Comparison between different frameworks for speech understanding of under-resourced languages

M. Graja,
College of Computer and Information Sciences,
Jouf University, KSA;

M. Jaoua
Miracl Laboratory, ANLP-RG, Sfax University, Tunisia

L. Hadrich Belguith
Miracl Laboratory, ANLP-RG, Sfax University, Tunisia

Summary
In this paper, we propose to compare four methods for speech understanding of under-resourced languages. The first method is knowledge based method which integrates ontology. The second one is statistical method which integrates CRF discriminative models. The third one is hybrid method which uses CRF models with integration of knowledge base. The fourth one is pattern based method. We have used a spoken Tunisian dialect corpus acquired and annotated to perform experiments. The evaluation is based on semantic representation generated by each method. The obtained results shows that the hybrid method is the best one compared to others, which proves that CRF models with ontology integration is suitable for under-resourced languages.

Key words:
Speech understanding, under-resourced language, knowledge base, domain ontology, pattern, CRFs.

1. Introduction

This work is part of the research work on human-machine oral dialogue and proposes to design an automatic understanding system of spontaneous speech. Automatic speech understanding constitutes the key link in a dialogue system since it makes it possible to clarify the meaning of the utterance by determining a semantic representation understandable by the machine [1][2]. It can also be defined as the process that generates the meaning of an utterance through the concept detection of the application. It could be also considered as the process of association between words and concepts of the domain, thus making it possible to generate a semantic representation of the useful meaning of oral utterances [3].

Speech understanding requires two steps. The first is based on concept understanding and consists in translating the current statement into a language of concepts. While the second step is to identify the semantic structure by transforming the set of concepts obtained, in the first step, into a semantic representation used by the dialogue manager [4] [2]. This second step can be defined as a representation converter [5]. The purpose of the semantic representation is to explicitly translate the meaning of an utterance, so it could be "understandable" by the dialogue manager.

To provide the meaning of an utterance, several semantic representation formalisms have been proposed in the literature. We cite as examples conceptual graphs, attribute/value pairs, logical formulas, semantic schemes, and FrameNet. Therefore, the choice of an explicit and efficient semantic representation is a critical task. It should be also noted that this representation does not obey to any standard and may vary according to the need of the intended application. For example, if it is a query of a database, the semantic representation in the form of an SQL query is preferred like the PEGASUS understanding system for air transport information [6]. The diversity of semantic representation makes difficult the evaluation and comparison of speech understanding systems [7].

In addition, the diversity of semantic representation formalisms makes it difficult to choose the representation that faithfully conveys the meaning of an oral utterance. However, the introduction of FrameNet as a unified frame network proves the interest of frame-based semantic representation [8] [9]. It is a representation able to represent rich dialogical information [8]. However, the FrameNet offers generic frames that may not meet the needs of the application [8]. Thus, it was essential to use a step of translation and adaptation of the FrameNet if we plan to exploit it for a specific dialect in a limited task. These facts direct us towards the choice of a frame-based representation.

In this work, we consider speech transcribed in textual form as the input of our understanding system. The semantic representation generated by our speech understanding system is analogous to that chosen by Bahou [10] and it is based on semantic schemas. This choice is motivated by the prospect of exploiting the same dialogue manager developed within our ANLP-RG research group. Finally, we consider the dialect as real form of

Manuscript received October 5, 2021
Manuscript revised October 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.10.22

communication for Arabic speakers and it is widely used to fulfill public services. Therefore, it is so crucial to consider dialects which are still quite processed in dialogue system. The Tunisian dialect (TD) is a representative example for Arabic dialect and it represents under-resourced language.

The main contribution in this work is to compare four methods for speech understanding of Tunisian dialect (TD). The evaluation of each method is based on the final output, which is semantic representation which is based on Frame. The first speech understanding method is a knowledge based method, which integrates domain ontology. The second method is based on CRF models learned from a little size corpus. The third method is a hybrid method, which integrates CRF models with ontology. The fourth method is pattern-based method which is an adaptation to the TD of an existing method already used for MSA (Modern Standard Arabic). The corpus used in this work is the TUDICOI corpus. It is a task-oriented spontaneous speech dialogue corpus in TD about railway request information.

