Semantic Language Model for Tunisian Dialect

Abir Masmoudi
MIRACL Laboratory
University of Sfax
masmoudiabir@gmail.com

Rim Laatar
MIRACL Laboratory
University of Sfax

Mariem Ellouze
MIRACL Laboratory
University of Sfax
Mariem.Ellouze@planet.tn

Lamia Belguith
MIRACL Laboratory
University of Sfax

Abstract

In this paper, we describe the process of creating a statistical Language Model (LM) for the Tunisian Dialect. Indeed, this work is part of the realization of Automatic Speech Recognition (ASR) system for the Tunisian Railway Transport Network. Since our field of work has been limited, there are several words with similar behaviors (semantic for example) but they do not have the same appearance probability; their class groupings will therefore be possible. For these reasons, we propose to build an n-class LM that is based mainly on the integration of purely semantic data. Indeed, each class represents an abstraction of similar labels. In order to improve the sequence labeling task, we proposed to use a discriminative algorithm based on the Conditional Random Field (CRF) model. To better judge our choice of creating an n-class word model, we compared the created model with the 3-gram type model on the same test corpus of evaluation. Additionally, to assess the impact of using the CRF model to perform the semantic labelling task in order to construct semantic classes, we compared the n-class created model with using the CRF in the semantic labelling task and the n-class model without using the CRF in the semantic labelling task. The drawn comparison of the predictive power of the n-class model obtained by applying the CRF model in the semantic labelling is that it is better than the other two models presenting the highest value of its perplexity.

1 Introduction

Generally, the development of such ASR system for a specific language requires first and foremost the construction of a large speech corpus. This corpus must be based on both an orthographic and a phonetic transcription. In addition, textual data for learning the LM of the system are also required. Nevertheless, these resources are not available directly for Arabic dialects. As a result, the ASR system development for Arabic dialects is fraught with many different kinds of difficulties that it faces. In this perspective, this work is integrated in the field of developing the Tunisian Dialect ASR system for the Tunisian Railway Transport Network. More precisely, we are interested in this paper in presenting our method for constructing a LM; one among the essential components of the ASR system. This model proposes to define the probability distribution on sets of word sequences. Due to the lack of learning data of the Tunisian dialect, it is necessary to find a method that maximizes the amount of information. This corresponds to the appearance of the n-class LM. The main idea of this model is to classify vocabulary words into lexical classes and to calculate the probability of a sequence of words, such as the probability of a sequence of lexical classes (5).

The primary contributions of this paper are as follows:

- Gathering our TARIC (Tunisian Arabic Railway Interaction Corpus) corpus to realize an ASR system for

Proceedings of Recent Advances in Natural Language Processing, pages 720–729,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_084

the Tunisian Railway Transport Network. This corpus is based on speech transcriptions. In order to obtain a standardized and normalized corpus, we employed our Tunisian Dialect CODA (Conventional Orthography for Dialectal Arabic) (30).

- We present our proposed method of creating the n-class LM. To achieve this, we show the different stages of our method.
- We evaluate the performance of the discriminative algorithm based on the CRF model in order to perform the semantic labeling task for spontaneous speech in the Tunisian Dialect in the n-class LM construction context.
- Testing the impact of using CRF model in the semantic labelling field. It is also important to assess its effect on creating semantic classes on the one hand, on developing an n-class LM and on its perplexity level on the other hand.
- We eventually disclosed the elaborated experiments we went through and the obtained results.

The remainder of this paper is organized as follows: Section 2 tries to show the main role of ASR system and its components. We also shed light on the main types LM. Section 3 discusses the related work in this field summarizing the main aspects of every work. Section 4 describes the TARIC corpus in Tunisian Dialect used for experiments. In section 5, we present the CRF discriminative model used to perform semantic labeling then, we explain our method to construct n-class LM for Tunisian Dialect. Then Conclusion is drawing in the last section.

