Statistical Approach for Spontaneous Arabic Speech Understanding Based on Stochastic Speech Recognition Module

Aymen Trigui, Naim Terbeh, Mohsen Maraoui, Mounir Zrigui

LaTice (Monastir team), Faculty of sciences of Monastir, Monastir, Tunisia
trigui.aymen@gmail.com, terbehnaim1987@gmail.com,
maraoui.mohsen@gmail.com, mounir.zrigui@fsm.rnu.tn

Abstract. This work is part of a big research project named "Oreillodule" aimed to develop tools for automatic speech recognition, translation, and synthesis for Arabic language. In this paper, our attention has mainly been focused on an attempt to present the semantic analyzer developed for the automatic comprehension of the standard spontaneous Arabic speech. We present a model of Arabic speech understanding system. In this model, both speech recognition module and semantic decoding module are based on statistical approach. In this work, we present and evaluate speech recognition module but we just explain the principle of Arabic speech understanding module.

Keywords: Speech understanding, Arabic language, Probabilistic model, semantic analyses, corpus.

1 Introduction

In the past 40 years there has been a significant research effort directed toward automatic speech recognition. Our work falls within the area of automatic understanding of the Arabic language, specifically in the context of finalized human / machine communication interfaces. The efficiency and performance of automatic spontaneous Arabic speech understanding system depend on its strength and its ability to overcome the difficulties of natural language processing among which some are linguistic and this concerns the understanding of written and spoken data. These problems are usually caused by the use of references, polysemic words, vague predicates, implicit form, etc. Others are due generally to the characteristics of spontaneous oral and in particular of the Arabic speech one.

The uses of statistical models for speech recognition and understanding have the advantage of being portable to other areas, or to multilingual applications [1]. In this work, we present a model for spontaneous Arabic speech understanding system. Both speech decoding and statement understanding are based on statistical approaches. We start in a first section by presenting the Arabic speech specificities and in the following section; we expose system architecture with a detail about each system components with its formal description.

2 Arabic Speech Specificities

In this section, we begin by listing the main characteristics of the Arabic speech and we detail its specificities.

2.1 Arabic Phoneme Set

Consonants. We can classify consonants according to several grammatical and phonetic criteria [2]: consonants articulated with vibration of the vocal cords and consonants that do not cause a vibration of the vocal cords, the crossing of air through the vocal tract gives rise to other varieties of sounds. The 28 Arabic consonants can also grammatically, be divided into two groups [3]:

- 14 solar consonants those are similar to the pronunciation of the « لـ » (Al) in the «الشمس» word (the sun). With this consonants category we must pronounce the letter "Al" before the word.
- 14 lunar consonants those are similar to the pronunciation of the « لـ » (Al) in the «القمر» word (the moon). With this consonants category we do not pronounce the letter "Al" at the beginning of the word.

Table 1. Consonants classification taking into account the transcription constraints

Lunar Consonants	أ ب ج ح خ ع غ ف ق ك هـ م و ي
Solar Consonants	ت ث د ذ ر ز س ش ص ض ط ظ ل ن

Vowels. In Arabic language, we distinguish three short vowels (« ُ » (ضمة /dhamma:/), « ِ » (كسرة /kasra:/), «َ » (فتحة /fatha:/)) and three long vowels (the fatha « َ » extended by an alif "ا", the dhamma « ُ » extended by a waw "و" and kasra « ِ » is extended by a "ي"). The duration of a long vowel is about twice the size of a short one. These vowels are characterized by the vibration of the vocal cords. They are represented in the following table:

Table 2. Arabic language vowels classification

Short	/ُ- /ِ - /َ
Long	ُو- ِي- ـَا

2.2 Other Vocalic Achievements

Semi-vowels. Arabic has two phonemes considered as semi-vowels: a bilabial spirant one « و » and a prepalatal spirant one « ي », called semi-vowels or sonants because of their kinship with the closed vowels « ُ » et « ِ ». These phonemes are used sometimes as consonants, sometimes as the corresponding vowel.

The "Soukoun" (السكو) . The "Soukoun" (السكون) is not a vowel itself but it is the absence of vowel. Indeed, even if it is part of "Haraket" (الحركات) of Arabic it is not comparable to the other six vowels. The "Soukoun" is noted by a small circle « ْ ».

