Dialog Act Classification from Prosodic Features Using Support Vector Machines.

Raul Fernandez & Rosalind W. Picard

MIT Media Laboratory
20 Ames Street. E15-120F
Cambridge, MA 02139
USA
{raul; picard}@media.mit.edu

Abstract

In this work we investigate the use of support vector machines (SVMs) and discriminative learning techniques on the task of automatic classification of dialogue acts (DAs) from prosodic cues. We implement and test these classifiers on solving an 8-DA classification task on the Spanish CallHome database and report preliminary recognition rates of 47.3% with respect to a 20.4% chance-level rate, which represents an improvement over previously reported work using decision trees and neural network classifiers. Although prosodic cues alone may not suffice for robust classification of DAs, we report results that suggest that SVMs offer an interesting alternative to previously explored models, and should be further explored to improve the contribution of prosodic models to the classification task.

1. Introduction

The automatic parsing and classification of spoken language into discourse structures is a task of fundamental importance for artificial systems that aim to achieve natural language understanding. Particularly in non-command-driven and non-task-oriented dialogue interactions, it is desirable to have a system that can make sense not just of the sequence of words uttered by a speaker, but also of the role they play in guiding the structure and course of the dialogue. The theory of speech acts provides a framework for analyzing the structure of dialogues in terms of self-contained units which convey information about the speaker's (hearer's) attitude with respect to the flow of the conversation, his understanding, intentionality, etc. Developing techniques for automatically identifying and classifying speech acts is therefore an important goal of any system aiming to understand the structure of spoken language, be it for communication in a human-machine interaction scenario, or for applications that involve browsing human-human dialogue and extracting relevant information (e.g, data retrieval).

Several studies, [3], [10], [12], [9], [11], have investigated different approaches for dialog act modeling. These approaches usually combine models that capture information about lexical constituency of speech acts, their prosodic realization, as well as the dialog act sequence to arrive at a parsing of a dialog into dialog acts (DAs) and a classification of these units. The results reported in these works strongly suggest that combining the outputs of these independent models improves the results otherwise obtained with any single model. In particular, prosody has been shown to aid in the task of disambiguating between speech acts that have similar lexical realization [10].

Although the contribution of lexical models (e.g., n-gram language models) to the overall classification task is typically greater than the contribution of a prosodic module [12], there is still a strong motivation for investigating alternate models that we can apply to improve DA classification from prosodic cues alone. Not only can an improved prosodic model contribute to improving the combined DA classification rate, they can also help when the performance of a lexical model (which in the fully automatic case would rely on the output of a speech recognizer) degrades in the presence of incorrectly transcribed words from speech. The purpose of this investigation is, therefore, not to cast the problem of DA identification as one of classification exclusively from prosodic features –since there may be upper bounds on what is achievable from prosody alone– but rather to investigate other machine learning algorithms different from the ones proposed in the literature (which thus far have been limited primarily to decision trees and artificial neural networks) for improving the recognition rates of a prosodic module. By doing this we can also examine how much prosody can contribute to this classification task, and gain some insight into what this upper bound might be. Although we do not address this problem here, we expect that the results of a learning algorithm for modeling DAs from prosodic features will have to be combined in parallel with the outputs of other models to arrive at a more robust classification scheme.

2. Dialog Act Tags

To investigate the automatic classification of DAs from prosodic cues, we have used extracts from the CallHome Spanish database and the corresponding set of manual annotations developed as part of the CLARITY project at Carnegie Mellon [3]. This corpus is tagged using a three-level coding scheme which attempts to describe the discourse structure using a hierarchical system. At the highest level, dialog segments are described in terms of *activities*. An activity is a portion of the dialog focusing on the purpose and goal of the speakers within a given dialog topic. The intermediate level, *dialog games*, provides a description of turn-taking or exchange sequences that shows how two dialog participants relate to each other. The lowest level of this description consists of *speech act* annotations. The speech act convention adopted in this work describes utterances (or portions thereof) in terms of the following categories [7].