2 ASR system and language model

An Automatic Speech Recognition (ASR) system aims to transcribe textual data of a speech signal. Indeed, in the context of statistical modeling of speech, an ASR system is composed of acoustic model, language model (LM) and phonetic dictionary. As our works aim at suggesting a method in order to construct a LM, we provided further explanations of this component by concentrating on its different types that have been proposed in the literature. As we mentioned earlier, the LM seeks to represent the language behavior in order to confirm or refute the propositions made by the acoustic module. In literature, several statistical LM types are recognized as being the most efficient models in ASR. Among these models, we can cite n-gram, sequences of n-words and factorial LM. We will present some examples of the aforementioned approaches in the following sub-sections.

2.1 N-gram model

Thanks to their simplicity and efficiency, n-gram models are the most widely used LM in the speech recognition field. They are based on the assumption that the appearance of a single word depends only on its history. In practice, estimating this probability is very difficult. In fact, no learning corpus can make it possible to observe all the sequences of possible words. As a result, the basic idea of n-gram models consists in considering only the sequences of words of length n, i. e the calculation is approached by a limited history consisting of the n-1 preceding words. The major drawback of this modeling type is that it can lead to assigning a zero probability to any n-gram that has never been encountered in the training corpus. This problem is serious especially when this n-gram could be perfectly valid in linguistics.

2.2 N-class model

Due to the lack of the learning corpus, it is vital to find a method that maximizes the amount of useful information on the one hand and reduces the model parameter space on the other hand. In order to meet this requirement, other methods have emerged recently. They consist in grouping words into classes. This corresponds

to the appearance of n-class LM. The chief idea of this model is to group vocabulary words into lexical classes and to estimate the probability of the word sequence, such as the probability of a sequence of lexical classes (5). One of the clearest conception of motivations for n-class models is that a word of a given class, not necessarily found in the learning corpus, inherits the probability of all the other representatives of its class. In addition, it is possible to add words to classes without having to re-estimate the model probabilities. However, the problems faced by n-class models are numerous. The first major difficulty is that this type of model requires the need for a prelabeled learning corpus (24). Nevertheless, manual labeling is particularly heavy despite its exact results.

2.3 Factorial LM

The factored LM is based on the principle that a word is no longer seen as a simple graphic chain but rather as a vector of characteristics (3). These characteristics can include the lemma and the grammatical class of a word, morphs, its kinds, its numbers, or Booleans indicating the word belonging to given semantic classes. On the theoretical side, factorial models have already shown good results for some tasks, such as machine translation (14). At the practical level, few works have relied on this model, especially in the speech recognition task.

To conclude, due to the limitation of our field of work "Tunisian Railway Transport Network", several words with similar behaviors exist, (semantic for example) but they do not have the same probability of appearance; their class groupings will therefore be possible. Moreover, the amount of learning data is reduced. In this context, the use of the n-class model is beneficial at several levels. For these reasons, we propose to build an n-class LM that is based mainly on the integration of purely semantic data. Indeed, our method will be used to create a LM based on semantic information for the creation of word classes. The figure 1 shows some words with the same semantic behavior.

Train /القِطَارْ	القِطَارْ , قِطَارْ ,أوتُورَايْ ,القِرِيْو
Ticket /بِيَايْ	بِيَايْ ,التَذْكِرَةْ ,التِسْكِرَةْ ,تِكَايْ
Schedule /تَوْقِيتْ	لُورَارْ ,تَوْقِيتْ,لَايْ-زُورَارْ,الاوْقَاتْ
First class /بِرُومْيَارْ - كْلَاسْ	فَرْسْتْ- كْلَاسْ ,دَرَجَةْ-أولَى ,بِرُومْيَارْ- كْلَاسْ
round trip /اَلَايْ-رُتُورْ	وَإِيَّابْ ذَهَابْ ,اَلَايْ-رُتُورْ

Figure 1: some words with the same semantic behavior.