This paper is organized as follows. Section 2 describes a spoken TD corpus. Section 3 the ontology-based method. Section 4 presents the CRF-based method. We present the hybrid method in section 5. Section 6 presents the patterns based method. Section 7 presents the evaluations measures. We present in section 8 the training and test corpus used in our experiments. Results and discussion are shown in section 9. The conclusion is drawn in the last section.

1. Spoken Tunisian dialogue corpus

Building a corpus of dialogue is a big challenge, especially when it comes to an under-resourced language that lacks resources [11][12]. It is for this reason that we have developed our oral dialogue corpus called TuDiCoI (Tunisian Dialect Corpus Interlocutor) for a limited task. This corpus was collected in collaboration with the Tunisian National Railway Company (SNCFT). It is about railway request information.

Table1: Main characteristics of the TUDICOI Corpus

# Dialogues	1825
# Speakers	1831
# Client turns	6533
# Staff turns	5649
# Words in client turns	21682

We performed a manual annotation in terms of semantic concepts. A concept is a semantic label attributed to a word or set of words in an utterance to express a minimal unity of meaning. The annotation scheme used to annotate the TUDICOI corpus is inherited from the Interchange Format (IF) [14]. We performed a two-level annotation. The first level deals with dialogue acts which covers the general intention of the utterance. The second level is intended to give more specific information about the task. Table 2 shows dialogue acts used in the first level and Table 3 shows semantic concept labels used to label our corpus.

Table 2: Dialogue acts used in the first level

Dialogue act	Example	Translation
Opening	عسلامة ، صباح الخير	Hi, good morning
Closing	بالسلامة	Bye
Undefined	شي ما عاد عندي حتى فرنك	I don't have any penny
Waiting	استنى شوية	Wait a bit
Request Information	وقتاش يخرج التران	When the train leaves
Acceptance	باهي خليهالي	Ok, leave it for me
Rejection	لالا	No No

Table 3: Semantic Concept labels used in the second level

Domain concepts		Requests concepts	
Train	Ticket_Numbers	Path_Req	Existence_Req
Train_Type	Ticket	Hour_Req	Trip_timeReq
Departure_hour	Hour_Cpt	Price_Req	Clarification_Req
Arrival_hour	Departure_Cpt		Booking_Req
Day	Arrival_Cpt	Dialogue concepts	
Origin	Price_Cpt	Rejection	Salutation_Begin
Destination	Class_Cpt	Acceptance	Salutation_End
Fare	Trip_time	Politeness	
Class	Ticket_type		
Link concepts		Out of vocabulary (Out_Vocab)	
Choice Coordination			

Given the complexity and time-consuming of the manual annotation task, we have only annotated 1476 dialogues. These dialogues consists of 5047 client turns (Table 4).

Table 4: Main characteristics of the annotated corpus

# Annotated dialogues	1476
# Annotated client turns	5047
# Annotated words in client turns	16772

2. Ontology based method

Since we deal with a limited task, we modeled the domain knowledge by an ontology. To build domain ontology, we have proposed in previous work a hybrid method for semi-automatic construction of domain ontologies. We generate the RIO ontology for our task (Railway Information Ontology) [13]. In order to integrate the RIO ontology into the speech-understanding module, we defined a method that distinguishes four stages (Figure 1).

