Detection of Questions in Arabic Audio Monologues Using Prosodic Features

Omair Khan
Department of Information and Computer Science
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
Email: umair@kfupm.edu.sa

Wasfi G. Al-Khatib
Department of Information and Computer Science
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
Email: wasfi@kfupm.edu.sa

Cheded Lahouari
Department of Systems Engineering
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
Email: cheded@kfupm.edu.sa

Abstract—Prosody has been widely used in many speech-related applications including speaker and word recognition, emotion and accent identification, topic and sentence segmentation, and text-to-speech applications. An important application we investigate is that of identifying question sentences in Arabic Monologue Lectures. Languages other than Arabic have received a lot of attention in this regard. We approach this problem by first segmenting the sentences from the continuous speech using intensity and duration features. Prosodic features are, then, extracted from each sentence. These features are used as input to decision trees to classify each sentence into either Question or Non Question sentence.

Our results suggest that questions are cued by more than one type of prosodic features in natural Arabic speech. We used C4.5 decision trees for classification and achieved 75.7% accuracy. Feature specific analysis further reveals that energy and fundamental frequency features are mainly responsible for discriminating between questions and non-question sentences.

I. INTRODUCTION

There has been a huge increase in the amount of data generated and stored as computers and Internet increasingly becoming part of our everyday life. This data is present in text as well as in audio and video format. Before the arrival of DSL and broad bandwidth connections, people used to limit themselves to text and image data. Now, with broader bandwidths available, there has been an increase in audio and video content on the Internet. Online video and audio sites such as Youtube, Google video, etc are among the most visited sites on the Internet. Audio and video content is widely shared through file sharing peer-to-peer networks. Multimedia content now constitutes the bulk of the Internet traffic in the form of IP-telephony, video and audio conferencing, Internet radio stations, music stores, lecture sites etc.

Audio data forms an important segment of this multimedia data present and transmitted through the Internet. It includes online lectures, music, radio programs, podcasts, news, Text-To-Speech(TTS) systems that translate textual websites into audio for visually impaired people etc. This content is present in downloadable as well as in streaming format. It is produced in a variety of languages. Most of it is in English and other Western languages because of the digital divide between the West and the Rest. Prominent Non-Western languages on the web include Japanese, Chinese, Korean, Turkish and Arabic. Arabic is the tenth most widely used language on the Internet [3].

In the Arabic language, huge repositories of lectures on the Islamic faith are available [1], [2]. An interesting problem in this regard is the semantic indexing of these lectures. This can be achieved by transcribing all the lectures manually or automatically through speech recognition techniques. However, neither solution is viable at this point. Manual transcription is a very laborious process requiring man power and time, and consequently comes with a high financial cost. Arabic speech recognition, on the other hand, has not matured enough to achieve high accuracy rates on such unrestricted domain. Therefore, one needs to look for other reasonable semantic content that can be automatically extracted. One such semantic content is that of questions posed within these lectures, whether by the speaker or the audience. In many such lectures, a question-answer session usually follows the main speech. These questions form a sizable and very useful knowledge-base on various issues, in which users may utilize. For example, one can look for the opinion of different scholars on a certain contemporary issue. Normally, the lectures are monologues, i.e. questions and answers are spoken by the same person. Therefore, the problem can be approached from a prosodic point of view, where the intonation of the voice is used for cues. We employ prosody to identify question segments in the speech in order to achieve high success rates. These segments can be later utilized by speech processing techniques in order to recognize important words and carry out the semantic indexing based on them.

In this paper, we report our findings with regard to identifying and employing prosodic features in detecting question and non-question segments in a monologue. The paper is organized as follows: Section II presents some background information that introduces the problem. Next, a literature survey is given in Section III. After that, Section IV presents the set of features considered in this study. This is followed by Section V describing the feature extraction process. The
classification technique employed and the experimental setup and results are shown in Sections VI and VII, respectively. Finally, our conclusions are drawn in Section VIII.

II. BACKGROUND

Prosody of a voice can be employed to differentiate a question from a statement. Prosody of a voice consists of its pitch, tone, pause duration, timing, loudness, intonation and other such acoustic or musical features of that voice. The distinctive variations of such features in a voice help a user to interpret utterances of a speaker. Native speakers use prosody to convey paralinguistic information such as emphasis, intention, attitude and emotion. Therefore prosody can be used in tasks such as identifying speech acts, topic segment boundaries, speaker emotions, locating and processing disfluency and so on.

