ISSN 1870-4069

Improvement a Transcription Generated by an Automatic Speech Recognition System for Arabic Using a Collocation Extraction Approach

Heithem Amich, Mounir Zrigui

LATICE Laboratory, Research Department of Computer Science,
University of Monastir, Tunisia
heithem07@gmail.com, mounir.zrigui@fsm.rnu.tn

Abstract. The following study propose a novel heuristic to improve an automatic speech recognition system for Arabic language. Our heuristic relies on the collaboration of two approach: the first one ensures the extraction of collocations from a voluminous corpus then stores them in a database. It uses a combination of several classical measures to cover all aspects of a given corpus in order to exclude bigrams having a high probability of occurring together. The second one constructs a search space on the relations of semantic dependence of the output of a recognition system then, it applies phonetic filter so as to select the most probable hypothesis. To achieve this objective, different techniques are deployed, such as the word2vec or the language model RNNLM in addition to a phonetic pruning system. The obtained results showed that the proposed approach allowed improving the precision of the system.

Keywords: automatic speech recognition, multi-level improvement, collocation, semantic similarity, phonetic pruning.

1 Introduction

Automatic speech recognition has been growing interest in recent years. It aims to facilitate communication between people and system and allows to moving from an acoustic signal of speech to the transcription of the signal in a written version. Indeed, how does a transcription system work? From a recording, the system starts by calculating a transformation of the signal in acoustic parameters adapted to a recognition engine [1]. This latter makes use of acoustic and linguistic knowledge to produce the transcription [2]. The performances of the transcription systems are good when two critical elements are well mastered, the quality of the sound recording and the availability of recordings representative of the context of use. Although an ideal transcription system remains always nonexistent, several research efforts have recently been made to come up with robust systems [3]. Automatic speech processing still has a few defect. In fact, the main limitations that hinder the development of efficient systems are generally linked to the great deal of variability in speech. On this respect, we remind of the intra-speaker variability [4], due to the elocution (singing voice, shouting, whispering, hoarse, husky, under stress), inter speaker variability (male voice, female voice, or child voice) as well as the variability caused by the signal acquisition device (type of micro-

phone), or by the environment (noise, cross talk) [5]. Moreover, the degradation of performance is generally due to the lack of precise rules to formalize knowledge to different decoding levels (including, syntax, semantics, and pragmatics). On statistical methods with learning techniques from oral corpora where the correct transcription is known in advance. A statistical ASR is made up of several components following the acoustic and linguistic modeling of speech signal with a view to its recognition.

Many Techniques have been developed to improve each component of the system so as take account of or reduce the problems related to speech variability. Never the less, each technique has certain weaknesses. This leads us to develop an approach which takes account neither of the recognition modules adopted by an ASR, nor its search algorithms, or its smoothing techniques, which is the strong point of this approach. As a matter of fact, we considered the ASR as a black box device of any power of decision. Its role is limited to providing the transcription that will trigger our correction process. Finally, our approach in the only one responsible for correcting mis-recognized hypotheses and irrelevant word [6,7]. Also, if possible, it try to predict the next word that speaker probably will uttered. After a brief state of the art on the technique of improving transcriptions, we describe our first approach in section 3 and the precision improvement approach in section 4, we evoke the global steps of our idea. In section 5, we integrate the concept of colocation into our system. Finally, we discuss different evaluation results. In the last section, we discuss different evaluation results in section 6.

2 State of the Art

Improving the performance of ASR caught the attention of specialists in many languages. Many works were carried out to improve the competency of the various components of the system such as the linguistic and acoustic models and to significantly improve the decoding quality and the transcription quality a priori. In this framework, Lecouteux [8] presents a combinational method allowing to exploit a priori manual transcriptions and to integrate then directly into the heart of a SARP. This method allows to effectively guiding the recognition system with the help of auxiliary information. He also combined SRALs based on guided decoding [9]. With reference to previous research works, Benoıt Favre [10] proposed a fusion system between an original sentence containing an error and sentence of clarification. Thus, he proposed many alignments of levenshtein variants [11] and a reranker to select the best hypothesis. Antoine Laurent [12] came up with a method allowing to help the user in the step of correcting ASR outputs and to correctly transcribe proper names to facilitate the automatic indexing of transcribed reunions.

