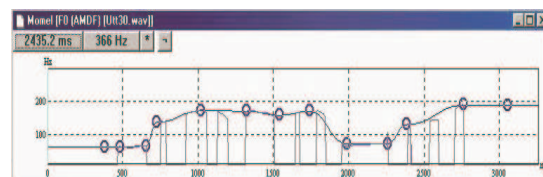


Figure 3: Stylized F0 (dark contour) after perceptual verification.

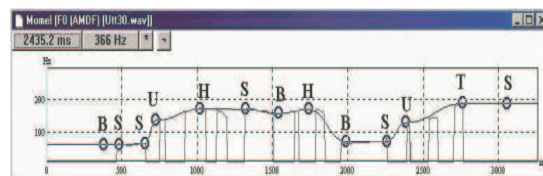


Figure 4: INTSINT annotation of the inflection points.

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 agreement was measured with the Kappa statistic (Carletta, 1996).

Obligation dialogue acts chosen for the experiment were action directive (*action-dir*) and information request (*info-request*), which belong to the forward looking function of DAMSL; in addition to mixed (*mixed*) 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. In addition, we identify a special kind of information request which is stated through an action directive; this is called mixed. We contrast these three dialogue acts with the *Other* label. Table 1 shows examples of utterances representing the four dialogue acts considered in the experiment.

UTTERANCE	DIALOGUE ACT TAG
utt3: u: Can you move the stove to the left?	action-dir
utt33: u: Can you show me the catalogue of sinks and machines?	mixed
utt53: s: Where do you want me to put it?	info-request
utt63: s: These are the four types of cabinets that we have.	other

Table 1: Dialogue act taggings.

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 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.

utt_id	intsint_tagging	duration	mod	dial_act
d12_utt1	BTLUTDLULUDUT	2564	int	other
d12_utt2	BT	837	dec	other
d12_utt3	MLHLTDBHLT	2671	int	action-dir
d12_utt4	MBT	1016	int	info-request
d12_utt5	BHDHTDB	3500	dec	other
d12_utt6	MTB	1276	dec	other
d12_utt7	BTDB	983	dec	other
d12_utt8	MTDDLHB	1725	int	info-request
d12_utt9	MUTDDLHDB	3103	dec	other
d12_utt10	MTB	582	dec	other

Table 2: A sample of the DIME Corpus annotations.

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

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

Table 3: Modality taggings.

Next, we present a general statistical description of the dialogue data set. 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 and LT.

Average duration of utterances is 2,228.1, with a maximum of 13,339.2 and a minimum of 211.9 milliseconds. 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 (98%) present dec or int modalities. The remaining 2% is imp.

Table 4 presents a statistical analysis of dialogue acts. The utterances annotated as *other* and *info-request* are almost 80% of the data set. The relation between dialogue acts and modalities is shown in Table 5.

DIAL. ACT	FREQ.	%	ACCUM. %
other	53	52.5%	52.5%
info-request	26	25.7%	78.2%
action-dir	14	13.9%	92.1%
mixed	8	7.9%	100.0%
TOTAL	101		

Table 4: Dialogue acts Pareto.

Most of the time action directives were not uttered as imperatives as would be expected but rather as interrogatives or declaratives; information requests were uttered in all cases as interrogatives; mixed dialogue acts (information request plus action directive) were interrogatives most of time; the rest was declarative mainly. There was a strong statistical relationship between dialogue act and modality; so modality is an important attribute to predict dialogue acts, but modality itself needs to be predicted from intonation data.

	dec	int	imp	SUM	%
action-dir	6	7	1	14	13.9
info-request	0	26	0	26	25.7
mixed	1	6	1	8	7.9
other	52	1	0	53	52.5
SUM	59	40	2	101	
%	58.4	39.6	2.0		

Table 5: Relation between modalities and dialogue acts.

The final part of the intonation contour in Spanish (the toneme) is important to define the utterance modality and the pragmatic interpretation. This part is represented with the last INTSINT tags of every string. In the experiment the last five tags are taken in order to determine how many of them are required to predict modality with the highest accuracy.