One may question our decision to choose only intonation or prosody of a voice to differentiate a question from a statement. The objective of this research is to closely study the tonal characteristics of an utterance that make it a question utterance in the Arabic language.

There have been different characterization of questions and statements by different researchers. Some define questions as “those utterance classes which average speaker would unreflectively label Questions” while others define questions as “utterances that convey perceived lack of information.” [4]. We will adopt the first definition of question as it is easier to work with. In our research, we have used Native Arabic speakers to classify an utterance as either Question or Non Question.

Linguistically, there can be two overlapping sets of questions: first, those which are syntactically defined as interrogatives whether or not they function as questions in the discourse; and second, those which are not syntactically defined as questions but do in fact have a pragmatic force of questions. We have only considered the first type of questions as there a greater chance of confusion and misclassification when working with the second type.

Questions can be divided into several subtypes depending upon the perspective of the researcher. One can divide questions based on phonetics, linguistics, grammar etc. We have considered the following question subtypes in our research.

1) Yes-no Questions (YNQ)

- A Yes/No Question is a question that can be answered with yes or no.
- e.g. “Do you have to have any special training?”
- Normally they have final rise intonation [19].

2) Wh Questions

- Wh Question is a question that uses a question keyword to ask a question. In English they begin with What, Where, Which, Who, Whom, When, Why, How etc. and in Arabic Aish, Mata, Man, Ay, Wain, Laish, Hal etc. It can not be answered by a yes or no.
- e.g. “How old are you?”
- Normally they have final fall intonation [19].

3) Rhetorical Questions

- A rhetorical question is one that requires no answer because the answer is obvious and doesn’t need to be stated. The speaker (of the rhetorical question) is not looking for an answer but is making some kind of a point, as in an argument.
- e.g. “Are you kidding me?”

4) Embedded Question

- An embedded question is a part of a sentence that would be a question if it were on its own, but is not a question in the context of the sentence
- e.g. “I don’t know where he has gone”.

In the next section, we briefly survey the literature for various research work related to the use of prosody in various applications.

III. LITERATURE SURVEY

Prosody is widely used in many speech related applications. Researchers have prosodic features for dialog act detection [30], [12], [36]; topic segmentation [34], [8]; sentence boundary detection [25], [31], [26], [13], [38]; emotion detection [29], [42], [20]; disambiguate sentences [27], detecting speech disfluencies [24]; remove mistakes from the output of their speech recognizer [22], [10] etc.

In the above approaches, researchers have used various types of classifiers such as decision trees, support vector machines (SVM), artificial neural networks such as MLP or RBF, and simpler classifiers such as k-NN, linear classifiers like LDC, Boosting algorithms like AdaBoost etc. Some authors such as Lee and Narayanan [20] have combined the output from multiple classifiers to give a single result.

Most of the researchers have used decision trees for prosody feature selection and classification tasks. Their advantage is that analysis of feature usage is easier because we can see exactly the usefulness of each feature for classification purposes. Mostly, CART style decision trees have been used by the researchers in the field.

Prosodic models have been discussed extensively in the field of Text-To-Speech (TTS) systems. In this field, various models for different languages can be found e.g. for English [7], for Mandarin Chinese [9] and for German [14]. However, these are models to generate prosody from text to speech. Nevertheless, they give us important information about how prosody is used in different languages.

In English and German Languages it has been shown that Yes/No questions have a final rising, Question-Word Questions have final falling tone [19]. It was found that besides utterance final position, question intonation has an effect on the pitch of the whole utterance especially Question Word question. For Chinese Language, the spectral balance of the final syllable is the most useful prosodic feature then the pitch at the end of the utterance which is the most useful prosodic feature for English. [41] In Swedish, question intonation has been primarily described as marked by a raised topline and a widened F0 range on the focal accent [17]. Studies in Danish
language have found that in addition to final rising tone, other features such as onset F0 and overall pitch range also help distinguish a question from a statement [15].

Shriberg et al. [30] report that for English Language, Questions have higher F0 means and higher end gradients than Statements. Questions are shorter in duration, have a lower percentage of frames in continuous region and have less variability of speaking rate than those in Statements. Declarative and Yes-No Questions typically have a final F0 rise. Wh-Questions often fall in F0 as do Statements. Wh-Questions show a higher average F0 than Statements.

