Automatic question detection: prosodic-lexical features and crosslingual experiments

Vũ Minh Quang – Laurent Besacier** - Eric Castelli**

*International research center MICA, CNRS/UMI-2954 – 1, Dai Co Viet - Hanoi, Vietnam
**LIG Laboratory, CNRS/UMR-5217 - 681 rue de la passerelle - BP 72 - 38402 Saint Martin d'Hères , France
{Minh-Quang.Vu ; Eric.Castelli}@mica.edu.vn ; Laurent.Besacier@imag.fr

10.21437/Interspeech.2007-613

Abstract

In this paper, we present our work on automatic question detection from the speech signal. We are interested in developing automatic detection system and investigate the portability of such system to a new language. The first goal of this paper is to propose and evaluate a combined approach for automatic question detection where prosodic features are augmented by the use of lexical features. It is shown that both early and late integration of theses features in a decision tree-based classifier improves the question detection performance compared to a baseline system using prosodic features only. The second goal of this paper is to conduct a crosslingual (French / Vietnamese) evaluation concerning the use of prosodic features. It is shown that our first system developed for French which uses an initial prosodic feature set can be improved using a new feature set that takes into account some specific prosodic characteristics of the Vietnamese tonal language. Both Vietnamese and French question detection systems obtain F-ratio performance around 80% on pre-segmented meeting and dialog utterances.

Index terms: prosodic / lexical features, automatic question detection, crosslingual experiments

1. Introduction

Automatically detecting questions from the speech signal is a challenging task that may be interesting for document summarization or to enrich a transcription by punctuation marks as well as in the objective of dialog act detection [1, 2]. For the non-tonal Western languages (French or English) it was validated that sentence prosody carries extra linguistic information, such as emotions, state of the speaker, or sentence type (affirmative, interrogative or exclamative [3, 4]). To automatically evaluate the type of sentences for detection or classification purposes, it is possible to analyze the speech signal directly by using its prosodic characteristics, as done in [1]. Using only prosodic information leads to acceptable performance but could be probably improved by using other cues. Thus, the first goal of this paper is to propose and evaluate a combined approach for automatic question detection where prosodic features are augmented by the use of lexical features that might be obtained by an Automatic Speech Recognition (ASR) engine for instance. The second goal of this paper is to conduct a crosslingual (French / Vietnamese) evaluation concerning the use of prosodic features. In fact, typical prosodic features take into account the evolution of the intonation during sentence statement: range of F0, increase of F0 at the end of the sentence or other parameters derived from the values of F0 [1, 5]. However, in the case of tonal languages (like Mandarin or Vietnamese), the melody contour of the intonation is complex [6]. It is composed of macro-variations corresponding to the intonation of the sentence and of micro-variations corresponding to the lexical tone applied to each syllable of mono (or bi) syllabic words. This is why the direct application to tonal languages of prosodic feature sets developed for non-tonal languages is very likely to be suboptimal since tonal micro-variations tend to scramble the extra-lexical information coded in the sentence prosody.

A former study from the authors of this paper [7] has shown that, as for non-tonal languages, the sentence type (question or not) information is mainly coded by the fact that the intonation goes up or not at the end of the sentence. However, this information can be scrambled by the modulation of prosodic contour due to the lexical tone. For instance, it was shown in a perception test that listeners (and so probably an automatic system) can badly classify assertions when produced sentences have a final syllable with the Vietnamese *rising* tone[1]. Similarly, questions can be badly classified if their final syllable carries a *falling*[2] tone. Thus, the second contribution of this paper consists in proposing an optimized prosodic feature set for Vietnamese that takes into account these problems, and to evaluate it through crosslingual experiments.

Section 2 of this paper presents the baseline features initially designed for question detection in French as well as our optimized feature set for Vietnamese. Section 3 shows briefly our lexical features while sections 4 and 5 present our crosslingual and prosodic-lexical experiments respectively. Finally section 6 concludes this work.

2. Prosodic features

2.1. Baseline features for French

In French language, the interrogative form of a sentence is strongly related to its intonation curve. Therefore, we decided to

[1] The *rising* tone is one of the 6 different tones of the vietnamese language

[2] The *falling* tone is one of the 6 different tones of the vietnamese language

use the evolution of the fundamental frequency (F0) to automatically detect questions in an audio input.
From this F0 curve, we derive a set of features which aim at describing the shape of the intonation curve. The parameters defined for our work are listed in Table 1. It is important to note here that, contrarily to classical short term feature extraction used in speech recognition, a unique long term feature vector is automatically extracted for each utterance of the database.
These features can be divided into 2 main categories: the first 5 features are the statistics on F0 values, and the 7 next features describe the contour of F0 (raising or falling). The F0 contour was extracted using the Praat[1] software.