The "Tanwin" (التنوين**)**. The three vowel diacritics may be doubled at the end of a word to indicate that the vowel is followed by the consonant n. These may or may not be considered harakāt, and are known as tanwīn (تَنْوِين), or nunation. The signs indicate, from right to left, -un, -in, -an.
In words containing "Tanwin" we listen the vocalic presence of the /N/ (Noon) phoneme but in reality when the word will be transcribed the (Noon) will not appear between the world's consonants [4] [5]. We can cite as an example the "Tanwin" in the end of the word "مَكْتَبٌ" (the office).

The Gemination. A consonant may be doubled by wearing the /chadda:/ « ّ » (الشدة) signe, which represents an intensification in the pronunciation of a consonant. All consonants may be geminated and this gemination has a sense différenciative function. The geminated consonant is considered equivalent to two identical single consonants immediately follow each other. A gemination appears only in the position where a group of two consonants is allowed, in other words, between vowels [6][7]. The beginning and the end of each geminated consonant may be owned by different significant item, in other words a word or morpheme boundary can pass between them. These two consonants can also be separated in one term of a morphological couple. It is therefore right to consider the geminate consonants as groups of two similar consonants.

3 Understanding System Presentation

3.1 System Architecture

We present in Figure 1 the general architecture of the proposed Arabic speech understanding system.

Speech Recognition Module. This module is responsible to providing a text from an acoustic segment. All details concerning this component on the system are described in our previous work [6] and [7].

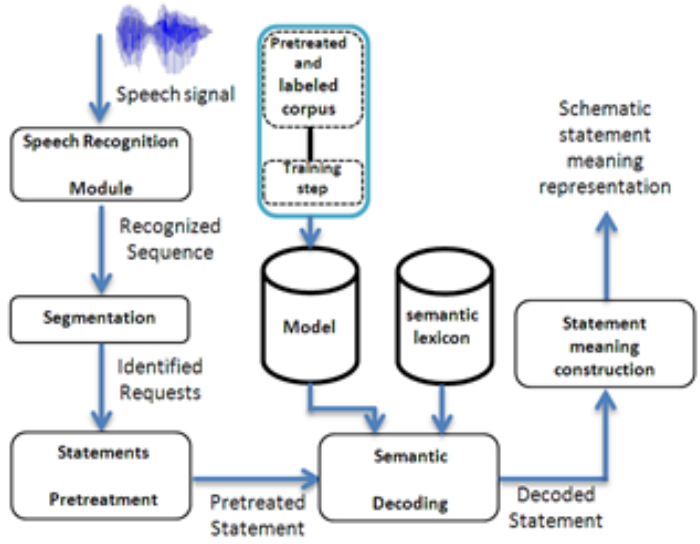

Fig. 1. The architecture of the proposed Arabic speech understanding system

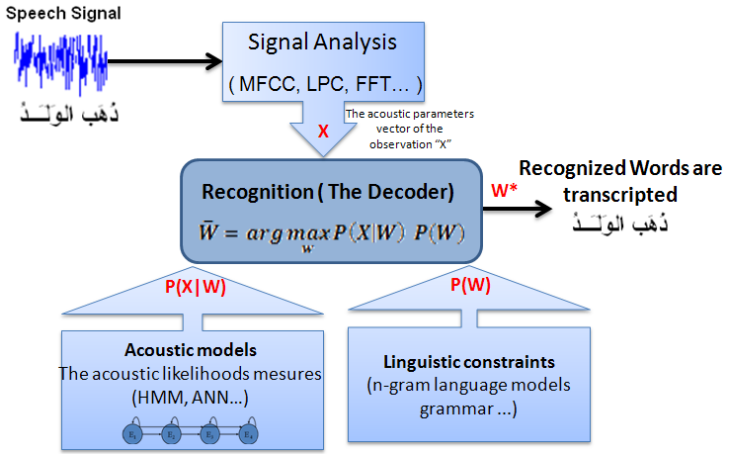

Fig. 2. General architecture of the speech recognition module

We present below the general architecture of the used Arabic speech recognition module.

From an acoustic segment the system generate a set of parameter vector (MFCC[1] in our case) the decoder will compare the result generated from the signal analysis module (MFCC) with different hypothesis using the both acoustic and linguistic modules; these too modules are previously trained using a tagged corpus. The module will provides the transcription corresponding to the most probable hypothesis.