IV. FEATURE SET

In order to employ prosody, one needs to determine the set of features that must be examined in order to determine whether they are “discriminating” features that significantly contribute to the success of the classification process. Based on the features that have been investigated in the literature survey, we decided to study the following features:

1) Pitch Features: Pitch features are considered the most important set of features to determine a particular sentence as a Question sentence. Most of the Pitch features are not straight forward to extract because of the ways in which pitch is used by different speakers and in different speaking contexts. We use pitch to recognize tone but the problem is that pitch is never fully measurable. Therefore, we always have missing data values. There are also problems of discontinuities in pitch tracking such as doubling errors and pitch halving [31]. These errors are estimated by lognormal tied mixture model of F0, which computes a set of speaker-specific pitch range parameters [33]. Then in many approaches these values are processed through median filtering, linear stylization before being used for feature extraction. This filtering and stylization of F0 helps us in robust extraction of F0 features such as value of F0 slope at a particular point, the maximum or minimum stylized F0 within a region, average slope, speaker specific values etc.

Pitch detection algorithms can be broadly divided into two types, those that are estimate pitch in the time domain and those that do it in the frequency domain. Some of the pitch detection algorithms can be found in Yin [11], Praat [5], Momel [16], and ESPS [35]. Praat and ESPS are most widely used tools in this field.

2) Energy Features: Several studies have proposed different ways to measure energy measures of voice signals. Some authors claim that spectral balance, measuring the energy distribution over the frequency spectrum, is vital to detect stress and accent [32]. Tsao claimed that question intonation in Chinese is a matter of stress [37].

3) Duration Features: The term duration refers to the length of a particular constituent of speech. Duration is widely regarded to be an important prosodic cue to the detection of questions. It has been studied that the utterance final syllable is usually longer in question intonation than in statement intonation whereas the other syllables in question intonation tend to be shorter [40]. Therefore Yuan and Jurafsky have extracted three duration features namely the duration of final syllable, the average duration of other syllables, and the length of the whole utterance.

4) Pause Features: Pause features are very important for dividing an utterance into sentences. Mostly signal energy is used to calculate the pause features but some authors have used other measures e.g in [26] they use perceptual loudness extracted from audio files encoded using MPEG-1.

5) Speaking Rate Features: Considerable improvement in the recognition can be made simply by taking speaking rate into account[6]. Speaking Rate is defined differently by different people. The two most common definitions are word per minute (WPM) and syllable per second (SPS) [23]. However, since we don’t have information about either word or syllable, therefore we can’t use this measure.

V. FEATURE EXTRACTION

The feature extraction process is described in this section. Before extracting features, we segment the speech into separate sentences using pause as a criterion. For detecting pauses, we use intensity and energy in the sound signal. If during a certain duration threshold, the intensity of the signal is below a certain minimum intensity threshold then that duration is considered a pause. Currently the default value for minimum duration is 0.6 seconds and 59 dB for maximum intensity. We have marked the sentence interval at the boundaries of the pause with time margin of 0.1 second left around the sentence segments. For speech segmentation and features extraction, we use a speech phonetics application called Praat [5]. For segmenting the speech, we have modified the program written by M. Lennes [21].

After segmentation, we identify questions in the speech and label each question according to its type viz. Yes/No, Wh, Rhetorical or Embedded. Non Question sentences are labeled automatically by the segmentation program. These sentences are then fed into our feature extractor that extracts the relevant features from those segments. We have modified program called ‘Prosodic Feature Extraction Tool For Praat’ by Huang et al. [18]. This tool requires vowel, syllable, word information of the audio signal in order to extract features but since we don’t have this information in our case therefore we had to modify this tool.

One of the restriction of using this tool is that there should only be one speaker in one audio file. In Praat, pitch and energy are calculated by the frame. The default length of each frame is 0.01 second or 10 milliseconds. The start, end and duration of objects in Praat are measured by the index or the number of frames in the waveform.

For detecting pitch, Praat uses a method described in [5]. This method is more accurate, noise-resistant, and robust, than methods based on cepstrum or combs, or the original autocorrelation methods. The reason why other methods were invented, was the failure to recognize the fact that if one wants to estimate a signal’s short-term autocorrelation function on the basis of a windowed signal, one should divide the autocor-
relation function of the windowed signal by the autocorrelation function of the window:
\[ r_x(\tau) \approx r_{xw}(\tau)/r_w(\tau) \]

Since we didn’t have any word information with us so we could utilize only F0 and energy features. We also used pattern features. We divided each sentence into three parts, the starting 20 frames, the ending 20 frames and the middle portion of the utterance. We calculate F0, energy and pattern features for each of these regions and their linear and logarithmic ratios and differences.