- *Questions*: describes acts that follow the form *and/or* the intentionality of a question.

- *Answers*: describes primarily answers to Yes/No questions.
- *Agreement/Disagreement*: describes acts used to accept or reject statements made by the other speaker.
- *Discourse Markers*: subsumes *backchannels* (acts used primarily to convey understanding or paying attention).
- *Forward Functions*: includes acts such as exclamations, apologies, formulaic wishes, thankings, etc.
- *Control Acts*: includes acts that expect an action on the part of the hearer or speaker; e.g., commands, requests, prohibitions, etc.
- *Statements*: encompasses opinion and non-opinion statements.
- *Other*.

These speech act tags are further refined in this taxonomy. For instance, the category *Statement* can also include descriptions about the speaker's attitude, the hypotheticality of a statement, doubt or uncertainty. However, discriminating between this extended set of tags is a task that lies beyond the scope of this investigation. In this work we investigate how well prosodic cues (alone) can discriminate these 8 tags.

3. Modeling

The classification models reported in the literature for the task of classifying DAs from prosodic cues include decision trees and neural networks [10], [3]. We investigate the use of a *newer* model, support vector machines, for this task.

3.1. Support Vector Machines

A support vector machine (SVM) implements an approximation to the structural risk minimization principle in which both the empirical error and a bound related to the generalization ability of the classifier are minimized. The SVM fits a hyperplane that achieves maximum margin between two classes, and its decision boundary is determined by the discriminant

$$f(\mathbf{x}) = \sum_i y_i \lambda_i K(\mathbf{x}, \mathbf{x}_i) + b \qquad (1)$$

where $\mathbf{x}_i$ and $y_i \in \{-1, 1\}$ are the input-output pairs, $K(\mathbf{x}, \mathbf{y}) \doteq \phi(\mathbf{x}) \cdot \phi(\mathbf{y})$ is a kernel function which computes inner products, and $\phi(\mathbf{x})$ is a transformation from the input space to a higher dimensional space. In the linearly separable case, $\phi(\mathbf{x}) = \mathbf{x}$. An SVM is generalizable to non linearly separable cases by first applying the mapping $\phi(\cdot)$ to increase dimensionality and then applying a linear classifier in the higher-dimensional space. The parameters of this model are the values λ_i, non-negative constraints that determine the contribution of each data point to the decision surface, and b, an overall bias term. The data points for which $\lambda_i \neq 0$ are the only ones that contribute to (1) and are known as support vectors.

Fitting an SVM consists of solving the optimization [8]:

$$\begin{array}{rrcl} \max & F(\Lambda) & = & \Lambda \cdot \mathbf{1} - \frac{1}{2}\Lambda \cdot D\Lambda \\ \text{subject to} & \Lambda \cdot \mathbf{y} & = & 0 \\ & \Lambda & \leq & C\mathbf{1} \\ & \Lambda & \geq & \mathbf{0} \end{array} \qquad (2)$$

where $\Lambda = [\lambda_1 \cdots \lambda_l]'$ and D is a symmetric matrix with elements $D_{i,j} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, and C is a non-negative constant that bounds each λ_i, and which is related to the width of the margin between the classes. Having solved Λ from the equations in (2), the bias term can be found:

$$b = -\frac{1}{2} \sum_i \lambda_i y_i \Big(K(\mathbf{x}_-, \mathbf{x}_i) + K(\mathbf{x}_+, \mathbf{x}_i) \Big) \qquad (3)$$

where $\mathbf{x}_-$ and $\mathbf{x}_+$ are any two correctly classified support vectors from classes -1 and $+1$ respectively [4]. In the work reported here we have use a Gaussian kernel of the form $K(\mathbf{x}, \mathbf{y}) = e^{\frac{||\mathbf{x}-\mathbf{y}||^2}{\sigma}}$ (where σ is a free parameter which was optimized on the training set).