No	Parameter	Description
1	Min	Minimal value of F0
2	Max	Maximal value of F0
3	Range	Range of F0-values of the whole sentence (Max-Min)
4	Mean	Mean value of F0
5	Median	Median value of F0
6	HighGreater-ThanLow	Is sum of F0 values in first half-length smaller than sum of F0 values in last half-length of utterance?
7	RaisingSum	Sum of F0i+1– F0i if F0i+1 > F0i
8	RaisingCount	How many F0i+1 > F0i
9	FallingSum	Sum of F0i+1– F0i if F0i+1 < F0i
10	FallingCount	How many F0i+1 < F0i
11	IsRaising	Is F0 contour rising? *(yes/no)*. Test whether RaisingSum > FallingSum
12	NonZero-FrameCount	How many non-zero F0 values?

Table 1: 12-dimensional feature vector derived from the F0-curve for each utterance

2.2. Optimized features for Vietnamese

In [7], it was shown that for Vietnamese, the differences between interrogative and affirmative sentences are characterized primarily by: a difference in F0 contour (increasing or decreasing) at the very end of the sentence; a higher register and a stronger intensity. These findings allow us to propose a new set of features which are more appropriate to describe these differences than those initially developed for French corpus in table 1. We do not detail all these features here but the main added parameters are related to the intensity (for instance minimum intensity of the current sentence), to the syllable durations, and above all, an important parameter was added which represents the F0 range in the last demi-syllable of the sentence. More precisely, it is the difference between F0 values of ending and beginning points of the last demi syllable. This parameter is particularly important to tackle the problem of the final tone (*rising* or *falling*) described at the end of section 1. All these features can be measured automatically; the parameters that need syllable begin / end labels were extracted after applying a forced alignment between the speech signal and its transcription[2].

[1] http://www.fon.hum.uva.nl/praat/

3. Lexical features

The goal of our lexical features is to represent interrogative terms or expressions. However, for many languages, the presence / absence of such terms is not sufficient, since their position in the sentence is also very important. For instance, French interrogative words like *pourquoi*[3] or *comment*[4] will probably indicate an interrogative utterance if they are the first word of it, while this might be different if they are not at the beginning of the sentence. Thus, bag-of-words techniques, like those used in information retrieval for instance, may be not adequate in our case. We need also to represent, in our lexical features, the position information of the words into a sentence.
Consequently, we defined parameters that describe both presence and position of interrogative words into the sentence. Our lexical features can be classified into three sub-categories :

- unigrams or bigrams present before a group of interrogative terms[5]
- the presence or absence of some interrogative terms in the utterance
- the unigrams present after a group of interrogative terms

These interrogative terms were chosen differently in each category and they are all specific for each language (French or Vietnamese). Moreover, to capture their positions (begin, middle or end in the sentence), two special tags “BEGIN” and “END” were added surrounding each sentence. In this way, both the presence and the position of interrogative words are correctly modeled by our lexical features.

4. Crosslingual Experiments

4.1. Experimental framework

4.1.1. French telephone meeting corpus: DELOC

Our telephone meeting corpus (called Deloc for “delocalized meetings”) is made up of 13 meetings of 15 to 60 minutes, involving 3 to 5 speakers (spontaneous speech). The total duration is around 7 hours and the language is French.
Different types of meetings were collected which correspond to three categories: recruitment interviews; project discussions in a research team; and brainstorming-style talking.
From this corpus, we have manually extracted a subset composed of 468 utterances: 234 question (Q) utterances and 234 non-question (NQ) utterances.