- **F0 features**: Mean and variance of the logarithmic F0 values are used to normalize F0 features and calculate the overall baseline and topline pitch of the speech. Pitch contour is stylized and several F0 are computed from it. F0 features are mainly about the pitch range of a sentence. They include the minimum, maximum, mean, and last F0 values of a specific region such as starting, middle and ending segment of the sentence. These features are also normalized by the baseline F0 values, the topline F0 values, and the pitch range using linear difference and log difference. They also include features which compare the pitch ranges of the starting, ending and middle sections of the sentence.

- **Energy features**: The energy features are based on the intensity contour produced by Praat. Similar to F0 features, range features are also calculated for energy using various normalization methods.

- **Pattern features**: These are the maximum, minimum, average pitch and energy slopes in the sentences and the ratios of these values in the starting, middle and ending segment of the sentence. Pitch and energy slope features indicate rising, falling, and unvoiced patterns of the stylized pitch values or the intensity contour in the sentence. We included the slope difference and dynamic patterns (i.e., falling, rising, and unvoiced) across the three sections of the sentence.

Raw pitch values are calculated by Praat’s autocorrelation method. These values are then smoothed and unvoiced/voiced regions are determined and stored in TextGrid. Praat’s pitch stylization function is used to stylize raw pitch values over voiced regions. Pitch slope values are generated from stylized pitch contours.

Since there is no intensity stylization function in Praat, intensity values are stylized by pitch stylization function. Stylization is performed over entire intensity contour, whereas in pitch case, this stylization only applies in voiced regions.

**VI. CLASSIFICATION**

For classification, we use C4.5 decision trees [28]. We have chosen decision trees because decision trees represent rules and rules can be explained so that humans can comprehend them and know the reasons for a particular classification of data. We divided the entire feature set into various subsets to gain an understanding of the relative importance of different feature types. Then we compared the performance of each feature subset with the all-feature set on the basis of accuracy of result, size of decision tree and training time. Finally, we compiled the distinguishing features of questions obtained from these classifications. We have used WEKA’s [39] implementation of C4.5 decision trees for this experiment.

In order to analyze feature contribution to the task of question identification, we consider different categories of features, viz.,

1. **F0 Features**
2. **Energy Features**
3. **Patterns Features**

Then, we considered all possible combinations of these subsets, which include

1. **F0-Patterns Features**
2. **F0-Energy Features**
3. **Energy-Patterns Features**

The full feature set is the combination of F0, energy, patterns subsets plus one additional feature, viz., the sentence duration. Binary features for a question class is common in all these subsets. This feature has value 1 if the sentence is a question and 0 otherwise. F0-Patterns subset has all the F0 features plus the pattern features related to F0 from the Patterns features subset. Similarly, energy-patterns subset has all the features from energy feature subset in addition to energy related pattern features from the patterns feature subset. These subsets and the full feature set are then classified using decision trees so that contribution of these feature sets, toward the overall goal of question identification, can be measured.

**VII. EXPERIMENTAL RESULTS**

We, first, describe the corpus used in this research. To our knowledge, this research work is the first of its kind regarding the Arabic language, and therefore, we did not come across any benchmark audio data. Hence, we had to acquire data on our own. We managed to include 15 Arabic audio lectures by three different male speakers with sample rate of 8KHz and having mono channel. We had in total 12571 sentences out of which only 514 could be classified as one of the three Question types that we are considering. These lectures had 152 Yes/No or Wh questions, 24 Rhetorical questions and 338 Embedded questions. In order to keep an equal prior probability of Question and Non-Question classes during the training stage, we had to downsample the Non-Questions to 514. So at the end we had 514 questions and 514 non-question sentences. 144 Features were extracted from each of these 1028 sentences. Although we had transcribed three different types of questions in our corpus but due to our small data size, we couldn’t downsample the data any further.

In our experiment, we used confidence factor value of 0.25. The confidence factor is used for pruning (smaller values incur more pruning). The minimum number of instances per leaf is 2. Three folds of data is used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.