3 Related Work

In this Section, we are going to review the existing works related to classify vocabulary words for the construction of an n-type class LM. In the context of training classes of words, (10) proposes a simple word classification algorithm for statistical LM in speech recognition. The classification criterion used in this approach is the similarity of words. Indeed, the principle is based on the criterion of substitution or replacement. According to this algorithm, two words are similar since they can be substituted in the learning corpus (10). According to this automatic word classification approach, the word accuracy rate was increased by 8.6% with a reduction in perplexity of about 6.9% (10). The method proposed by (8) is essentially based on the principle of combining different sources of information at the class formation level. In his work, (8) uses two types of information: contextual information and prior information. The former is the most commonly used, corresponds to n-gram dependencies. This information can be collected not only at the words level, but also at the level of previously constructed classes of words (8). It is fundamental to take into account the contextual information in order to better distribute the words into the classes. Thus, the use of contextual information is of interest in the context of improving the predictability of the model. It makes it possible to offer a better distribution of words into classes and thus, a more balanced distribution of distributions (8). The second type, either semantic or syn-

tactic information, is formalized by categories or grammars. In the approach proposed by (8), the used information a priori is extracted from a learning corpus labeled in grammatical categories. The approach proposed by (29) is based on contextual information (left context and right context), so words that appear frequently in similar contexts should be assigned to the same class. According to (29), different vocabulary words are classified using the k-means algorithm. The particularity of this approach is based on the fact that the number of words in a class is set to k and if there is a class whose number of words is less than k then that class will be merged with another. The main advantage of this algorithm is its simplicity to find centroids and suddenly, the cost of merging words or classes becomes less expensive. The approach developed by (2) proposes to integrate semantic information for the formation of word classes in the statistical LM of ASR system. This approach is based on a pivot language (called IF for Interchange Format), which represents the meaning of the sentence regardless of the language (2). Thus, the criterion of choice of classes is guided by the definition of the pivot language and the most used concepts in the IF.

4 Tunisian dialect corpus collection

We create our own corpus of real spoken dialogues corresponding to the information request task in railway stations in collaboration with the Tunisian National Railway Company (SNCFT). This corpus is called TARIC, for Tunisian Arabic Railway Interaction Corpus [16]. The main task of the TARIC corpus is information request in the Tunisian dialect about the railway services in a railway station. These requests are about consultation, train type, train schedule, train destination, train path, ticket price and ticket booking. The creation of the corpus was done based on two steps: the production of audio recordings and the manual transcription of these recordings.This corpus consists of 21,102 statements and 66,082 words.

The Tunisian dialect the Tunisians' mother tongue, which is used in their daily oral communication. It is becoming more and more useful not only in interviews, talk shows and public services but also in blogs, chat rooms, SMS, e-mails, etc. However until now the Tunisian dialect has no standardized spelling. Indeed, this dialect differs from MSA and it does not have a standard spelling because there are no academies of Arabic dialect. Thus, to obtain coherent learning data and to have a robust and powerful language model, it is necessary to utilize a standard spelling. Indeed, there are words with many forms. For example, the word رزرفسيون /reservation/ can be written in four different ways: رَازَارفسيون, رَازَارفسيون and ريزرفسيون. As a result, each word has a unique form. Spelling transcription guidelines CODA (Tunisian Dialect writing convention), (30), were adopted.

The normalization step is essential because it presents a key point for the other steps of our method. Among the normalisation Tunisian Dialect words we have: *(i)*Number "sixteen" is written as ستطَاش. *(ii)*To define the future, we must follow the following form: بَاش + verb, for example: بَاش نمشي. *(iii)* To define the negation, we must follow the following form: مَا + verb.

The Tunisian Dialect is characterized by the presence of foreign words, such as for instance: French, English, Spanish, Italian, etc.To transcribe these words, Arabic alphabets have been used. These foreign words have a unique form, for example: رتور [Back], ترَان [train]... At the end of this step, we obtain a standardized corpus, the figure 2 represents a corpus extract before the normalization step.

Figure 2: Corpus extract before the normalisation step

The figure 3 represents a corpus extract after the normalization step.

Figure 3: Corpus extract after the normalisation step

5 N-class LM construction

In this section, we are going to describe our method to create n-class LM based on semantic information for establishing word classes. This method is made up of five distinct stages.