Fathi Bongares [13] studied the methods of combining transcription systems of large vocabulary speech. His study focuses a on the coupling of heterogeneous transcription systems with the aim of improving the transcription quality. Combining different transcription systems is based on the idea of exploiting the strengths of each system in order to obtain a final improved transcription. In order to overcome the essential problem of natural language processing that resides in the manipulation of large volumes of texts long Med Achraf presents a collocation extraction approach based on clustering technique. He used a combination of several classical measures which cover all aspects of

a given corpus in order to drew out the consecutive pairs (w_i,w_{i+1}) of a word commonly used from a voluminous corpus. Likewise, Christopher manning exposes a number of approaches to capturing collocations such as selection of collocation by frequency or the method based on the mean and variance of the distance in more than the t-test method and mutual information.

3 The Proposed Approach

In this section, we will present our system in details.

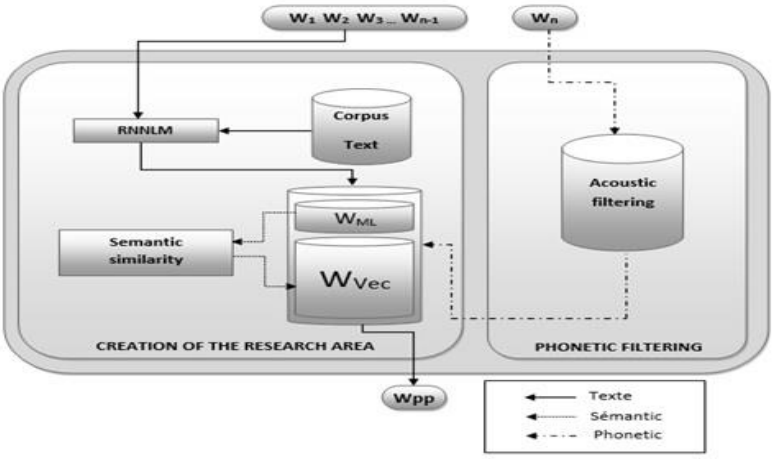

Fig. 1. The Verification and Correction System of Transcription (SyMAT).

The process of automatic correction of mis- spelt words from Arabic will be done in two main phases, as shown in figure 1. The steps of the left block scheme represent the first phase. It is particularly appropriate for extending the search space for the word to correct. The second stage is it at the right scheme. This phase is responsible for selecting the most likely word scheme.

3.1 Creation of Search Space

We expose to you the following case: the ASR has succeeded to transcribe the following word: w_0,w_1,…,w_{n-1} . By using our approach, we want to find the next word w_n badly recognized by the system ASR. The first step is to build a research space that may contain the word which we are seeking. This part is essential to develop the search space that will contain the words generated by the RNNLM language model and the semantic similarity.

Rnnlm. Let S=w_0,w_1,…,w_{n-1} be the context at a given instance our approach aims to estimate all of the most likely hypotheses w_n by using an RNNLM language model. This preliminary phase consists of passing the set of observations S to a language model in order to retrieve the set of the most likely words which could complete S. The

RNNLM model is based on the association of neural networks at word level. In what follows, we briefly remind of the mathematical strategies relevant to the model. Recently, deep neural networks have made a great success in the fields of image processing, acoustic modelling [13], language modelling [14,15], etc. Language models based on neural networks do better than standard back off n-gram models. Words are projected into low dimensional space similar words are grouped together. RNNLM could be a deep neural network LM due to its recurrent connection between input layer and hidden layer [16]. The network has an input layer x, a hidden layer S and an output layer y. We denote input to the network in time t as x (t) and output as $y(t)$. $S(t)$ refers to the state of the network (hidden layer). In put vector (x) is formed by concatenating vector $w(t)$ which represents current word. Output is made from neurons in context layer S at time $(t-1)$ [17]. The architecture of the neural network used to calculate conditional probabilities is organized in three layers [18]. The input layer reads a word $w(t-1)$ and a continuous $S(t-1)$. The hidden layer compresses the information of these two inputs and calculates a new representation S (t) for the input of the next propagation. The value is then passed on to the output layer, which provides the conditional probabilities P (w (t) $\mid w(t-1), S(t-1)$). RNNLM can be expressed as follows:

$$x(t) = w(t-1) + S(t-1), \tag{1}$$

$$S_j\ (t) = f\left(\sum_i\ U_i(t)U_{ij}\right), \tag{2}$$

$$y_k = g\left(\sum_i\ S_j\ (t)k_j\right), \tag{3}$$

where f (z) is a function of sigmoid activation:

$$f(z) = \frac{1}{1+e^{-z}}, \tag{4}$$

and g (z) is a softmax function:

$$g(z_m) = \frac{e^{zm}}{\sum K\ e^{zm}}. \tag{5}$$

Semantic Similarity. Identifying the similarity between words is an important TAL task regarding the domains where this technique could be useful, such as the search for information, automatic translation or even the automatic generation of text. The ability to correctly identify the semantic similarity between words is essential for our system. This is because of its contribution to the reconstruction of research space. The search for similarity is based on the word2vec techniques [19]. Word2vec is a neural network with two layers having as an input a text corpus and as an output a set of vectors representing the characteristics of the input word in this corpus. Word is then taken to measuring the cosines similarity where an angle of 0 degree expresses a total similarity, whereas an angle of 90 degrees expresses no similarity. The following table present a list of words associated with the word «July» rising word2vec, in order of proximity.

Table 1. A list of Words Associated with the Word "July= جويلية " using Word2vec

ASR	Cosine values
June (جوان)	0.9557317
April (افريل)	0.9386088
August (اوت)	0.9324805
March (مارس)	0.9314448
May (ماي)	0.9097166

Word2vec assigns a value equal to 0.6230781 to the word «France», so we deduce that France does not admit any semantic dependence with the word «July». The next step is to apply the text corpus learning and display the figure that shows the location of the words in a two dimensional space by a projection of the main component PCA, we notice that words with the same semantic meaning are adjacent. The figure below illustrates the locations of a set of words having the same semantic context.

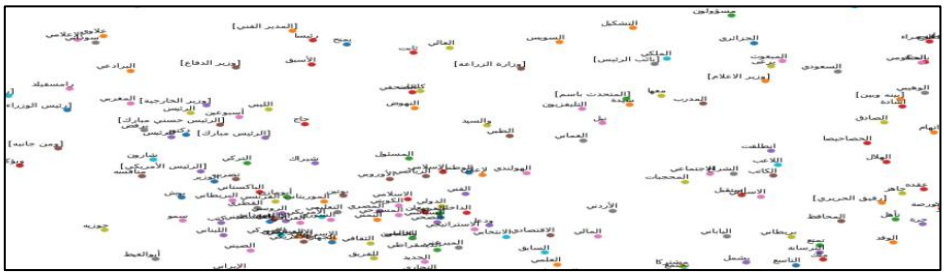

Fig. 2. The Distribution of Words According to the Cosine Value using PCA.

3.2 Selection of the Most Probable Word

Having collected a well-defined number of lexicons constituting the search space, we highlighted the techniques allowing filtering, classifying and finding the most appropriate hypothesis. We adopted two filtering methods: the syntactic filtering and the phonetic, filtering.

Phonetic Comparison. Having obtained a set of word W_{vec} +W_{ML}, we introduced another filtering mechanism operating at a phonetic level. This tool compares the frequency spectrum of the word W_n coming from an ASR and the frequency spectra of the word W_{vec} +W_{ML}. This method consists in aligning the signals of two words, then measuring the degree of similarity of two spectra. At the end of this phase, we estimate the word W_n having the most likely label and the highest degree of acoustic similarity. This example shows how to measure the similarities of signal. Whether they are correlated or not? The black and blue signals show the signals of two most likely words generated by search space. The third signal corresponds to the word signal generated by ASR. This figure shows that there is no phonetic similarity between the two candidates with the third signal. Just by looking at the time series, the signal seems not to

correspond to one of both models. A closer look reveals that the signals did different lengths and sample rates.

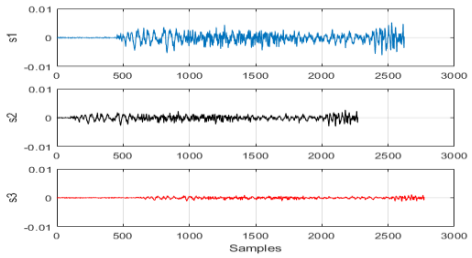

Fig3. Comparing the Similarity of Two Signals.