This experiment is run 10 times using 10-fold crossvalidation, then the values were averaged to give the final statistic.
given in Table I. Table I compares makes all-feature tree as a base system and compares other decision trees with it. It marks the statistically significant improvement or degradation with respect to the base system in each feature subset decision tree. We measured statistical significance using Paired t-test with confidence of 0.05 (two tailed).

Table I shows the comparison of trees of these feature sets using various measurement statistic. The True Positive (TP) rate is the ratio of exemplars classified as members of a particular class and total number of exemplars which are truly member of that class. It shows the proportion of that class captured by the classifier. It is equivalent to Recall. The False Positive (FP) rate is the ratio of examples which were classified as class x, but belong to a different class, among all examples which are not of class x. The Precision is the proportion of the examples which truly have class x among all those which were classified as class x. The F-Measure is a combined measure for precision and recall that is given by

$$ F_{\text{Measure}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1) $$

An ROC plot is a plot of true positive rate vs. false positive rate as the prediction threshold sweeps through all the possible values. ROC of 1 is perfect prediction – all positive cases sorted above all negative cases. ROC of 0.5 is random prediction – there is no relationship between the predicted values and truth.

As we can see in that table that the highest accuracy is achieved by decision tree with full feature set yielding an accuracy of 75.66% which is significantly higher than the chance level of 50%. Although Energy-only, F0-Patterns and Energy-Patterns have higher True Positive rate than that of all-feature tree, but they also have significantly higher False Positive rates. So, overall all-feature trees have higher Precision, F-measure and ROC area figures although F-Measures of Energy-only, F0-Energy and Energy-Patterns are almost similar to that of all-feature tree. However we see that all-feature tree takes significantly higher time for training than other trees. Its tree size and number of leaves are also significantly greater in number than other feature subset trees with the notable exception of Patterns-only tree. Patterns-only tree performs significantly worst than other feature set trees and its tree size and number of leaves is also significantly higher than other trees. Therefore, we can say that Patterns feature subset is completely useless for classification of Question and Non-Question sentences.

It can be seen from Figure 1 that full dataset achieves highest accuracy rate (75.66%) but tree made by only energy features also achieves similar level of accuracy(74.53%). The size of all-feature trees 108.4 (average of 10 all-feature decision trees) whereas the size of Energy-only tree is 47.12. So the size of energy-only tree is less than half of that of all-features tree and its performance is almost similar. Performance of energy-patterns-only tree is almost identical with 71.59 % accuracy and its size is just 6.36. Very simple compared to the energy-only and all-feature trees and achieving comparable accuracy. Interesting thing is that this tree doesn’t use pattern features in its tree.

F0-only tree achieves (69.91%) accuracy and its size is 78.36. Its performance is below energy-only tree but considerably better than pattern-only tree which achieves 58.15%, just above chance level of 50%. We have used only derived patterns set for this tree because tree with basic and derived patterns failed to do classification beyond the chance level. Pattern-only tree is not only worst in performance but its tree has largest size among all the feature trees. It can be seen from F0-pattern and energy-pattern trees as well that pattern seldom contribute to question identification. F0-pattern and energy-pattern trees reduce the accuracy of F0-only and energy-only trees. However, even when pattern inclusion reduces the accuracy by about 3%, it reduces the size of the tree by 10 times the size of original F0-only or energy-only tree. Paradoxically, even these reduced trees use only F0 or energy features. Figure 2 compares the sizes of these feature trees.
TABLE I
COMPARISON BETWEEN ALL-FEATURE DECISION TREE AND FEATURE SUBSET TREES