3.2. Prosodic Features

The feature set used in these experiments is derived from the F0 and energy contours of a segmented speech act. The F0 extraction algorithm, described at length in [1], implements a normalized autocorrelation method with a Gaussian window to extract a set of pitch candidates, and uses dynamic programming as a post-processing step to select the best sequence of pitch values by suitably defining an optimization function which penalizes large octave and voicing-to-unvoicing transitions.

The feature set includes the following measurements related to pitch, energy, and duration. Let F0, F0', and F0" be the pitch estimate and its first and second differences, and EC, EC', and EC" the corresponding measurements from the energy contour.

Pitch Features

- Unbiased estimate of F0 variance
- Unbiased estimate of F0 skewness
- Unbiased estimate of F0 kurtosis
- F0 interquartile range
- F0 range
- Difference between max(F0) and the F0 sample mean
- Difference between the F0 sample mean and min(F0)
- Unbiased estimate of F0' variance
- Unbiased estimate of F0" skewness
- Unbiased estimate of F0" kurtosis
- F0' range
- Unbiased estimate of F0" variance
- Unbiased estimate of F0" skewness
- Unbiased estimate of F0" kurtosis
- F0" range

Energy Features

- EC sample mean.
- Unbiased estimate of EC variance
- Unbiased estimate of EC skewness
- Unbiased estimate of EC kurtosis
- EC interquartile range
- Difference between max(EC) and the EC sample mean
- Unbiased estimate of EC' variance
- Unbiased estimate of EC' skewness
- Unbiased estimate of EC' kurtosis
- EC' range

- Unbiased estimate of EC" variance
- Unbiased estimate of EC" skewness
- Unbiased estimate of EC" kurtosis
- EC" range

Duration Feature

- Length of voiced portions in F0

By looking only at voiced portions, the duration feature automatically excludes pauses. Although this feature may be dependent on the particular segmental content of an utterance, it is expected that for most utterances (particularly longer ones), it will be proportional to the duration of the articulated portions of an utterance (assuming there's roughly a constant voiced-to-unvoiced ratio across utterances).

4. Discussion and Results

Automatic dialog act classification is a learning task that involves very non-equal class priors: Some of the DA categories, such as *Discourse Markers* or *Questions*, are better represented in the CallHome Spanish corpus than others. Since this is a corpus of non-scripted spontaneous speech, we have decided to attempt to model the priors, rather than selecting an equal number of samples for learning each category. For training the model and evaluating its performance, we constructed independent training and testing sets of DAs by sampling various dialogues from the database, trying to include a variety of speakers and dialectal differences. The number of DAs included in each set approximately reflected the frequency of occurrence in the corpus, and both training and testing sets were designed in proportion to reflect these priors. To evaluate the performance of the algorithm we have compared the recognition errors against the Bayes error $B = \sum_i P_i(1 - P_i)$, where P_i is the i^{th} class prior estimated from the frequencies in the training set (in what follows, the figure reported as chance is $1 - B$).

The task that we have considered in this investigation is that of classifying a DA *given* a segmentation of continuous speech into unknown DA units. In other words, we have not considered the problem of automatic parsing into DAs. Although prosodic cues may also be used to segment speech into these units as discussed in [3], we have assumed an error-free segmentation so as to be able to assess the classification results independently of the segmentation results (and the errors that may occur at this stage). However, it is worth observing that a segmentation algorithm is likely to incur errors, and that the values in the feature set may be sensitive to a faulty segmentation. In particular, duration features (such as the length of a DA) are likely to vary considerably when the segmentation is not properly done. Since the work reported in [10] and [12] has shown that duration features play an important role in properly classifying DAs, we also investigate whether and how the results obtained with the baseline set can change when duration features are absent.

Table 1 shows the overall recognition rate on the 8 DA classification task for both the training and testing sets when the duration measure was included and left out of the training phase. The overall baseline recognition of 47.3% represents a relative increase of 56.9% with respect to the chance rate of 20.4%, and is comparable to what is reported in [3] for the CallHome Spanish database using additional 4-gram word models and 1-gram discourse grammars (a 48% rate on a 26%-chance task is reported there). This rate also represents an improvement over the 38.9% recognition rate using prosody alone on a 35%-chance task reported in [12] for the SWITCHBOARD database.

