Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 225–232

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5–6 November 2017, Dubai, United Arab Emirates

Towards the supervised machine learning and the conceptual segmentation technique in the spontaneous Arabic speech understanding

Younès Bahou [a,b*], Mohamed Hédi Maaloul [b,**], Emna Boughariou [b]

[a] *Hail University, Kingdom of Saudi Arabia*
[b] *Sfax University, Tunisia*

Abstract

The understanding task of an utterance meaning depends mostly on its concepts extraction. In this paper, we propose a method for the spontaneous Arabic speech understanding, in particular a conceptual segmentation of spontaneous Arabic oral utterances. It takes a transcribed utterance as input and provides conceptual labels as output in the form of a set of Conceptual Segments (CSs). This method is a part of the numerical approach and is based on supervised machine learning (ML) technique. The originality of our work lies in the processing of Out-Of-Vocabulary (OOV) words whether before and/or after the utterance segmentation task. Furthermore, this work is a part of the improvement of the understanding module of SARF system [2]. Indeed, we aim to compare our numerical method with the symbolic one proposed by [2] and the hybrid one proposed by [1].

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

Keywords: Arabic speech understanding, conceptual segmentation, out-of-vocabulary words, supervised machine learning.

1. Introduction

In the current context, we have to deal with the conceptual segmentation of spontaneous Arabic oral utterances using a numerical method based on the supervised ML technique. Thus, we propose an original method, which aims at extracting the utterance concepts in the presence of the speech recognition errors (i.e., OOV words). We use the ML technique for both the OOV words processing and the utterances conceptual segmentation.

Several methods have been explored in the automatic understanding context of the spontaneous speech, particularly the conceptual segmentation. We have clustered these methods into three approaches, namely, symbolic approach,

* Younès Bahou. Tel.: +966-50-934-7656 ; Tel.: +216-98-211-189.
** Mohamed Hédi Maaloul. Tel.: +216-99-015-015.
E-mail address: bahou.younes@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.113

numerical approach and hybrid approach. We have adopted the numerical approach based on the supervised ML technique because it has the advantage of greatly reducing the use of human expertise. Indeed, it is based on statistical models that rely on parameters estimated from a huge number of annotated data [7]. Statistical methods based on ML technique have been widely applied to the Arabic speech understanding axe such as in the Jamoussi system [8] (by neural networks and bayesian network) and the classifier YOSR [6] using the Support Vector Machine (SVM) multi-class classifier.

This work is a part of the realization of the interactive vocal server SARF [2] that offers users an oral access in Modern Standard Arabic (MSA) language to Tunisian railway information. We aim, through this work, to test our conceptual segmentation method of Arabic oral utterances through the understanding module of the SARF system, and to compare our method with its other versions.

The remainder of the paper is organized into five sections. Section 2 presents an overview of the contribution of the use of the conceptual segmentation technique in the speech understanding context. Section 3 presents our proposed conceptual segmentation method of Arabic oral utterances. Section 5 presents the implementation of our proposed method on the CSM module. Section 6 presents the evaluation of CSM along with a comparaison with the current literal understanding module of SARF system and AOCSM module [1].

2. Conceptual segmentation in the automatic speech understanding

The automatic speech understanding field can be seen as an intention of mapping between a string of words and a semantic representation conveying the main ideas of an utterance [8]. It requires linguistic knowledge allowing the semantic interpretation of this utterance in the presence of extralinguistic phenomena [5]. Hence, we are faced with the extraction and the labelling of semantic concepts from transcribed utterances in order to perform the task of the automatic spontaneous speech understanding. We note that an *utterance* refers to a complete unit of speech bounded by the speech recognition errors, the spontaneity of the interaction and the processed language challenges.

Consequently, the conceptual segmentation technique has been widely applied to the speech understanding. It presents the first level of automatic oral understanding [7]. Its aim is to extract the set of concepts in the form of a semantic representation. Indeed, this technique consists of splitting an utterance into CSs that have a sequence of words with the same sense entity. These CSs must be minimal and of very fine granularity [4].
However, the conceptual segmentation technique has proved its performance in automatic speech understanding systems in the presence of several speech specificities such as disfluencies that are due to the spontaneity of the interaction (self-correction, repetition, hesitation and word-fragment) [1] and OOV words that are due to the speech recognition errors (unknown words and miss-recognized words) [3].

This paper present a contribution to treat only the OOV words. By the fact, most of the works have dealt with the processing of both unknown and miss-recognized words after the conceptual segmentation task. Still, unknown words do not depend for their detection on the labeled CSs contrary to the miss-recognized words. That is why we explore the possibility of processing those unknown words before the conceptual segmentation task.

3. Proposed method for the conceptual segmentation

In this section, we detail our proposal as a part of the improvement of the literal understanding module of the SARF system [2]. As previously mentioned, we use the ML technique for both the OOV words processing and the conceptual segmentation of Arabic oral utterances. The decision taken for using the numerical approach, is guided by examining the contribution of the ML technique instead of the symbolic approach currently implemented in the SARF system for both the conceptual segmentation and the OOV words processing tasks.

Indeed, our method is carried out using the ARDIC (Arabic Railway DIalogue Corpus) [4] which is a corpus of Arabic dialogues acquired using the Wizard of Oz technique. This corpus includes 360 dialogues recorded between