DIALOG ACT LABELING IN THE DIHANA CORPUS USING PROSODY INFORMATION

Vicent Tamarit, Carlos-D. Martínez-Hinarejos

Instituto Tecnológico de Informática, Universidad Politécnica de Valencia
Camino de Vera s/n, 46022, Valencia, Spain

ABSTRACT

We propose a dialog act classification based on the prosody of the audio signal in combination with the course of the dialog. The work is applied to the Spanish corpus DIHANA. As far as we know, it is the first experiment made with prosody in this corpus. To do the labeling, we used two features that had been extracted from the user speech (pitch and energy) in a HMM classifier combined with an n-gram of dialog acts. The results shows a sightly improvement in the tagging when prosody is included in the classification.

1. INTRODUCTION

In a speech based dialog system, it is necessary to recognize the user's speech and to understand the true meaning of the uttered sentence. Both of them are used by the machine to generate an appropriate response. When obtaining the relevant information that aids the system, the speech is segmented into utterances (the minimal significant unit from the dialog viewpoint) each of which is labeled with a dialog act (DA). Dialog acts typically represent types of sentences or communicative intentions. Common dialog acts are: question, answer, response,... One of the research fields in dialog systems is the identification of dialog acts.

One way to extract dialog acts from speech is using the automatic speech transcription, so the recognition accuracy may affect the correct labeling. To improve the tagging other authors have proposed the use of some prosody features that can identify different types of sentences. On the one hand, this method has one important advantage: it could be used before the speech recognition, and the dialog act identification may aid the speech recognizer in the recognition of the words; but, on the other hand, the signal is more difficult to interpret than the transcription.

Some studies have proved the influence of prosody in dialog acts identification. In [1], results are presented for the SwitchBoard corpus, based on spontaneous conversations between English speakers. These results show an improvement on the DA identification when using prosodic features. The CallHome Spanish corpus, with telephonic conversation in Latin American Spanish, has been used in a similar test [2]. In this last work, pitch and energy features are computed to classify acts through Support Vector Machines (SVM).

In this article we describe the results of dialog act labeling in the Spanish corpus DIHANA, using pitch and energy values. This corpus is recorded only by Spanish speakers, seeking those who do not have a strong accent. Instead of SVM or K-Nearest Neighbours (K-NN) techniques, that do not capture the continuity of the features, we used Hidden Markov Models (HMM) with gaussian output distributions, in a similar way to the speech recognition process. Furthermore, we improved the prosodic classification with a dialog act n-gram.

2. DESCRIPTION OF THE CORPUS

The Spanish corpus DIHANA [3] is composed of 900 dialogs about a telephonic train information system. It was acquired by 225 different speakers (153 male and 72 females), with small dialectal variants. There are 6,280 user turns and 9,133 system turns. The vocabulary size is 823 words. The total amount of speech signal was about five and a half hours.

The acquisition of the DIHANA corpus was carried out by means of an initial prototype, using the Wizard of Oz (WoZ) technique [4]. This acquisition was only restricted at the semantic level (i.e., the acquired dialogs are related to a specific task domain) and was not restricted at the lexical and syntactical level (spontaneous-speech). In this acquisition process, the semantic control was provided by the definition of scenarios that the user had to accomplish and by the WoZ strategy, which defines the behaviour of the acquisition system.

The annotation scheme used in the corpus is based on the Interchange Format (IF) defined in the C-STAR project [5]. Although it was defined for a Machine Translation task, it has been adapted to dialog annotation [6]. The three-level proposal of the IF format covers the speech act, the concept, and the argument, which makes it appropriate for its use in task-oriented dialog.

Based on the IF format, a three-level annotation scheme of the DIHANA corpus utterances was defined in [7].

WORK SUPPORTED BY THE EC (FEDER) AND THE SPANISH MEC UNDER GRANT TIN2006-15694-CO2-01.

This DA set represents the general purpose of the utterance (first level), as well as more precise semantic information that is specific to each task (second and third levels).

All of the dialogues are segmented in turns (User and System), and each turn is also segmented into utterances. Finally, each utterance is labelled with a three-level label. Obviously, more than one utterance can appear per turn. In fact, an average of 1.5 utterances per turn was obtained.