[1] MFCC ; Mel Frequency Cepstral Coefficient

Segmentation. The role of this module is to segment statements transcribed by the speech recognition module. This treatment helps to identify the different requests of the speaker's message. The same message can consist of one or more requests at once. In this sense, it is necessary that the system can identify the different requests of the message, in order to interpret the user's request in its entirety.

Statements Pretreatment. The role of this module is to segment statements transcribed by the speech recognition module. This treatment helps to identify the different requests of the speaker's message. The same message can consist of one or more requests at once. In this sense, it is necessary that the system can identify the different requests of the message, in order to interpret the user's request in its entirety.

Semantics Decoding. It allows determining the meaning of each word in the statement.

Statement Meaning Construction. This module allows us to generate the entire attribute/value pairs.

Decoding pretreated semantic statements is based on a numerical model which encodes the rules of grammar, and on a semantic lexicon (see Figure 3 below). The semantic lexicon is a set of form associations: word/semantic features SEF describing the word meaning (see 1st definition)+ a set of syntactic feature SYF describing the word characteristics (gender, number and type). For example, the meaning of the word "الذاهب" (ranging) is described as follow:

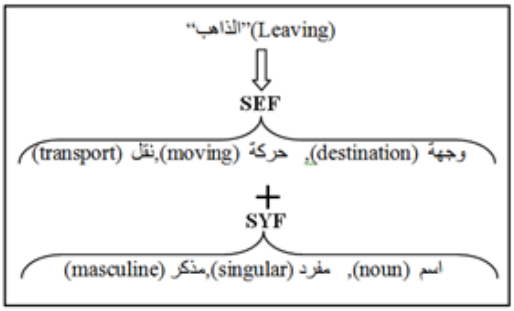

Fig. 3. Decoding pretreated semantic statements Using SEF and SYF

First definition: The semantic features set (SEF): An SEF is a set of three semantic features {D, C, and TM}, allowing the representation of the meaning of each non-empty word.

The first feature is concerning domain, referenced by "D", it can specify the finalized treated application domain, for example it can provide information on touristic or Railways information etc.

The second feature is concerning semantic class, referenced by "C", this feature specify the semantic class to which a word belongs to,This type of feature is used to

group synonymous words orwords having the same semantic role in a specific application.

The third feature is a differential one and “TM” references him. It allows to oppose words to each other and to specify in what way they have a different meaning.

In their study [8], the authors create an automatic system of understanding of Arabic spontaneous speech. However, the proposed system is just for railway statements. Just one area. While in this work, we have tried to generalize their work. The authors presented the world with the triple (D, C, and TM), where "D" present the Domain (Area), and for them the domain is fixed (D = 1). "C" shows the semantic classes, this parameter is used to group sets of synonymous words or words that are semantically similar. For authors in these classes are created manually for the rail sector. "TM" allows words to oppose to each other and to clarify how they have different meanings.
Our idea is to:

- Determine the maximum possible existing domains (religious, sport, economics, computer science, politics, social sciences ...). In this way "D" will be an integer greater than 1. This allows us to cover the maximum possible of domain.
- "C" shows the semantic classes of each domain. Here, since it is impossible to determine the classes of all fields manually (as is done by the authors for the rail sector). So we used an automatic segmenter. It is an approach that allows to group synonymous words in one class. Where we get a set of semantic classes for each domain.
- "TM" will keep the same role as in [5].

4 Formal Description

Systems based on the language models attempt to determine the numerical score of a word sequence (statement) S = m1, m2, ..., mi, with the general formula is described by the equation (1) below:

$$P(S) = P(m_1) \cdot P(m_2 / m_1) \cdot ... \cdot P(m_i / m_1, m_2, ..., m_{i-1}) \quad (1)$$

In the case of the interpretation of a significant words sequence [12] M1, M2 ,…Mn using a set of semantic features SEF1,SEF2,…,SEFn the model is trying to determine the score interpretation of each of these words, for each of these semantic features sets.