[2] for the experiments reported here, the forced alignment was done using the verbatim transcriptions.
[3] why
[4] how
[5] among these terms we find for instance for French : *pourquoi, qui, quand, comment, combien, où…*

4.1.2. French client / agent dialog corpus : NESPOLE

The NESPOLE project was a common EU NSF funded project exploring future applications of automatic speech-to-speech translation in e-commerce and e-service sectors. The scenario of NESPOLE involves an Italian speaking agent, located in a tourism agency in Italy discussing with a client (English, German or French speaking) located anywhere via Internet and using audio-conferencing tools like NetMeeting. The client wants to organize a trip in the Trentino (Italia) area, and asks the agent for information concerning his trip. More information on this database can be found in [8]. We use in this experimentation a subset of the French-speaking part of this corpus; it consists in 650 Q- and 650 NQ-sentences

4.1.3. Vietnamese databases : VietP and Assimil

The *VietP* corpus was used in our early study on Vietnamese prosody [7]. This corpus is made up of 14 pairs of Q/NQ sentences which were extracted from significant dialogues. Each dialogue is repeated five times by six native speakers (3 men and 3 women) from Hanoi (the North region, considered to be official pronunciation of the Vietnamese language). The corpus is finally composed of 420 Q- and 420 NQ-sentences.
The *Assimil* corpus was extracted from the "Assimil language learning CDrom" for Vietnamese and contains 168 Q- and 168 NQ-sentences. It is used in the prosodic-lexical experiments for Vietnamese since the *VietP* corpus transcriptions are too poor (only 14 different Q/NQ pairs) to be used in experiments where lexical information is taken into account.

4.1.4. Decision tree-based classifier

Traditionally, statistical-based methods such as Hidden Markov Model (HMM) or Gaussian Mixture Model (GMM) can be used to solve classification problems in speech processing. These statistical methods generally apply on short term features, extracted for instance at a 10ms frame rate. However, in our case, statistical methods are hard to use since we do not use short term feature vectors, as explained in section 2.1: one feature vector only is extracted for the whole utterance to be classified, which excludes the use of conventional statistical classifiers.
Thus, decision trees, which correspond to another classical machine learning (ML) method [9,10] are a good alternative.
Decision tree is a divide-and-conquer approach to the problem of learning from a set of independent examples (a concrete example is called instance). Nodes in a decision tree involve testing a particular condition, which usually compares an attribute value with a constant. Some other trees compare two attributes with each other, or utilize some functions of one or more attributes. Leaf nodes give a classification for all instances that satisfy all conditions leading to this leaf, or a set of classifications, or a probability distribution over all possible classifications. To classify an unknown instance, it is routed down the tree according to the values of attributes tested in successive nodes, until it reaches a leaf. The instance is classified according to the class assigned to this leaf.

For this work, we have used an implementation of decision-tree algorithms that is included in the open-source toolkit Weka[1] which is a collection of algorithm implementations written in Java for data mining tasks such as classification, regression, clustering, and association rules.

4.1.5. Protocol and evaluation measure

In all of these experiments and for each corpus, we use 10-folds cross validation procedure: the whole corpus is equally and randomly divided in 10 parts. In the first repetition, the first part is taken apart and used for test data while the 9 remaining parts are used for training data. In each of these next repetitions (from 2^{nd} to 10^{th} repetition), the test data is changed to the 2^{nd} part, 3^{rd} part…10^{th} part while all remaining 9 parts are for training data. The performance given for each experiment is then the mean of the scores obtained for each training / test configuration.

For the training data, a decision tree is constructed (the decision-tree algorithm used in our experiments is called "C4.5") and the obtained classifier is evaluated on the remaining test data. The evaluation is based on measures coming from the information retrieval domain such as recall (R), precision (P) and F-ratio, where :

$$R=\frac{N_{\text{correctly detected questions}}}{N_{\text{total questions in the test set}}} \qquad P=\frac{N_{\text{correctly detected questions}}}{N_{\text{total questions detected}}}$$

$$FRatio=\frac{2P\cdot R}{P+R}$$

The final detection performance is thus measured as the mean of F-ratio on test data for all 10-folds cross validation.

4.1.6. Crosslingual experiments results

Table 2: Crosslingual experiments for automatic question detection using prosodic features

	Feature set FR (Fratio)	Feature set VN (Fratio)
Deloc database (FR)	**74%**	67%
Nespole database (FR)	**73%**	68%
VietP database (VN)	77%	**81%**

Table 2 shows our crosslingual experiments results where both baseline (French) and adapted (Vietnamese) feature sets are evaluated on french and vietnamese test sets. The results show that the use of the matched feature set is important to optimize performance: while baseline feature sets are optimal for non tonal languages, it is important to use optimized features for the Vietnamese tonal language (81% Fratio instead of 77% for Vietnamese on *VietP* database). It may be found surprising that the overall performance is higher on the VietP database than on the other corpora. The main explanation is that both French corpora (FR) are made of utterances extracted from real dialogs

[1] http://www.cs.waikato.ac.nz/~ml/weka/

or meetings while the VietP data (VN) was recorded in a much more controlled situation.