Only the first level contains linguistic information that can be learned by a prosodic classifier, so we preprocessed the audio corpus to cut the turns into first level utterances. The final tags we used were: Afirmación (Yes-answer), Negación (No-answer), Pregunta (Question), Respuesta (Generic Answer), Cierre (End dialog), Indefinida (No tagged). The $42\,\%$ of the first level utterances are questions; thus, our baseline classification error is $58\,\%$. There are 7,373 first level utterances. We used, on average, 6,118 for train and 1,255 for test.

3. MODELING

Other authors have used some different techniques to classify DAs from prosodic cues, like decision trees [1], neural networks [8], and SVM[2]. We investigated the performance of HMMs for classification based on prosody.

3.1. Feature extraction

There is a classification of sentences in Spanish based on the speaker intonation [9]. The intonation of the sentences are quite different, i.e., for a question or for a statement. In Figures 1 and 2, we show the pitch evolution for one sentence with different intonations.

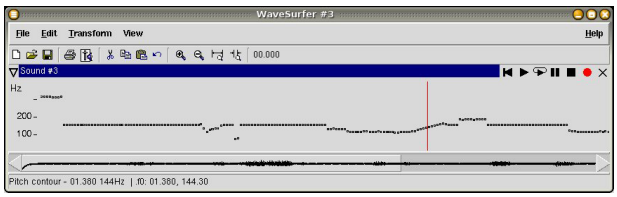

Figure 1. Pitch evolution for the sentence ”Se puede ir desde Santurce a Bilbao” recorded like a statement.

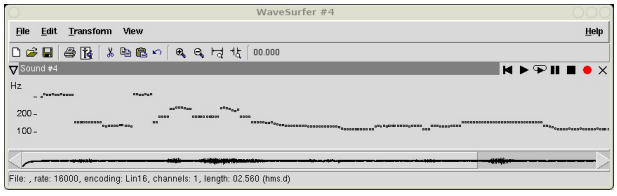

Figure 2. Pitch evolution for the sentence ”Se puede ir desde Santurce a Bilbao” recorded like a question.

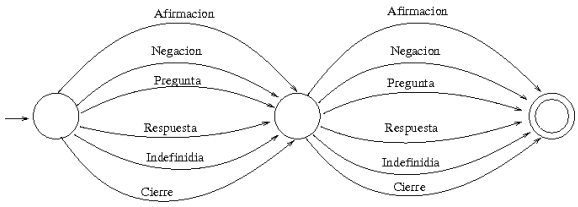

Figure 3. Finite State Machine used as language model in the task.

The feature extraction is based in only two features: energy measure and pitch. For every 10 ms of signal, we computed the frame energy and an estimation of the F0. We used the Snack library [10], developed by the Department of Speech, Music and Hearing, in the Royal Institute of Technology in Sweden, to estimate the fundamental frequency. The purpose of the library is to develop in a short time sound tools using scripting languages such as Tcl/TK or Python. In addition to these two values, we computed the first and second derivative of the features. Therefore, we obtained vectors with six elements.

3.2. Hidden Markov Models

The HMMs with gaussian distributions in the states are used in speech recognition to model the sound units, usually phonemes. Their structure allows them to model the time-variation of the features so they can represent the prosodic variation in time.

We used a three-state HMM for each dialog act. This structure was selected due to the minimum number of vectors obtained from the audio signal. They were trained using the HTK software [11].