Data showed that dialogue acts can be predicted by using modality and duration; however, modality is not available as a data in the speech, but it can be identified by using intonation and duration. In this section we describe a two-layer processing

architecture, where modality is predicted first, and this information is used to predict dialogue act.

For the experiment a decision tree using J48, a CART-style algorithm, and WEKA software were used. 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.

Initial trees to predict dialogue acts were created by using INTSINT tones and duration as predictors (without modality), but the accuracy was not satisfactory (average less than 70%). Other trees were created by using tones only as predictor (without duration and modality), and accuracy was also low (average less than 70%).

Forty-five decision trees were created to predict modality by using duration and tones as predictor data. The most representative tree had an accuracy of 85.1%, and Kappa of 0.70390, and it was obtained when using the last 2 tone tags; 10 fold cross validation was used to produce it. This tree is presented in Table 6, where the numbers in parentheses are the number of cases complying/not complying each rule.

Utterance duration was not useful to predict modality. Also, accuracy decreased considerably (it was less than 77%) when using the last 3, last 4 or last 5 INTSINT tones as predictors. The confusion matrix of the tree is presented in Table 7. Dec had the highest prediction accuracy (52 against 7+0) and then int (34 against 5+1). Imp utterances were too scarce to produce a concluding result.

The best of the 45 trees to predict modality had accuracy of 100.0% and Kappa of 1.0; the last 1 INTSINT tag was used to create it. It was built by training with a representative subset of 70% and testing with the remainder.

MODALITY	RULES
int	If last_2 = UT, then int (20.0/1.0)
	If last_2 = BT, then int (15.0/5.0)
	If last_2 = LT, then int (3.0)
	If last_2 = BU, then int (2.0)
	If last_2 = LU, then int (2.0/1.0)
	If last_2 = HD, then int (1.0)
	If last_2 = MT, then int (1.0)
	If last_2 = TL, then int (1.0)
dec	If last_2 = DB, then dec (20.0)
	If last_2 = TB, then dec (15.0/1.0)
	If last_2 = MB, then dec (9.0)
	If last_2 = DD, then dec (1.0)
	If last_2 = DL, then dec (2.0)
	If last_2 = UH, then dec (2.0)
	If last_2 = LL, then dec (1.0)
	If last_2 = TD, then dec (1.0)
	If last_2 = HL, then dec (1.0)
If last_2 = UM, then dec (1.0)	
imp	If last_2 = HB, then imp (3.0/1.0)

Table 6: Tree for predicting modality.

In the second stage of the experiment, the predictor data were INTSINT cleansed strings (taking the last 5, last 4, last 3, last 2, and last 1 labels from every string), in addition to duration and modality; 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 whole data were used, where 70% was for training and 30% for testing; 2) subsets which were randomly defined but not strictly representative 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 they were not strictly representative. The combination of different attributes and training/testing modes permitted the creation of forty-five decision trees.

CLASSIFIED AS \ ACTUAL	int	dec	imp
int	34	5	1
dec	7	52	0
imp	2	0	0

Table 7: Confusion matrix of the tree in Table 6 to predict modality.

The tree produced in mode 1, using the last 1 INTSINT tag, produced promising results: accuracy of 80.6% and Kappa of 0.6880 (average). This could be considered the most representative tree of all 45 because its training and testing sets were created by a manual stratified sampling; also, because the results of this tree are approximately the average of all trees; this tree is presented in Table 8.

Modality and duration were the useful attributes to predict dialogue act, while INTSINT tags were not. This is evident by observing that no INTSINT attribute is in the tree. Its confusion matrix is presented in Table 9 and depicts that *info-request* is the dialogue act which had the highest classification accuracy (8 against 0+0+0), then *other* (15 against 0+1+0), then *action-dir* (2 against 0+2+0) and finally *mixed* (0 against 0+1+2).

The best tree of all 45 for predicting dialogue acts had an accuracy of 82.1%, while the average of all 45 trees was 79%; Kappa was 0.7128 for this tree and 0.6640 was the global average. The tree was obtained by using the last 3 tags of INTSINT sequences and by training and testing with the full data set on the 10-fold cross validation mode.