5.1 Pre-processing stage

When we studied our corpus, we noticed that there are words that have no semantic value when they are figured all alone. Only the grouping of these words with other words can give a better semantic value to words that may be insignificant and subsequently useless for our field of work. As a result, we have decided to create semantic blocks that consist of grouping one or more words into a single word. Semantic block is defined as a group of two or more words. Indeed, this pretreatment consists of adding a (-) between two or more words to build a single word. Among the words that can be grouped together to form a semantic block, we find مَاضِي followed by another word to indicate the time for example مَاضِي سَاعَة [1 PM]. Cities whose names are composed such as سيدي بوزيد [name of Tunisian city] مَا [negation] followed by a verb with a negative form to express negation. This step is necessary because it will be used for semantic labeling and later for the formation of word classes. Indeed, the main objective of this step is to give a better semantic value to words that may be insignificant and subsequently useless for our work. Following this step, we obtain a corpus that contains all the possible semantic blocks. Below (figure 4) is an excerpt from this corpus.

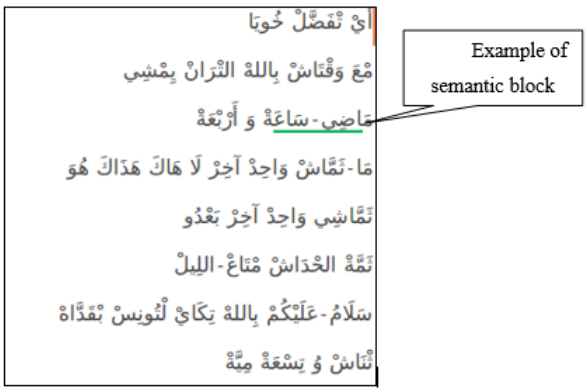

Figure 4: Extract of our corpus after semantic blocks formation

5.2 Statistical Semantic labeling

In this sub-section, we present in more details the way we integrate semantic order information for the formation of word classes. According to the previous steps, the semantic labeling step consists in giving a label to each single word or semantic block. We have performed the labeling

task of words or semantic blocks extracted from a sentence and their corresponding concepts in the field of railway request information. In this task, each word or semantic block is labeled with a concept indicating its appropriate semantic nature. Thus, semantic labeling is not done word by word because we can find words that can have several meanings depending on the context in which they are used. Subsequently, the labeling of a word or a semantic block is done while taking into account its left and right neighborhood in a sentence. The figure 5 shows examples of semantic labeling.

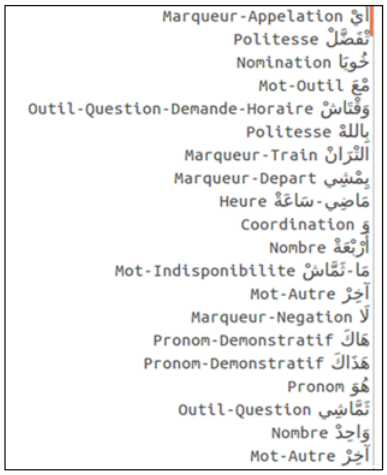

Figure 5: Extract of semantically labelled corpus

In order to improve the sequence labeling task, we proposed to use a discriminative algorithm based on the Conditional Random Field (CRF) model. Thanks to their ability in learning the sequence tagging tasks efficiently, CRF have been applied to a wide range of NLP applications, such as morpho-syntactic tagging (POS), chunking and language modeling. CRF are probabilistic models for computing the conditional probability of a possible output giving an input sequence also called the observation sequence. To train semantic labeling associations and some predefined feature sets, CRF learns a set of weights w. Learning the parameter set w is usually done by a maximum likelihood learning for $p(\bar{x}|\bar{y};w)$:

$$p(\bar{x}|\bar{y};w) = \frac{1}{z(\bar{x}|w)} exp \sum_j w_j F_j(\bar{x},\bar{y}) \quad (1)$$

According to these equations, $\bar{x}$ represents the sequence of words or semantic blocks (observation), $\bar{y}$ represents the sequence of labels, and w stands for the weights. F_j corresponds to a feature function. The latter depends on the sequence of words, the current label, the previous label and the current position in an utterance. We utilized the CRF++ toolkit (16) in our experiments. It is a customizable and open source, which implement the CRF for segmenting and labeling sequential data. It is written in C++, employs fast training based on gradient descent and generates n-best candidates. In order to measure the performance of the labelling task, three evaluation metrics were generally adopted. The latter allow expressing Recall and Precision. These measures could be calculated as follows:

$$Precision = \frac{\#word\ correctly\ Labelling}{\#word\ Labelling} \quad (2)$$

$$Recall = \frac{\#word\ correctly\ Labelling}{\#total\ of\ word} \quad (3)$$

Our purpose is to create training and testing sets decently. We outlined the available datasets for the languages under investigation. We randomly selected 15826 sentences and 49562 words for training, 5276 sentences and 1652 words for testing. The outcomes of the CRF system show a Recall of 88% and a Precision of 87%. Based on the manual examination of the automatic labeling result using CRF, we found that the CRF have the ability to detect composed token specific to the task and label them correctly.