	Training (%)	Testing (%)
No Duration	64.8	42.0
Duration	63.7	47.3

Table 1: Recognition Rates on the Classification Task for 8 DA Classes (Chance = 20.4%)

(Cross-database comparisons, particularly if they involve cross-language comparisons, should be made with caution, however, as these two corpora may be prosodically quite different.) Although the recognition rate on the test set decreased when duration information was omitted from the basic feature set, this change is not statistically significant given the number of DAs used in the test set ($p < 0.5$). A larger sample set would be needed in order to further qualify whether omitting duration features adversely affects the performance of the SVM models.

Although the overall recognition rate exceeds the chance classification rate, the model did not succeed in modeling all categories equally well. In particular, there was overlap between the *Statements* and *Questions* category. At first, this is a somewhat surprising result considering that Spanish exploits intonational features to encode the difference between otherwise lexically equivalent statements and questions. The *Questions* category, however, encompasses not just utterances that follow the form and intention of a question; it is also used to label utterances where a question may be implied (though perhaps not in the form of a question) as well as utterances that follow the form of a question when no answer is expected (e.g. rhetorical). To rule out interference from these last two cases, the model was also trained and tested using only utterances in the *Questions* category that followed both the form and intention of a question. Although this did not solve the problem, it seems that most of the missclassifications arise from *Open-ended Questions* being confused with *Statements* and *Discourse Markers*. *Open-ended Questions* are a large subset of the *Questions* category. Some of the examples used in the training database are

- *¿Qué tal? (How is it going?)*
- *Pero tú, ¿cómo te sientes tú?*
  (And you, how are you feeling?)
- *¿Y cómo le va a la Candice en el colegio?*
  (How is Candice doing in school?)

Open-ended Questions can be realized by a speaker without the distinctive pitch patterns that are commonly the trademark of other types of questions (e.g. *Yes/No Questions*). We hypothesize that this may be a source of confusion for this model, and suggest that this issue deserves further consideration.

The model was able to provide better classification rates for the *Statements* category. Table 2 shows the confusion matrix for a *Statements* detection subtask, where the trained model is used to isolate this category from the rest. The chance classification rate is 67% in this case, and with this system we are able to correctly classify 77% of the speech acts. When omitting duration information from the feature set (Table 3), the classification rate drops to 71%, which is a considerable reduction (although not statistically significant given the number of samples in the set). The reduction in performance when duration is missing is consistently more noticeable in the testing set (see Tables 1 and 3). This suggests that duration features may play an important role in the generalization ability of the model (the ability of the model to maintain its performance on a set of unseen data).

	Training		Testing	
	Statements	Rest	Statements	Rest
Statements	29	11	23	17
Rest	19	134	14	77

Table 2: Confusion Matrix for Statements *Classification Task. (Duration Information Included)*

	Training (%)	Testing (%)
No Duration	84	71
Duration	84.5	77.0

Table 3: Recognition Rates for Statements *Classification Tasks (Chance = 67%)*

4.1. Feature Selection

It is clear that several of the features in the set described in Section 3.2 are not independent measures. In addition to the experiments with the baseline feature set, we investigated whether we could reduce the dimensionality of the feature set and eliminate colinear features without affecting the performance by applying Principal Component Analysis (PCA) to this data set. PCA finds a linear transformation to project the data onto a lower dimensional space. The *representation error* (not to be confused with the classification error reported in the paper previously) incurred in this transformation is related to the eigenvalues of the covariance matrix of the data set. Hence, by properly monitoring the rate of decay of the eigenvalues, one can choose a suitable projection onto a lower dimensional space that is still able to explain much of the variance in the data. (See [2] for PCA details.) By applying PCA we were able to reduce the dimensionality from 30 to 18 and still represent 99.5% of the data's variance in this lower space. (Note that this method does not *pick out* the best features, but rather builds a new set by taking linear combinations of the original set.) This reduction in dimensionality at such low cost indicates the degree of colinearity of the original set. With PCA we are able to retain similar recognition rates as obtained with the baseline feature set. These results are summarized in Table 4 for the 8 DA classification, as well as the *Statements* classification subtask