The decodification process was made using iATROS. This recognition software was developed in the PRHLT Group in the Instituto Tecnológico de Informática. It is based on the Viterbi algorithm and uses three models:

- Acoustic models: Each model represents a phonetic unit as a continuous HMM. In our task, we have one acoustic model for each dialog act which represents the prosodic variation.
- Lexical model: Each word is described as a Finite State Machine (FSM), that defines the acoustic models that compose the word. In our case, no words are actually defined. Therefore, each lexical model correspond to an only acoustic model.
- Language Model: Defines the relations between the words. We used a FSM to model the structure of a turn. Figure 3 shows the model we used. It has three states, since the utterances have two dialog acts at most. The transition probabilities between states are equal for all the edges.

```
Pregunta|Respuesta Indefinida 0.02 -3.9121
Pregunta|Respuesta Afirmacion 0.22 -1.5141
Pregunta|Respuesta Pregunta 0.34 -1.0788
Pregunta|Respuesta Respuesta 0.27 -1.3093
Pregunta|Respuesta Cierre 0.07 -2.6569
Pregunta|Respuesta Negacion 0.07 -2.6569
```

Figure 4. Some estimations of the 3-gram used in this task. The second number is the log-probability.

Obviously, some turns have only one tag. We used partial decodification to solve this problem and allow decodification with only one utterance.

3.3. Word Graphs

A word graph (WG) is a direct graph, no cyclical and with weights, where each node represents a discrete point in time. The edges of the graph are a set [w, s, e], where w is the hypothetic word from node s to e. The weights are scores associated to the edges. The best path from the initial state to the final state is the most likely hypothesis. In short, a WG is like a "picture" of the recognition process.

The word graph is necessary to add n-gram information to the model. The n-gram is estimated using the sequence of dialog acts in the dialogues, i.e., using system and user turns. In Figure 4 is showed a example of 3-gram. However, recognition is only performed on user turns (audio records from system utterances does not exist). Therefore, after the recognition process we obtain a WG with the calculated acoustic probabilities and equal language model probabilities, which includes all possible dialog act sequences. We can incorporate the n-gram information by changing the language model probabilities in the WG by the corresponding n-gram probabilities. After this change, we searched the best path in the WG using the combination of the acoustic and the new language model probabilities.

In Figure 5 there is an example of a word graph for this task. In this case there is only one dialog act, and the graph shows us the probabilities for each class. In the shown WG, lets assume that the previous dialog acts for that turn were "Pregunta" and "Respuesta". The new WG with n-gram probabilities (using the probabilities of Figure 4) is showed in Figure 6.

4. RESULTS

To obtain significant results in the labeling task with the DIHANA corpus, a cross-validation approach was adopted and 5 different partitions were used. Each of them had 720 dialogues for training and 180 for testing. The statistics for the corpus are presented in Table 1.

For each partition we combined the prosody classification with a 3-gram trained with the utterances' evolution within the dialogs. The 3-grams included all the utterances (user and system), because in a real dialog system we always know the tag of the previous system turn.

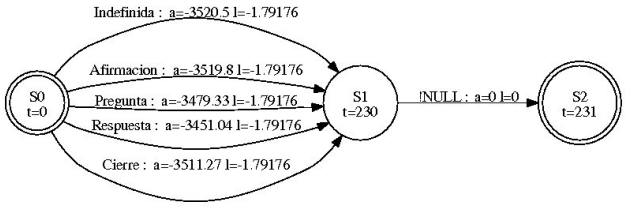

Figure 5. Example of word graph for this task. Each edge is represented with the label and the log-probability of the acoustic model (a) and language model (l).

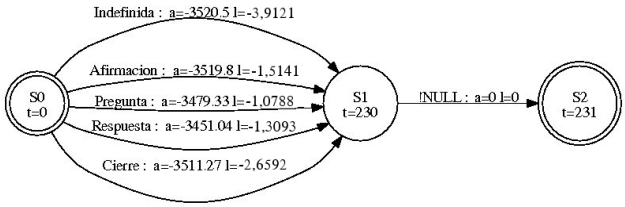

Figure 6. The original language model probabilities are replaced by the n-gram ones.

Table 2 shows the results of the experiments. We included the tagging using only the calculated 3-grams, and the combination with prosody. In this table, the Word Error Rate (WER) measures the accuracy of the act labeling, whereas the Sentence Error Rate (SER) shows the accuracy of the whole turn tagging.

Labeling acts using the 3-gram produced an improvement of 20 points from the baseline (that we fixed in 58 %). The inclusion of prosody information reduces the WER only in one point.