5.3 Construction of semantic classes

The present work, being mainly dedicated to building n-class-based LM, focuses essentially on the formation of semantic

classes. Each class represents an abstraction of similar labels. In fact, a semantic class may correspond to a label or a group of labels, whereas a label cannot belong to only one class. Thus, if we consider the two statements: **Sentence 1:** تُونِس دُوزيَام كلَاس [Tunis second class]. **Sentence 2:** سوسَة پروميَار كلَاس [Sousa first class]. They become similar from this point of view because the words تُونِس [name of city] and سوسَة [name of city] belong to the same semantic class grouping "city" and the words دُوزيَام and پروميَار belong to the same semantic class grouping "Class". Indeed, the number of classes must be limited while the number of occurrences of words belonging to same class must be large enough to come up with correct learned probabilities. We therefore choose to limit ourselves to the selection of classes most frequently encountered in our corpus, corresponding to our Field of study "Tunisian Railway Transport Network". In this step, we identify the most common label and group them into classes. A class then corresponds to a set of words leading to the same semantic labelling representation.

Table 1: Examples of Semantic Classes

Semantic classes	Variants semantic tags
City	Destination-Station..
Schedule	Trip-time, Arrival-hour ..
Response	Confirmation, Negation..

In our case, this classification provides about 20 semantic classes such as [City], [Response], [Request-Concept], [Comparison], [Schedule]. After obtaining the list of semantic classes, as shown in Table 1, we can then directly associate each word of our corpus with the class to which it belongs.

5.4 N-class calculation

We have already done the first experience for n-class LM calculation for Tunisian Dialect. After obtaining the list of semantic classes, we use it in combination with the data of LM to build our new "semantic" model. In the LM learning corpus all words (or semantic blocks) are replaced by class names. The result is a "prepared" corpus that contains both words (or semantic blocks) and Semantic classes. Finally, we use the SRILM [1] toolbox to learn LM including semantic classes. SRILM is a toolkit for building and applying statistical LM, primarily for use in speech recognition, statistical tagging and segmentation, and machine translation.

5.5 Evaluation of a LM

Several measures are used to evaluate the quality of LM. We present perplexity as the most used method. Perplexity (PPL) is a quick method to evaluate LM. It is commonly used for several years to judge the quality of LM (15). This evaluation metric is used to measure the prediction ability of LM on a test corpus not seen during learning. The principle of perplexity is to check how much LM can predict the word sequences of the language it is supposed to model. Perplexity is defined by:

$$PPL = 2^{-\frac{1}{n}} \sum_{t=1}^{n} logP(w_t|h) \quad (4)$$

Where $P(w_t|h)$ represents the probability proposed by the LM for the word knowing the history h. Indeed, the perplexity of LM is between 1 and V, V is the size of vocabulary, that is to say the number of words that compose it. A reduced value of perplexity leads to better LM prediction capability.

As we have already mentioned, a low value of perplexity reflects a strong predictive power of LM. Thus, to better judge our choice of creating an n-class word model, we compared the created model with the 3-gram type model on the same test corpus of evaluation. Additionally, to assess the impact of using the CRF model to perform the semantic labelling task in order to construct semantic classes,

[1] http://www.speech.sri.com/projects/srilm/

we compared the n-class created model with using the CRF in the semantic labelling task and the n-class model without using the CRF in the semantic labelling task. The table below shows a comparison between the n-class model based on the CRF model in the Semantic labeling task and the n-class model without applying the CRF model in the Semantic labeling task together with the n-gram model in terms of perplexity.

