Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 128 (2018) 38–46

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Natural Language and Speech Processing, ICNLSP 2015

Detection of Sentence Modality on French Automatic Speech-to-text Transcriptions

Luiza Orosanu[a,*], Denis Jouvet[a]

[a] *Speech Group, LORIA, Université de Lorraine, LORIA, UMR 7503,, Nancy 54600, France*

Abstract

This article analyzes the detection of sentence modality in French when it is applied on automatic speech-to-text transcriptions. Two sentence modalities are evaluated (questions and statements) using prosodic and linguistic information. The linguistic features consider the presence of discriminative interrogative patterns and two log-likelihood ratios of the sentence being a question rather than a statement: one based on words and the other one based on part-of-speech tags. The prosodic features are based on duration, energy and pitch features estimated over the last prosodic group of the sentence. The classifiers based on linguistic features outperform the classifiers based on prosodic features. The combination of linguistic and prosodic features gives a slight improvement on automatic speech transcriptions, where the correct classification performance reaches 72%. A detailed analysis shows that small errors in the determination of the segment boundaries are not critical.

Keywords: Question detection; speech-to-text transcriptions; prosody; likelihood ratio; part-of-speech tags

1. Introduction

The work presented in this paper is part of the RAPSODIE project, which aims at studying, deepening and enriching the extraction of relevant speech information, in order to support communication with deaf or hard of hearing people. The detection of sentence modality (questions versus statements) is a key problem here, the deaf or hard of hearing people must be informed when a question is directed to them, and that they should respond or ask for further clarifications (if needed). Therefore, we are interested in finding a solution that performs well on automatic speech-to-text transcriptions (its accuracy depends on the sound quality and on the performance of the speech recognition system).

The automatic detection of questions has been studied in the past decades with different objectives: to model and detect the speech structure [1], to detect the sentence modality (mainly statements versus questions) [2, 3, 4, 5, 6, 7],

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: luiza.orosanu@loria.fr

10.1016/j.procs.2018.03.006

to create the summary of documents or meetings [4], to enrich an automatic transcription with punctuation marks [8], etc.

Most of the studies are only based on correct data (textual and/or audio), without being related to any automatic speech recognition systems. A detector of French questions (versus declarative and exclamatory sentences) was designed in [2] using only prosodic features (pitch and energy) computed on the last 700 milliseconds of speech. Prosodic based classifiers were also studied for the Arabic language [6]. The energy and the fundamental frequency were the key features in their classifier.

Another detector of French questions (versus declarative sentences) was designed in [4, 5]. They first started off with only prosodic features and soon realized that using other cues like the lexical information can improve its performance on spontaneous speech.

In [9] the English question asking behavior was designed in order to improve the intelligent tutoring systems. Their study concluded that the most useful features were prosodic, in particular the pitch slope of the last 200 milliseconds of a turn.

The studies related to automatic speech recognition systems have to additionally take into account word error rates, poor sound qualities, spontaneous speech, which can highly decrease the classification performance. In [1], 42 dialog acts were used to model and detect the discourse structure of natural English speech (human-to-human telephone conversations). They used three different types of information (linguistic, prosodic and a statistical discourse grammar) and achieved an accuracy of 65% on ASR transcripts versus 72% on reference manual transcripts. Combining recognized words with the discourse grammar was the most useful for this task.

The detection of questions in English meetings was addressed in [10]. They used lexico-syntactic, turn related and pitch related information and achieved an accuracy of 54% on ASR transcripts versus 70% on reference manual transcripts. The lexico-syntactic features were the most useful for this task. The automatic punctuation (comma, period, question mark) of French and English speech-to-text data was studied in [8]. Their boosting-based model uses linguistic (based on word n-grams) and prosodic information and was tested under real world conditions.

All of the mentioned classifiers are applied on different languages, on different data, on different conditions, and with different features. Some use data sets that are already manually classified in dialog acts, others extract sentences from their data sets based on the punctuation marks (with or without posterior manual relabeling). Some, based on the general opinion that a question's intonation has a final rising pitch [11], are interested only in the last part of the speech signal, others on the whole sentence. Some compute various prosodic coefficients (even up to 123 different coefficients), others keep only the most classic values: mean, maximum, minimum, delta, slope, etc. Every analysis is unique and very dependent on the data and on the choice of features.

In our study several approaches are analyzed: creating a classifier with only prosodic features (extracted from the acoustic signal) or one with only linguistic features (extracted from the word and part-of-speech sequences) or one that combines both linguistic and prosodic features. The classifier evaluations are carried out using linguistic features stemming out from automatic speech-to-text transcriptions (to study the performance under real conditions) and from manual transcriptions (to study the performance in ideal conditions - i.e. when there are no word errors).

The novelty of our approach consists in combining 3 different types of linguistic features (word-based n-grams, PartOfSpeech-based n-grams and the presence of discriminative interrogative patterns) to detect questions in French automatic speech-to-text transcriptions. The first experiments are conducted on perfect (predefined) sentence boundaries. Then we evaluate the performance loss when sentence boundaries are not perfect (by changing those boundaries randomly or relative to silence/noise decoded units).

The paper is organized as follows: section 2 is devoted to the description of the data and tools used in our experiments, section 3 provides a description of the features used for question detection, and section 4 presents and analyzes the results.

2. Experimental setup

2.1. Textual data for training language models

Textual punctuated data is necessary for modeling the lexical and syntactic characteristics of questions and statements.