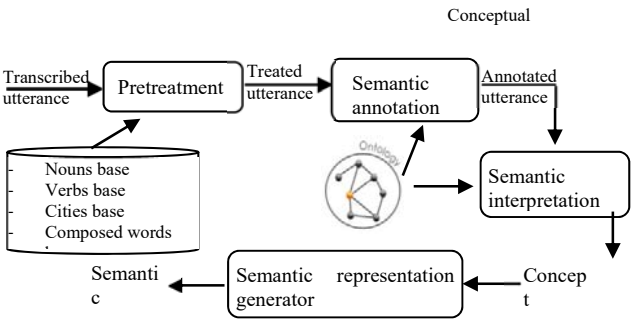

Figure 1. Ontology-based method for speech understanding of the TD

- Pretreatment: this step consists in treating the client's turn in order to reduce the structure complexity and standardize words.
- Semantic annotation: it consists of labelling utterances based on the RIO ontology concepts. It should be noted that it is possible that a word can be labeled by two concepts or by one concept, or may be labeled as "Out" if we do not find the word in ontology.
- Semantic interpretation: it helps improving semantic annotation by exploiting the semantic relations of the RIO ontology.
- Semantic representation generator: this step is responsible to generate a semantic schema. A semantic schema is defined by its name, by one or more reference concepts and a list of concepts. The schema name is the type of the request, while the reference concepts facilitate the detection of the schema. The semantic schema are instantiated by the words of the utterance. Figure 2 illustrates the general structure of a semantic scheme.

After analyzing our corpus, we determined a list of semantic schemas that respond to the utterances expressed by the client. It should be noted that an utterance can contain several requests identified through the reference concepts, which generates a set of semantic schemas. In the absence of a reference concept, the default schema is selected. This last case occurs when an utterance depends on the dialogical context where the reference words are explained in the previous utterances of the dialogue.

```
<Scheme_Name>
 <Reference_Concept/>
    <Concept1 />
    <Concept2 />
    :
</Scheme_Name >
```

Figure 2. General structure of semantic scheme

The step of generating the semantic representation results from a simple conversion from the conceptual representation to the corresponding scheme. If we consider the following example " وقتاش تران إكسبراس يخرج من صفاقس/ "When the Express train leaves Sfax?". The result of the semantic annotation is as follows:

صفاقس	من	يخرج	إكسبراس	تران	وقتاش ◄
SfaAqis	min	yuxrij	ĂkspraAs	traAn	wqtaAš
Origin	Semantic Rel	Semantic Rel	Train_Type	Train	Dep_Hour_ Req

In this annotation, the Dept_Hour_Req is the reference concept which instantiate the Dept_Hour_Req scheme. The schema filling process consists of aligning the identified concepts with the different cases of the schema. The result of this alignment is an instance of the Dept_Hour_Req represented by Figure 3.

```
<Scheme_Dep_Hour_Req>
    <Reference_concept>
        Dep_Hour_Req
    </Reference_concept>
    <Origin>صفـاقس</Origin>
    <Destination />
    <Train_Type>إكسبـرا<Train_Type/>
    <Day />
    <Dept_Hour />
    <Arrival_Hour />
</Scheme_Dep_Hour_Req>
```

Figure 3. Instanciation of Dep_Hour_Req scheme

It should be noted that this step of the semantic representation generation constitutes a common dominator for all the methods of understanding that we will describe later in this paper. For this reason, we are limited to describe only the conceptual decoding steps of the different proposed methods.

3. CRF Based Method

In our work, we are interested in the CRF (Conditional Random Fields) models, which are, up to now, the most successful and most used models for conceptual labeling [19].

In order to experiment these models for the TD, we used the same turns of speech coming from the TuDiCoI corpus and which are exploited by the method based on the ontology. It should be noted that these speech turns are not segmented into utterances, unlike almost all understanding methods based on stochastic models. To our knowledge, the only work that use unsegmented turns is the work of Marintèze et al. [24] which exploits the HMM models. The idea of unsegmented speech turns derives from the fact that the speech recognition module outputs unsegmented speech turns, which further facilitates the transcription task and the segmentation task dialogues.

To use the CRF models, we performed a learning step followed by a test step.