Table 2: Value of perplexity calculated on the same test corpus

Type of model	**Perplexity**
3-gram	74.46
n-class without CRF	4.17
n-class with CRF	3.87

Table 2 shows the very significant relative reduction in perplexity by applying the CRF model in the semantic labelling compared to the other models. These results are consistent with what could be expected: (1) it is the classification based on semantic data that has minimized the perplexity of the obtained LM. The value of the n-class model perplexity remains well below that of the 3-gram model on the test corpus. Interestingly, the same models as for the learning corpus have the lowest perplexity value on the test corpus. (2) the application of CRF model to perform the semantic labelling task affects the improvement in creating semantic classes, by taking into account the n-class LM, and calculating the perplexity rates.

6 Conclusion

In this work, we have interested in constructing a statistical LM as one of the components of ASR system. The main role of this model is to give the probability of distribution on sets of word sequences. In particular, we are interested in n-class LM by using semantic information for the creation of word classes. Our choice is justified by the fact that some words are similar but they do not have the same probability of appearance, so their class groupings will be possible because of the limitation of our field of work ”the Tunisian Railway Transport Network”. The main idea of this model is to group vocabulary words into semantic classes and to consider mainly the calculation of the probability of a sequence of words such as the probability of a sequence of these semantic classes. To do this, we have followed these steps:We firstly construct semantic blocks that consists in grouping one or more words into a single word that we call ”semantic blocks”. Secondly, we do the semantic tagging. So in order to obtain a labeled corpus, the semantic labeling step consists in giving a label for each single word or for each semantic block. To improve the sequence labeling task, we proposed to use a discriminative algorithm based on the Conditional Random Field (CRF) model. Thirdly, we form semantic classes. In fact, a semantic class may correspond to a label or a group of labels, whereas a label cannot belong to only one class.LM calculation based on the SRILM tool. Evaluating the constructed model by calculating its perplexity. In order to test the model generated by our statistical LM system, we compared the created model with the 3-gram type model on the same test corpus. Secondly, to better judge the impact of using the CRF model to perform the semantic labelling task in order to construct semantic classes, we compared the n-class created model based on the CRF in the semantic labelling task and the n-class model without using the CRF in the semantic labelling task. According to this evaluation, we can say that the classification based on semantic data has minimized the perplexity of the LM obtained as compared to the rapport 3-gram LM. Moreover, the use of the CRF model to perform the semantic labelling task has an impact on the improvement in creating semantic classes, by taking into account the n-class LM and calculating the perplexity rates.

References

[1] Bahl L.R., Brown P.F., Souza P.V., Mercer R.L.: A tree-based Statistical Language Model for Natural Language Speech Recognition, IEEE Transactions on Acoustics, Speech and Signal Processing (1989).

[2] Bigi B : Modle de langage smantique pour la reconnaissance automatique de parole dans un contexte de traduction, Laboratoire Parole et Langage - Aix-en-Provence (2012).

[3] Bilmes J. A., Kirchhoff K.: Factored Language Models and Generalized Parallel Backo, Proc. of Human Language Technologies, North American (2003).

[4] Bouagres F. : Attelage de systmes de transcription automatique de la parole, thse de doctorat, universit du Maine-Le Mans-France (2012).

[5] Brown P. F., DellaPietra V. J., Souza P. V., Lai J. C., Mercer R. L.: Class-Based N-Gram Models of Natural Language, Computational Linguistics (1992).

[6] Chelba C., Engle D., Jelinek F., Jimenez V., Khudanpur S., Mangu L., Printz H., Ristad E., Rosenfeld R., Stolcke A., Wu D.: Structure and Performance of a Dependency Language Model, Dans Proc. of the European Conf. on Speech Communication and Technology (Eu-rospeech) (1997).

[7] Chelba C., Jelinek F.: Structured Language Modeling , Computer Speech and Language (2000).

[8] Damnati G. : Modles de langage et classification automatique pour la reconnaissance de la parole continue dans un contexte de dialogue oral homme-machine, thse de doctorat, uni-versit dAvignon et des pays du vaucluse (2000).

[9] Graja M., Jaoua M. and Belguith L.: Discriminative Framework for Spoken Tunisian Dialect Understanding. SLSP(2013)

[10] Farhat A., Isabelle J.F., O'Shaughnessy D.: Clustering Words for Statistical Language Mod-els based on Contextual Word Similarity, Dans Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Atlanta, USA (1996).