We denote by “I” the interpretation score that describe the meaning of the M1, M2, … Mn words respectively with the semantic features sets SEF1,SEF2,…,SEFn. I is measured as below:

$$P(I) = P(SEF_1 ... SEF_n / M_1 \dots M_n) \quad (2)$$

$$= P(SEF_1 / M_1) \cdot P(SEF_2 / SEF_1, M_1M_2) \cdot ... \cdot P(SEF_n / SEF_1 \dots SEF_{n-1}, M_1 \dots M_{n1})$$

$$= P(SEF_1 / M_1) \cdot P(SEF_2 / SEF_1, M_2) \cdot ... \cdot P(SEF_n / SEF_1 \dots SEF_{n-1}, M_n)$$

The transition from the first to the second line of the equation above is an approximation of the model, which considers that the probability of a SEF_i is conditionally dependent

only on the features sets of the current word M_i and not to those of the complete sequence. Fixing in advance the application domain, each significant words M_i can be interpreted through a set of semantic features using the form SEFi = (C_i, TM_i) and the above equation is transformed to equation 3 below:

$$P(I) \approx P((C_1,TM_1) / M_1) \times P((C_2,TM_2) / (C_1,TM_1),M_2) \times ... \times P((C_n,TM_n) / (C_1,TM_1) \times \ldots \times (C_{n-1},TM_{n-1}), M_n) \quad (3)$$

To minimize the number of candidates features sets SEF and improve the performance of semantic decoder, we integrated, into the equation above, other sources of information (illocutionary nature and type of the statement) which can participates in the selection of SEF. Thus, the probability of interpreting each word M_i by a given features set SEF_i = (C_i, TM_i), taking into account the nature and the type of statement noted by NT_j, is given by equation 4:

$$P(I) = P(SEF_1, ... SEF_n \mid NT_j, M_1 \ldots M_n) = P(NT_j / M_1 \ldots M_n) \times P((C_1,TM_1) / NT_j,M_1) \times P((C_2, TM_2) / NT_j, (C_1, TM_1), M_2) \times ... \times P((C_n,TM_n) / NT_j,(C_1,TM_1), \ldots (C_{n-1},TM_{n-1}),M_n) \quad (4)$$

In the equation above $P(NT_j / M_1, M_2, \ldots, M_{i-1}, M_i)$ is the probability that the statement is a NT_j type, knowing that the statement is made by the meaningful words M_1, ..., M_n. Note that the first word is treated separately, by annotating it using the default class *C= request*.

The figure below shows an example of meaning construction using the designed model.

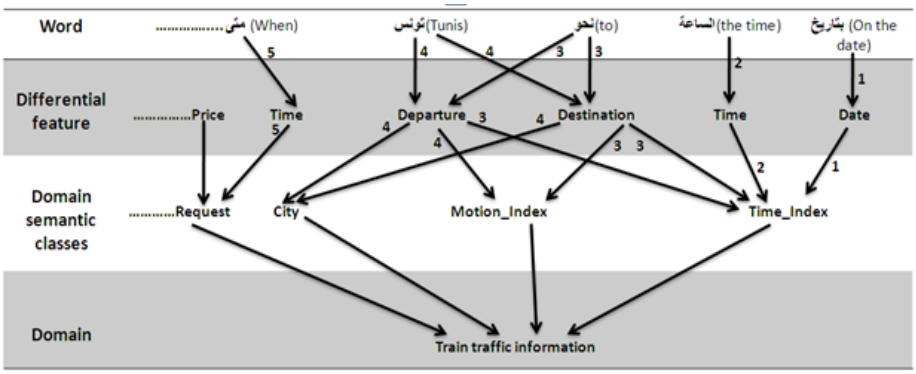

Fig. 4. Sample of statement meaning construction using our model

Figure 4 describes the structure and content of the semantic lexicon. The numbers used indicate the semantic features that can be used for the semantic representation of each word and each level of description.

5 Tests and Results

In order to evaluate our system we made an experiment using a corpus containing three domain; train reservation as the single domain used in [5] and we add to other domain; Book request in a library and touristic information as done in [9].

5.1 Corpus Collection

This corpus was collected by asking thirty-five different people to record vocal utterances relating to train reservation information request, touristic information request and book request. The following table provides information about the details about this task.