DIALOGUE ACT	RULES
info-request	if modality=interrogative and duration≤3103.125, then information request (18.0)
action-dir	if modality=interrogative and duration>3103.125, then action directive (7.0/3.0)
	if modality=declarative and duration>1276 and duration≤2106.9375, then action directive (5.0/2.0)
	if modality=imperative, then action directive (2.0/1.0)
other	if modality=declarative and duration≤1276, then other (27.0)
	if modality=declarative and duration>2106.9375, then other (11.0/2.0)

Table 8: Tree for predicting dialogue acts.

CLASSIF. AS ACTUAL	other	action-dir	info-request	mixed
other	15	0	1	0
action-dir	0	2	2	0
info-request	0	0	8	0
mixed	0	1	2	0

Table 9: Confusion matrix of the tree in Table 8 to predict dialogue acts.

Although few data were available (one dialogue only) we consider that these preliminary results seem to be promising. Results depict that identifying modality to identify dialogue act could be a useful method to be evaluated in a prototype dialogue management system. Other interesting attributes to be evaluated in experiments for the short term are speaker type (user or Wizard), and dialogue act tag corresponding to the previous utterance.

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 greater amount of attributes and data will be available.

6 Discussion and Further Work

The present methodology promises a simple and efficient way to identify dialogue act types for the construction of dialogue managers for practical dialogues; the present investigation will be continued with experiments focusing on the identification of other obligation acts and also common ground dialogue acts, and then 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 Llisterrí and Monserrat Riera at Universitat Autònoma de Barcelona. We also thank Hayde Castellanos, Varinia Estrada, Fernanda López, Isabel López, Ivan Meza, Iván Moreno, Patricia Pérez, Carlos Rodríguez, Issac Castillo, Javier Cuétara, Laura Irene Gonzáles, Ivonne López, Laura Lorena Rosales and Arturo Wong 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

- James Allen and Mark Core. 1997. *Draft of DAMSL: Dialog Act Markup in Several Layers. Technical report*, The Multiparty Discourse Group. University of Rochester, Rochester, USA, October.
- Jean Carletta. 1996. *Assessing agreement on classification tasks: the kappa statistic*. Computational Linguistics, 22(2):249-254.
- Robert Espesser. 1999. *M.E.S. Motif Environment for Speech*. http://www.lpl.univ-aix.fr/ext/projects/mes_signal.htm
- Eibe Frank, Mark Hall and Len Trigg, 2004. *WEKA. Waikato Environment for Knowledge Analysis*. <http://www.cs.waikato.ac.nz/~ml/weka>
- Daniel Hirst, Albert Di Cristo and Robert Espesser. 2000. *Levels of representation and levels of analysis*

- for the description of intonation systems.* in M. Horne (ed) *Prosody: Theory and Experiment* (Kluwer, Dordrecht).
- Daniel Hirst, and Robert Espesser. 1993. *Automatic modeling of fundamental frequency using a quadratic spline function.* CNRS (URA 261), Institut de Phonétique d'Aix, Université de Provence.
- Luis Pineda, Antonio Massé, Iván Meza, Miguel Salas, Erick Schwarz, Esmeralda Uruga and Luis Villaseñor. 2002. *The DIME Project.* In *Lecture Notes in Artificial Intelligence 2313*, Springer, pp. 166-175.
- Luis Pineda, Haydé Castellanos, Sergio Coria, Varinia Estrada, Fernanda López, Ivonne López, Iván Meza, Iván Moreno, Patricia Pérez and Carlos Rodríguez. 2005. *Balancing Transactions in Practical Dialogues. Technical report.* Department of Computer Science, IIMAS, UNAM, México.
- Shriberg, E., R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer y C. Van EssDykema. 1998. *Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?*. *Language and Speech* 41(3-4): 439-487, Special Issue on Prosody and Conversation.
- Villaseñor, L., A. Massé y L. Pineda. 2001. *The DIME Corpus.* ENC 01, 3er Encuentro Internacional de Ciencias de la Computación, SMCC-INEGI, Aguascalientes, Mexico.
- Witten, I. y E. Frank. 2000. *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan-Kaufman Publishers. San Francisco, CA. USA: 89-97.