[11] Jelinek F.: Continuous Speech Recognition by Statistical Methods, Proc. Of the IEEE (1976).

[12] Ji G., Bilmes J.: Dialog Act Tagging Using Graphical Models, Proc. Of the IEEE Intl Conf. on Acoustics, Speech, and Signal ProcessingICASSP (2005).

[13] Habash, N., Diab, M. and Rambow, O.: Conventional Orthography for dialectal Ara-bic,Proceedings of the Eighth International Conference on Language Resources and Evalua-tion, LREC-2012, (2012).

[14] Kirchhoff K., Yang M.: Improved Language Modeling for Statistical Machine Translation, Proc. of the ACL Workshop on Building and Using Parallel Texts (ParaTex), Morristown, NJ, USA. Association for Computational Linguistics (2005).

[15] Kneser R., Ney H.: Improved Clustering Techniques for Class-Based Statistical Language Modeling. Proc. European Conference on Speech Communication and Technology, Berlin, Allemagne (1993).

[16] Kudo, T.: crf++, http://chasen.org/ taku/software/CRF++/

[17] Lecorv G. : Adaptation thmatique non supervise d'un systme de reconnaissance automatique de la parole, thse de doctorat, Universit europenne de Bretagne (2010).

[18] Masmoudi A., Ellouze M., Estve Y., Hadrich Belguith L., Habash N.: A Corpus and Pho-netic Dictionary for Tunisian Arabic Speech Recognition, LREC'2014, Reykjavik, Iceland (2014).

[19] Masmoudi A Ellouze M., Estve Y., Bougares F HadrichBelguith L: Phonetic tool for the Tunisian Arabic, SLTU'2014, Russia, (2014).

[20] Masmoudi, A., Habash, N., Khmekhem, M., Esteve, Y. and Belguith, L.: Arabic Transliteration of Romanized Tunisian Dialect Text: A Preliminary Investigation, Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, p.608-619, (2015).

[21] Masmoudi, A., Bougares,F., Ellouze, M., Estve, Y., Belguith, L.: Automatic speech recognition system for Tunisian dialect. Language Resources and Evaluation 52(1): 249-267 (2018).

[22] Abir Masmoudi, Rim Laatar, Mariem Ellouze, Lamia Hadrich Belguith: N-Class Language Model for Tunisian Dialect Automatic Speech Recognition System. LPKM (2018)

[23] Rosenfeld R.: Adaptive Statistical Language Modeling: A Maximum Entropy Approach, PhD Thesis, (1994).

[24] Smali K., Conception et ralisation dune machine dicter entre vocale destine aux grands vocabulaires : Le systme MAUD, Thse de doctorat, Universit de Nancy 1, 1991.

[25] Smaili K., Jamoussi S., Langlois D., Haton J.-P.: Statistical feature language model, Proc. of the 8th Intl Conf. on Spoken Language ProcessingICSLP, (2004).

[26] Vergyri D., Kirchhoff K., Duh K., Stolcke A.: Morphology-Based Language Modeling for Arabic Speech Recognition, Proc. of the 8th Intl Conf. on Spoken Language Processing (ICSLP), Jeju Island, South Korea, (2004).

[27] Wang, Y., Acero, A., Mahajan, M., Lee, J.: Combining statistical and knowledge-based spoken language understanding in conditional models. In: Proceeding COLING/ACL, Sydney, Australia (2006)

[28] Xu P., Chelba C., Jelinek F.: A study on richer syntactic dependencies for structured lan-guage modeling, Proc. of the 40th Annual Meeting on Association for Computational Linguistics (ACL), Morristown, NJ, USA. Association for Computational Linguistics, (2002).

[29] Zitouni I.: Linearly Interpolated Hierarchical N-gram Language Models for Speech Recogni-tion Engines, IBM T.J. Watson Research Center, NY, Bell Labs Alcatel-Lucent, NJ,USA, (2008).

[30] Zribi I., Boujelben R., Masmoudi A., Ellouze M., Belguith L., Habash N.: A conventionnal Orthography for Tunisian Arabic, LREC'2014, Reykjavik, Iceland, (2014).