Learning CRF models consists in estimating the parameter vector θ = (λ1, λ2, ... λjn,μ1, μ2, ..., μkn) from the training data (x(i), y(i)), i=1..N based on the following model:

$$p(y/x)=\frac{1}{z(x)}\exp\left(\sum_j \lambda_j t_j(y_{i-1},y_i,x,i)+\sum_k \mu_k s_k(y_i,x,i)\right) \quad (1)$$

With:

$$z(x)=\sum_y \exp\left(\sum_j \lambda_j t_j(y_{i-1},y_i,x,i)+\sum_k \mu_k s_k(y_i,x,i)\right) \quad (2)$$

z(x) is the normalization factor that makes the sum of all probabilities equals to one. $t_j(y_{i-1},y_i,x,i)$ represents a

transition feature function of the entire observation sequence and the labels in positions i and i-1 in the label sequence. $s_k(y_i,x,i)$ represents state feature function of the label in position i in the observation sequence. λ_j and μ_j are parameters which are estimated from training data.

Given this model defined in the Equation 1, the most probable labeling sequence y* for an input x, is:

$$y^* = \arg\max_y p(y / x) \quad (3)$$

In the context of conceptual labeling and in order to take into account the dependence of the words (or group of words) of the same utterance, we can adopt several models that combine the n-grams of adjacent words. For each adopted CRF model, an estimate of underlying parameters is required based on a prior learning step. In order to achieve this, pretreatment and manual conceptual labeling sub-steps are required.

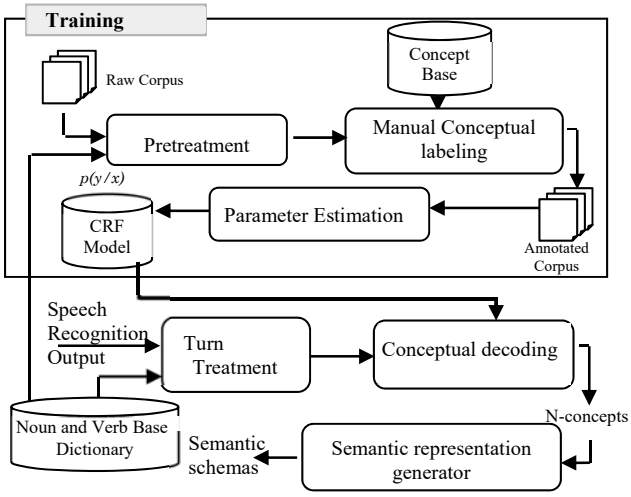

Figure 4. CRF based method for speech understanding of TD

4. Hybrid method

Ontologies bring together the knowledge of a domain and connect it semantically. However, ontology can alter understanding in the absence of an explicit semantic relationship in the utterance. This particular case can be solved by exploiting stochastic models given their ability to model the dependencies between words in the utterance. On the other hand, the erroneous annotations generated by the CRF models result from words rarely encountered in the corpus. So, ontology can resolve this problem. For this, we have proposed a method of understanding the TD which is based on the coupling between the CRF models and the domain ontology [15].

The hybrid method that combines the CRF models and the domain ontology inherits the same steps as the one based on the CRF models. In addition, it includes a new step entitled "Knowledge Integration" which consists of exploiting the ontology if decoding failures. Failure at this level results in the assignment of the "Out" label (Out_Vocab). The knowledge integration step is integrated after the “Conceptual Decoding Step”. This integration is used to improve the result of the automatic conceptual decoding step by integrating domain knowledge. This knowledge is either general knowledge (days of the week, months and numbers) or knowledge specific to the railway information field (train schedules, cities and train stations, etc.). This step uses the semantic relations of the ontology to better interpret the words identified by this step.

5. Pattern based method

Bahou [10] conducted a research work on MSA dialogue systems. This work proposes a method of literal understanding of the MSA within the SARF system (Arab Voice Information System for Railway Transport). SARF's method for speech understanding is part of the semantic approach and it is based on the formalism of case grammars for the generation of semantic schemas. We focused on adapting this work to propose a speech understanding method based on patterns. The adaptation that we propose is to explicitly replace the resources of the SARF system with those of the TD and aims to compare the performance of the methods we have proposed for understanding in TD with that translated from the MSA.