	Training (%)	Testing (%)
All Classes	52.3	48.1
Statements	79.3	76.0

Table 4: Recognition Rates on the 8 DAs and Statements *Tasks with Reduced Feature Set (Chance = 20.4% and 67%)*

5. Conclusions

In this work we have explored the use of support vector machines for the classification of dialog acts from prosodic cues alone. Unlike other other approaches reported in the literature for solving this task (e.g., decision trees and neural networks), SVMs are trained discriminately not just to find a solution that minimizes the error on the training set, but one that has good generalization properties to data points not used while training. Although there seem to be clear bounds on what is achievable on DA classification from prosody alone, we report preliminary overall recognition rates on the CallHome Spanish database that represent an incremental improvement on previously reported methods for this corpus. Further consideration needs to be given to the use of more independent measures (not colinear) on the feature set to improve the method's performance, particularly on modeling individual categories of DAs for which there is still much overlap. Although the incremental change reported here suggests that we may be reaching the bound of what is possible to detect from prosody alone, SVMs offer a promising alternative that should be explored further to increase the performance of prosodic models and their contribution to more robust DA classification schemes that make use of lexical models.

6. Acknowledgements

We would like to thank Margaret Whitman for all her help with segmenting the CallHome Spanish database, Yuri Ivanov for many valuable discussions, and the Digital Life Consortium at the MIT Media Lab for their support.

7. References

[1] Boersma, P., 1993. Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound. *IFA Proceedings*, 17, 97-110.

[2] Duda, R.O.; Hart, P.E.; Stork, D.G., 2001. *Pattern Classification*, 2nd. Ed. New York: John Wiley.

[3] Finke, M.; Lapata, M; et al., 1998. CLARITY: Inferring Discourse Structure from Speech. In *Proc. AAAI '98 Spring Symposium on Applying Machine Learning to Discourse Processing.*

[4] Gunn, S., 1998. Support Vector Machines for Classification and Regression. Technical Report. Image, Speech and Intelligent Systems Group. University of Southampton.

[5] Jurafsky, D.; Shriberg, E.; Fox, B.; Curl, T., 1998. Lexical, Prosodic, and Syntactic Cues for Dialog Acts. In *ACL/COLING Workshop on Discourse Relations and Discourse Markers.*

[6] Levin, L.; Ries, K.; Thymé-Gobbel, A.; Lavie, A., 1998. Tagging of Speech Acts and Dialogue Games in Spanish Call Home. In *Proceedings of ACL-99 Workshop on Discourse Tagging.*

[7] Levin, L.; Thymé-Gobbel, A,; Lavie, A.; Ries, K.; Zechner, K., 1998. A Discourse Coding Scheme for Conversational Spanish. In *International Conference on Spoken Language Processing (ICSLP '98)*.

[8] Osuna, E. E.; Freund, R.; Girosi, F., 1997. Support Vector Machines: Training and Applications. A.I. Memo 1602/C.B.C.L. Paper 144. MIT.

[9] Ries, K., 1999. HMM and Neural Network Based Speech Act Detection. In *Proc. International Conf. Acoustics and Signal Processing (ICASSP '99)*.

[10] Shriberg, E.; Bates, R.; et al., 1998. Can Prosody Aid in the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech*, 41(3-4), 439-487.

[11] Stolcke, A.; Coccaro, N.; Bates, R.; Taylor, P.; Van Ess-Dykema, C., 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3), 339-373.

[12] Stolcke, A.; Shriberg, E.; et al., 1998. Dialog Act Modeling for Conversational Speech. In *Proc. AAAI '98 Spring Symposium on Applying Machine Learning to Discourse Processing.*