	Training		
	User	System	Total
Dialogues	720		
Turns	5,024	7,206	12,330
Running words	42,806	119,807	162,613
Vocabulary	762	208	832

	Test		
	User	System	Total
Dialogues	180		
Turns	1,256	1,827	3,083
Running words	10,815	29,950	40,765
Vocabulary	417	174	485

Table 1. DIHANA corpus statistics (average of the five cross-validation partitions).

WER/SER	3-gram	Combined
Partition 1	41.4/40.2	40.5/39.3
Partition 2	43.5/42.2	42.8/41
Partition 3	40.8/39	39.8/37.6
Partition 4	40.9/39.5	40.6/39.4
Partition 5	34.5/33.2	32.4/30.7
Total	40.3/38.9	**39.3/37.6**

Table 2. Results for the experiments in the five partitions.

5. CONCLUSIONS AND FUTURE WORK

The corpus DIHANA has a labeling oriented to the human-machine interaction. This tagging is useful for the system to understand the requests and generate a response, but it is not based on the intonation of the sentence. This task-oriented labeling could be the reason of the little improvement in the classification using our prosody-based classifier. As far as we know this is the first time prosody is used in the dialog act classification in the DIHANA corpus, so we can not conclude that prosody does not improve the dialog act tagging, as more experiments should be performed.

Future work is directed to improve the intonation extraction, as can be seen in [12], and test new prosody features in this corpus, such as those proposed in [13], as well as other classification techniques like K-NN, neural networks or decision trees, which are proved in other corpora, but not in DIHANA. The classification structure based on HMMs could be applied on other corpora like CallHome or SwitchBoard. These corpora are annotated with a different set of dialog acts that could be more suitable for the prosody-based classifier. The use of a Spanish corpus annotated with labels based on the intonation of the sentences may help us to determine the utility of the prosody in Spanish. Restructuring the dialog acts in DIHANA is another possibility.

6. REFERENCES

[1] E. Shrinberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?," *Language and Speech. Special Issue on Prosody and Conversation*, vol. 41 (3-4), pp. 439–487, 1998.

[2] Raul Fernandez and Rosalind W. Picard, "Dialog act classification from prosodic features using support vector machines," *Speech Prosody 2002, International Conference*, pp. 291–294, 2002.

[3] J.-M. Benedí, E. Lleida, A. Varona, M.-J. Castro, I. Galiano, R. Justo, I. López de Letona, and A. Miguel, "Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana," *Fifth International Conference on Language Resources and Evaluation (LREC)*, pp. 1636–1639, May 2006.

[4] M. Fraser and G. Gilbert, "Simulating speech systems," *Computer Speech and Language*, , no. 5, pp. 81–89, 1991.

[5] Lavie A., L. Levin, P. Zhan, M. Taboada, D. Gates, M. M. Lapata, C. Clark, M. Broadhead, and A. Waibel, "Expanding the domain of a multi-lingual speech-to-speech translation system," *Proceedings of the Workshop on Spoken Language Translation, ACL/EACL-97*, 1997.

[6] T. Fukada, D. Koll, A. Waibel, and K. Tanigaki, "Probabilistic dialogue act extraction for concept based multilingual translation systems," *ICSLP 98*, pp. 2771–2774, 1998.

[7] N. Alcácer, J. Benedí, F. Blat, R. Granell, C. D. Martínez, and F. Torres, "Acquisition and labelling of a spontaneous speech dialogue corpus," *Proceeding of 10th International Conference on Speech and Computer (SPECOM). Patras, Greece*, pp. 583–586, 2005.

[8] Finke M. amd Lapata M., "Clarity:infering discourse structure from speech," *Proc AAAI '98 Spring Symposium on Applying Machine Learning to Discourse Processing*, 1998.

[9] Antonio Quilis and Joseph A. Fernández, *Curso de fonética y fonología española*, CSIC, 1993.

[10] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," *Proc of ICSLP, Beijing*, pp. 464–467, October 2000.

[11] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, CUED, UK, v3.2 edition, July, 2004.

[12] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-tür, and Gökhan Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, pp. 127–154, 2000.

[13] A. Stolcke, N. Coccaro, R. Bates., P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer, "Dialogue act modelling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 1–34, 2000.