The proposed adaptations are to replace lexical resources and treatments that are related to the MSA language without affecting the body of the proposed method based on patterns. During these adaptations, we exploited our TuDiCoI corpus to generate the lexicon, the patterns and the conceptual segments [16]. In the following, we describe the different adaptation of the SARF system:

- Adaptation of lexical resources: the lexical database, extracted from the TuDiCoI corpus that we have integrated into the SARF system, contains the names base, compound word base and the verb base. We have also integrated a thesaurus built from these three bases.
- Adaptation of conceptual segments: because of the difference of utterance structure between MSA and TD, we adapted conceptual segments by adding, deleting, or changing the order of concepts.
- Adaptation of treatments: some additional treatments need to be considered in order to take into account the specificities of the TD in relation to MSA.

6. Evaluation measures

In all our experiments, we will use the same part of the TuDiCoI corpus for evaluation and adopt the same

evaluation measures. Our evaluation aims to measure the overall performance (Global evaluation) of the speech understanding system in terms of acceptability of understanding. This is to measure the accuracy of the semantic schemas generated by the whole system (black box) [17]. These approaches use assessment measures that have been developed to compare the understanding methods between them.

Bahou [10] proposed other global assessment measures to calculate False-understanding and Acceptable understanding. The False-understanding consists of calculating the error rate in terms of utterances that have generated incorrect semantic representations.

$$False - understanding = C.I + C.E \quad (4)$$

With C.I indicates the number of utterances that generate incomplete semantic representations and C.E indicates the number of utterances that generate erroneous semantic representations.

Acceptable understanding is calculated by the following formula:

$$Acceptable\ understanding = C.C + C.I \quad (5)$$

With C.C denotes the number of utterances that generate complete semantic representations and C.I indicates the number of utterances that generate incomplete semantic representations.

It should be noted that these global evaluation measures have been adopted to ensure the possibility of comparing the results obtained with the method based on patterns adapted from MSA to TD.

7. Training and test corpus

For our evaluation, we divided the annotated TuDiCoI corpus into two parts. The first one is used for learning step. It represents about 80% of the total size. While the second part represents 20% of the corpus used for the test. Table 5 provides features of training and test corpus.

Table 5: Features of training and test corpus

	Training corpus	Test corpus
Dialogues	1202	267
Turns	4131	906
words	13555	3217

In our work, we classified all the turns of the test part into three types, according to the recommendation proposed by the ARPA community [18] namely the series A, D and X. Tables 6 and 7 describe the characteristics of these different series. This classification provides an overview of the types of statements contained in the test portion.

Table 6: Characteristics of the test set over three series A, D and X

Client turn number			
A	**D**	**X**	**Total**
379	482	45	906
41.83%	53.21%	4.96%	100%

Table 7: Characteristics of the evaluation corpus

#Dialogues 267	#User words 3217	Client turns			
		A	**D**	**X**	**Total**
		379 41.83%	482 53.21%	45 4.96%	906

The first set (Series A) corresponds to context-independent client speech turns. This set contains oral utterances that do not relate to the dialogue history. In the second set (Series D), utterances correspond to those dependent on the context. This set contains oral statements that relate to the dialogical context. The third set (X Series) corresponds to out of context statements of the dialogue. It includes the marginal utterances that have no relation to the domain.

8. Results and Discussion

In order to compare the different methods, we present in Table 8 a summary of the results obtained by the different methods proposed.

Based on results obtained by the speech understanding method based on adapted patterns for the TD, we noticed that this method recorded the lowest rates compared to the results obtained by all the proposed methods. This is due to the difficulty of modelling patterns for all TD utterance structures. In fact, TD utterances are stretcher than those in MSA.