5. Combining lexical and prosodic features

The combination of prosodic and lexical features was done both in early and late integration fashion:
-in the early integration, both prosodic and lexical feature sets where merged into a single vector during training and a new decision tree was built. The same process was applied to the feature set during testing before questioning the decision tree.
-in the late integration, both lexical and prosodic features were used separately to train two different decision trees during training. During the test, each prosodic and lexical feature set was sent to the corresponding decision tree and the final decision was based on a composite score which was a linear combination between prosodic and lexical scores.
Table 3 summarizes the results obtained for the features used separately or combined. For this experiment, in order to test the potential of the approach without taking into account the quality of the ASR recognizer that would be used to decode the utterances, we used the verbatim transcription of the test utterances for the lexical approach. We are aware that a noisy ASR output may have a bad influence on the lexical part of the combined approach.
The results show that the combined approach may be interesting to slightly improve the overall question detection procedure. However, further investigation is needed to 1) have conclusive idea of the better combination to use (differences between early and late integration are not always significant) 2) check if the improvement remains when using a noisy ASR output to get the lexical features. It is also interesting to note that the lexical features seem more important, compared to the prosodic features, for the tonal language. One explanation may be that, for tonal languages, the prosody also contributes to the tone applied to each syllable, and is consequently less used to encode other informations like sentence type.

Table 3: Combining prosodic and lexical features for automatic question detection (%Fratio)

	Prosodic feat.	Lexical feat.	Combined (early integration)	Combined (late integration)
Deloc database (FR)	74%	58%	77%	**78%**
Nespole database (FR)	73%	65%	**77%**	75%
Assimil database (VN)	64%	81%	82%	**88%**

6. Conclusion

This paper presented our ongoing work concerning automatic question detection. First, we proposed to use a new prosodic feature set which takes into account the specific prosodic characteristics of tonal language (like Vietnamese) and have shown that it can lead to better results than a feature set originally developed for a non-tonal language (French). We have also shown that, despite the fact that in a tonal language, the prosody is complex due to the presence of lexical tones, an automatic question sentence detection from the speech signal can be obtained with satisfying performance (81%). Secondly, it was shown that a better performance can be obtained by combining prosodic and lexical features. Future works must include : identifying the best way to combine these two models, seeing how our new prosodic features generalize to other tonal languages, using lexical features extracted from an ASR output instead of verbatim transcriptions and evaluate our question detection system on a continuous audio flaw using an automatic utterance segmenter.

7. References

[1] Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Cocarro, N., Martin, R., Meteer, M. & Van Ess-Dykema, C. "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *Language and Speech 41*, pp. 439-487, 1998.

[2] Vu M.Q., Castelli E., Boucher A. & Besacier L. "Classification de parole en Question et Non-Question par arbre de décision" *SFC 05, 12èmes Rencontres de la Société Francophone de Classification* - Montréal, 2005

[3] Rossi M. "*L'intonation, le système du français : description et modélisation*" Editions Ophrys, 1999, ISBN : 2-7080-0912-5.

[4] Hirst, D.J. &Di Cristo, A. (Eds.) "*Intonation Systems. A Survey of 20 Languages*" Cambridge: Cambridge University Press.

[5] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," Computational Linguistics, vol. 26, pp. 339–373, 2000.

[6] Yuan, J.; Shih, C; Kochanski, G. P., Comparison of Declarative and Interrogative Intonation in Chinese. *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, pp. 711-714.

[7] Vu M.Q., Tran D.D. & Castelli E. (2006) *Prosody of Interrogative and Affirmative Sentences in Vietnamese Language: Analysis and Perceptive Results.* Interspeech2006 ICSLP, International Conference on Spoken Language Processing. 17-21 September 2006, Pittsburgh PA, USA

[8] Mana N., Burger S., Cattoni R., Besacier L., Maclaren V., Macdonough J., Metze F., "The NESPOLE! VoIP Multilingual Corpora in Tourism and Medical Domains" Eurospeech 2003, Geneva, 1-4 Sept. 2003.

[9] Marquez L., "Machine learning and Natural Language processing", Technical Report LSI-00-45-R, Universitat Politechnica de Catalunya, 2000.

[10] Witten I.H., Frank E., Data mining: Pratical machine learning tools and techniques with Java implementations, Morgan Kaufmann, 1999.