<table>
<thead>
<tr>
<th></th>
<th>Full Energy</th>
<th>F0 Patterns</th>
<th>F0 Energy</th>
<th>F0 Patterns</th>
<th>EnergyPatterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Correct</td>
<td>75.66</td>
<td>74.53</td>
<td>69.91</td>
<td>58.15</td>
<td>66.10</td>
</tr>
<tr>
<td>True Positive</td>
<td>0.78</td>
<td>0.85</td>
<td>0.74</td>
<td>0.57</td>
<td>0.78</td>
</tr>
<tr>
<td>False Positive</td>
<td>0.26</td>
<td>0.36</td>
<td>0.34</td>
<td>0.41</td>
<td>0.28</td>
</tr>
<tr>
<td>Precision</td>
<td>0.75</td>
<td>0.71</td>
<td>0.69</td>
<td>0.58</td>
<td>0.74</td>
</tr>
<tr>
<td>Recall</td>
<td>0.78</td>
<td>0.85</td>
<td>0.74</td>
<td>0.57</td>
<td>0.78</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.76</td>
<td>0.77</td>
<td>0.71</td>
<td>0.58</td>
<td>0.75</td>
</tr>
<tr>
<td>ROC</td>
<td>0.79</td>
<td>0.71</td>
<td>0.71</td>
<td>0.58</td>
<td>0.79</td>
</tr>
<tr>
<td>Elapsed Training Time</td>
<td>2.69</td>
<td>0.64</td>
<td>1.18</td>
<td>0.58</td>
<td>2.07</td>
</tr>
<tr>
<td>Tree size</td>
<td>108.36</td>
<td>47.12</td>
<td>78.36</td>
<td>176.28</td>
<td>89.48</td>
</tr>
<tr>
<td>No. of leaves</td>
<td>24.58</td>
<td>24.58</td>
<td>24.58</td>
<td>24.58</td>
<td>24.58</td>
</tr>
</tbody>
</table>

* Clearly significant improvement or degradation

subset. Apparently presence of pattern features didn’t do any improvement in full-feature tree. Patterns data successfully reduced tree size in case of F0-patterns tree and energy-patterns tree. F0 and energy feature sets are important to classification even though Energy-only feature set (38 features) seem to do just fine for classification task. Summary of size of each feature set is given in Figure 3.

Figure 4 shows the comparison of F-measure among the all-feature decision tree and all the feature subsets. All-feature decision tree and other feature subset decision trees, except F0-only and patterns-only decision trees, have almost similar performance for this precision/recall measure. F0-only and patterns-only decision trees perform significantly worse than other trees with patterns only tree only getting 0.58 score as compared to 0.76 by all-feature tree and 0.77 by energy-only decision tree.

Table II gives detailed comparison of various accuracy measures by class. This table also shows that the all-feature decision tree gives the most accurate and precise classification results followed closely by F0-energy and energy-only decision trees.

VIII. CONCLUSION

We compared feature subsets performance with the complete feature set to find out if the whole feature set is useful or is there any particular part of it that is mainly responsible for question identification or if some part is completely useless for this task. In this comparison, we have seen that all-feature dataset gives us the most accurate results with decision trees. Among feature subsets, we found that energy feature subset is most helpful for question identification and patterns feature subset doesn’t contribute to this task. However, pattern feature subset, when combined with F0-only or energy-only subset, produce surprisingly simple and small decision trees with good accuracy rate. Even though energy-only feature set give high accuracy almost matching accuracy achieved by the full feature set, we get good accuracy rated for F0-only feature set too. It shows that question cues are redundantly present in these feature subsets such that if one set of features is removed, other features compensate their removal.

ACKNOWLEDGMENT

We would like to thank King Fahd University of Petroleum and Minerals (KFUPM) for supporting this research work and providing the computing facilities.

REFERENCES

<table>
<thead>
<tr>
<th>Class</th>
<th>F0 Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>ROC Area</th>
<th>TP Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>NQ</td>
<td>0.253</td>
<td>0.752</td>
<td>0.768</td>
<td>0.747</td>
<td>0.755</td>
</tr>
<tr>
<td>F0</td>
<td>NQ</td>
<td>0.325</td>
<td>0.683</td>
<td>0.698</td>
<td>0.69</td>
<td>0.715</td>
</tr>
<tr>
<td>Energy</td>
<td>NQ</td>
<td>0.385</td>
<td>0.69</td>
<td>0.856</td>
<td>0.764</td>
<td>0.754</td>
</tr>
<tr>
<td>EnergyPattern</td>
<td>NQ</td>
<td>0.535</td>
<td>0.642</td>
<td>0.961</td>
<td>0.77</td>
<td>0.703</td>
</tr>
<tr>
<td>Pattern</td>
<td>NQ</td>
<td>0.43</td>
<td>0.565</td>
<td>0.558</td>
<td>0.562</td>
<td>0.552</td>
</tr>
<tr>
<td>EnergyF0</td>
<td>NQ</td>
<td>0.270</td>
<td>0.741</td>
<td>0.706</td>
<td>0.794</td>
<td>0.794</td>
</tr>
</tbody>
</table>

Q represents Questions category and NQ represents Non-Question category of sentences

---