Examination of the results obtained by the RIO ontology-based method, we notices also a very high rate of False-Understanding. The cases of failure result mainly from the absence of semantic relations in the statement. As an example, we noticed that several city names were annotated by two different concepts (the Origin concept and the Destination concept).

Table 8: Comparison between the different proposed methods

	Global evaluation				
	C.C	**C.I**	**C.E**	**False-Unders**	**Acce. Unders**
Pattern based	52.81%	20.51%	26.68%	47.19%	73.32%
Ontology based	45%	45.76%	9,23%	55.00%	90.76%
CRF based	78.46%	16.01%	5.51%	21.53%	94.48%
Hybrid	81.70%	13.41%	4.87%	18.29%	95.12%

Problems encountered with pattern based method and ontology based method is fixed with CRFs models. In fact, CRFs are able to estimate the probability of a word appearing in its neighborhood even in the absence of

semantic relation in the utterance. These models have shown a robustness against noisy data. These models increase in the acceptable understanding to 94.48%. These good results using CRFs models have been obtained for a little size corpus which confirms the performance of these models for under-resourced languages.

However, we have noticed that CRF models cannot correctly label low frequency words which are rarely encountered in the training corpus. So, the idea behind the hybrid method is to integrate knowledge from the ontology. This is made it possible to overcome the limitations of the CRF models by identifying the words rarely used in the training corpus. This integration has reduced the number of incorrect predictions which reduces the error rate. As consequence, we improve the overall evaluation.

Through a manual examination of the generated semantic schemas for all proposed methods of speech understanding, we observed that certain analyzes are failed due to the presence of out-of-vocabulary and the presence of utterances typed D and X. The presence of out-of-vocabulary words do not help to detect the correct semantic schema to represent correctly the user needs. In addition, the existence of out-of-context (X-type) utterances generates errors in semantic schemas identification, resulting in a misunderstanding of utterance, since the default schema will be generated. Finally, type D utterances produce a misidentification of the speaker's request needs since they represent utterances which rely on the dialogical context.

9. Conclusion

This paper encapsulated all the experiments conducted to evaluate the different methods for speech understanding of the TD. We observed that the results of the pattern-based method are insufficient because of the complexity of modeling all possible patterns for dialect utterances. Similarly, we found that the CRF models do not allow consider words seldom used in the training corpus. While the ontology based method cannot interpret the words in the absence of semantic relations. It is for this reason that hybridization has made it possible to improve the results obtained by identifying the dependence between the words by using CRF models and domain knowledge on the other hand. We can conclude that the hybrid method proposed for speech understanding is adequate for a limited domain and makes it possible to overcome the lack of linguistic resources especially for under-resourced language.

References

[1] H. Aust, M.Oerder, F. Seide and V. Steinbiss, “The Philips automatic train timetable information system”, Speech Communication. Vol. 17, no. 3, pp. 249_262, Nov. 1995.

[2] F. Bechet and A. Nasr, “Robust Dependency Parsing for Spoken Language Understanding of Spontaneous Speech”. In proceeding of the international speech communication association (Interspeech), Sept. 2009.

[3] D. Lee, M. Jeong, K. Kim, S. Ryu, and G. Lee, “Unsupervised Spoken Language Understanding for a Multi-Domain Dialog System”, IEEE transactions on audio speech and language processing, Vol. 21, no. 11, pp. 2451_2464, Nov. 2013.

[4] Y. Y. Wang and A. Acero, “Discriminative Models for Spoken Language Understanding”. In proceeding of international conference on spoken language processing (ISCA). 2006.

[5] S. Jamoussi, “Méthodes Statistiques pour la Compréhension Automatique de la Parole". Thèse de Doctorat, université Henri Poincaré. 2004.

[6] V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goddeau, J. Glass and E. Brill, “ Pegasus: A Spoken Language Interface for On-Line Air Travel Planning”, Speech Communication, Vol. 15, pp.331-340. 1994.