Table 3. Corpus collection details

Number of utterance	250
Number of speakers	35
Queries types	38
Data for Training	62%
Data for Test	38%

5.2 Evaluation

Some languages such as English, French, and German have platforms for evaluation understanding modules of dialogue systems. These platforms give to the community a large set of corpus of real annotated dialogues. However, this is not the case for the Arabic language where these resources are absents, with the exception of a few corpus distributed by ELDA/ELRA [10] [13]. The evaluation of corpus involves about 100 queries of different types (negation, affirmation, interrogation and acceptance), uttered spontaneously and manually transcribed. These requests correspond to scenarios dealing with information on the tourism fields. These scenarios are inspired from corpus MEDIA [11] and try to cover the input space The evaluation of the understanding module, with this evaluation corpus showed that this system generates 20 errors (average one error by 5 items). Measures of recall, precision and F-measure are respectively 72.00%, 69.00% and 75.69% and the average time to execute an utterance of 12 words is 0.279 seconds. Comparing these results with results obtained by other understanding modules [1], our system has provided fewer errors than many official sites such as UNISYS and MITRE.

6 Conclusion

We present in this paper a semantic decoder based on a hybrid language model, which allows integrating contextual, lexical, and semantic and illocutionary information at the same time. It allows, moreover, considering only the relevant sets of semantic features "SEF" in the history of the word to interpret. For this, we have developed a method to

automatically extract the relevant SEF which describe the meaning of words with semantic influence on the word to interpret. This is achieved, based on the concept of average mutual information brought by Rosenfeld (Rosenfeld, R. 1996). We intend eventually to evaluate our model by comparing it with other deployed models as models obtained by linear combination of language models well known as the maximum entropy.

References

1. Minker, W. (1999). *Compréhension automatique de la parole spontanée*. Editions Le Harmattan.
2. Saidane, T., Zrigui, M., & Ahmed, M. B. (2005, May). Arabic speech synthesis using a concatenation of polyphones: the results. In *Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 406-411). Springer Berlin Heidelberg.
3. Mallat, S., Zouaghi, A., Hkiri, E., & Zrigui, M. (2013). Method of lexical enrichment in information retrieval system in Arabic. *International Journal of Information Retrieval Research (IJIRR)*, *3*(4), 35-51.
4. Maraoui, M., Antoniadis, G., & Zrigui, M. (2009, July). SALA: Call System for Arabic Based on NLP Tools. In *IC-AI* (pp. 168-172).
5. Charhad, M., Zrigui, M., & Quénot, G. (2005, March). Une approche conceptuelle pour la modélisation et la structuration sémantique des documents vidéos. In *SETIT-3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, TUNISIA*.
6. Trigui, A., Maraoui, M., & Zrigui, M. (2010). Acoustic Study of the Gemination Effect in Standard Arabic Speech. *IPCV*, *2010*, 192-196.
7. Trigui, A., Maraoui, M., & Zrigui, M. (2010, June). The gemination effect on consonant and vowel duration in standard Arabic speech. In *Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing (SNPD), 2010 11th ACIS International Conference on* (pp. 102-105). IEEE.
8. Zouaghi, A., Zrigui, M., & Antoniadis, G. (2008). Compréhension automatique de la parole arabe spontanée. *Une Modélisation Numérique, Traitement Automatique des Langues (TAL 2008)*, *49*(1), 141-166.
9. Lhioui, C., Zouaghi, A., & Zrigui, M. (2013). A combined method based on stochastic and linguistic paradigm for the understanding of arabic spontaneous utterances. In *Computational Linguistics and Intelligent Text Processing* (pp. 549-558). Springer Berlin Heidelberg.
10. Bahou, Y., Belguith, H. L., & Ben Hamadou, A. (2008). Towards a human-machine spoken dialogue in Arabic. In 6th Language Resources and Evaluation Conference (LREC 2008), Workshop HLT Within the Arabic World. Arabic Language and Local Languages Processing Status Updates and Prospects, Marrakech, Morocco.
11. Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., & Mostefa, D. (2005). Semantic annotation of the french media dialog corpus. In *Ninth European Conference on Speech Communication and Technology*.
12. Zouaghi, A., Zrigui, M., & Ben Ahmed, M. (2005). Un étiqueteur sémantique des énoncés en langue arabe. In Actes de la 12ème Conférence sur le Traitement Automatique des Langues Naturelles (TALNRECITAL 2005), Dourdan, France.
13. Villaneau, J., Ridoux, O., & Antoine, J. Y. (2004). LOGUS: compréhension de l'oral spontané. *Revue d'Intelligence Artificielle (RIA)*, *18*(5-6), 709.