[7] B. Jabaian, “Systèmes de Compréhension et de Traduction de la Parole : Vers une Approche Unifiée dans le Cadre de la Portabilité Multilingue des Systèmes de Dialogue", Thèse de doctorat, Université d’Avignon, 2012.

[8] M.-J. Meurs, F. Duvert, F. Béchet, F. Lefevre, and R. de Mori, "Annotation en Frames Sémantiques du corpus de dialogue MEDIA". Les actes de la conférence sur le traitement automatiques des langues naturelles (TALN), 2008.

[9] J. Trione, F. Bechet, B. Favre, and A. Nasr, “Rapid FrameNet annotation of spoken conversation transcripts”, Joint ACL-ISO Workshop on Interoperable Semantic Annotation, 2015.

[10] Y. Bahou, "Compréhension Automatique de la Parole Arabe Spontanée: Intégration dans un Serveur Vocal Interactif". Thèse de doctorat, Université de Sfax, 2014.

[11] N. Habash, M. Diab, and O. Rambow, “Conventional Orthography for Dialectal Arabic”. In proceeding of the international conference on language resources and evaluation (LREC), 2012.

[12] R. Zbib, E. Malchiodi, J. Devlin, D. Stallard, S. Matsoukas, R. Schwartz, J. Makhoul, O.F. Zaidany, and C. Callison-Burch, “Machine Translation of Arabic Dialects”. In proceeding of the conference of the North American chapter of the association for computational linguistics: Human Language Technologies (HLT-NAACL), 2012.

[13] C. D. Martínez-Hinarejos, J. M. Benedí, and R. Granell, “Statistical Framework for Spanish Spoken Dialogue Corpus”, Speech communication, Vol. 50, no.11, pp.992_1008, Nov. 2008.

[14] N. Alcacer. J. Benedi. F. Blat. R. Granell. C.D. Martinez. F. Torres. "Acquisition and labeling of a spontaneous speech dialogue corpus". In Proc. of 10th International Conference on Speech and Computer (SPECOM). Patras. Greece. 2005.

[15] M. Graja, M. Jaoua, and L. H. Belguith, “Statistical Framework with Knowledge Base Integration for Robust Speech Understanding of the Tunisian Dialect”, IEEE/ACM Trans. on Audio, Speech, and Language Processing, Vol. 23, no. 12, pp. 2311_2321, Dec. 2015.

[16] W. Neifar, Y. Bahou, M. Graja and M. Jaoua, “Implementation of a Symbolic Method for the Tunisian Dialect Understanding”. In proceeding of the international conference on Arabic language processing (CITALA), 2014.

[17] R. Zajac, S. Helmreich and K. Megerdoomian, “Black-Box/Glass-Box Evaluation in Shiraz”. In proceeding of the workshop on machine translation evaluation at the international conference LREC, 2000.
[18] W. Minker, and S. Bennacef, “Speech and Human Machine Dialog”. The springer international series in engineering and computer science, 2004.
[19] I. Touati, M. Graja, M. Ellouze and L. Belguith, “Opinion Target Extraction from Arabic News Articles Using shallow Features”. In proceeding of the Mediterranean Conference on Pattern Recognition and Artificial Intelligence MedPRAI, 2018.

M. GRAJA: received her PhD degree respectively in 2015 from University of Sfax. Currently, she is assistant professor in College of Computer and Information sciences, Jouf University-KSA. Her research interests include NLP, dialogue systems and opinion summarization.

M. JAOUA: received his PhD degree in computer science from the University of Tunis in 2004. He is currently an assistant professor - University of Sfax (Tunisia). His main research activities are text summarization and dialogue systems.

L. HADRICH BELGUITH: received her PhD degree in 1999 from University of Tunis. Currently, she is a professor in Computer Science and Management - University of Sfax - and the head of the ANLP-RG of MIRACL laboratory. Her research activities are Arabic text analysis, Automatic Abstracting, Question-Answering systems, and spoken dialogue system.