The Case of the First Word of the Sentence. Concerning the previous steps of our approach, we recalled the different phases of the automatic correction of transcriptions provided by an automatic speech recognition system. We elaborated architecture capable of sending back the next most likely hypothesis w_n after taking the n-1 hypotheses produced by an ASR as input: its worth mentioning that it is evident to find the words having indices between 2 and n given that there is data to manipulate. However, at the start of our procedure, we had w_0 data to activate our approach, so as to find the first word of the sentence. To overcome this limitation, we have partially changed our strategy. Indeed, we temporarily accepted the two most likely words generated by an ASR w_{11} and w_{12}. We remind that a speech recognition system uses these three pillars lexicon, the language model and the acoustic model to provide a text representing the transcription of a sound signal (the best one). It is also possible to retain several recognition hypotheses. The output world, then, be a list of best hypotheses N, a word graph or a confusion network. We limited ourselves to extracting the two most likely words among the retained N best hypotheses of an ASR of the first word of a sentence. This is simple due to the lack of data, which obliges us to accept w_{11} and w_{12}. However, the choice is not final. We have designed the method that reviews and verifies the first word of the sentence. The final result can accept w_{11} or rather w_{12} as well as a new lexicon retained by our approach based on a set of probabilities.

4 The Global Steps

In this section, we will present a detailed representation of our automatic correction system of the transcript provided from a speech recognition system. This procedure is carried out in 4 steps:

- The first step consists in extracting the two best hypotheses of first word of the sentence 1 from an ASR.
- Having acquired the two hypotheses W_{11} andW_{12}, we accept W_{11}. Then, we pass W_{21} to our search approach.

- It is essential to indicate the origin of the word. That is to say, if it is the result of the language model $W_{2\mathrm{M1}}$ or rather the result of word2vec $W_{2\mathrm{vec1}}$.

- Of the word comes from the language model, we pass W11 and $W_{2\mathrm{ML1}}$ to our approach in order to determine $W_{3\mathrm{ML1}}$ or $W_{3\mathrm{vec1}}$. Otherwise, shift back to by using an inverse language model choose either W_{11} or W_{12} or even another word proposed by the language model. This back shift is done only when the word, retrieved by our approach, comes from the tool word2vec. Needless to remind that we could also define a sort of in versed language model whose words were generated in a reverse order (from right to left):

$$P_{\mathrm{reversed}}\ \overrightarrow{(w)} \stackrel{\mathrm{def}}{=} \mathrm{P}(w_n)\ \mathrm{P}(w_{n-1}|w_n).\ \mathrm{P}(w_{n-2}|w_{n-1}w_n).$$
$$\mathrm{P}(w_{n-3}|w_{n-2}\ \mathrm{P}(w_{n-2}|w_{n-1}w_n).w_n).\ \ldots.\ \mathrm{P}(w_2|w_3w_4)\ \mathrm{P}(w_1|w_2w_3)$$

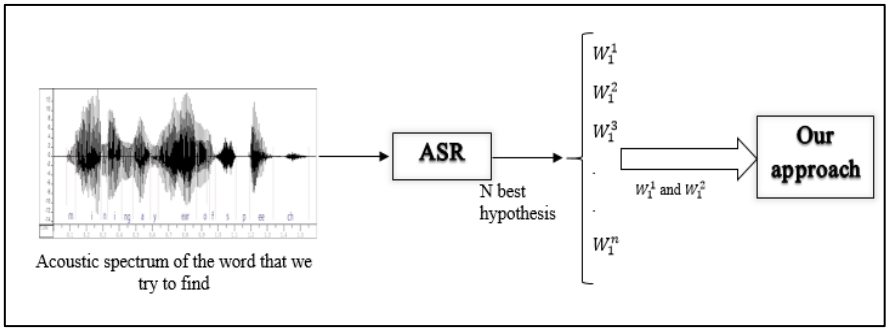

Fig. 4. First Phase of Our Approach.

Following each word generated by an ASR, it is susceptible to change the old word found by our approach during a back shift. The final choice is decided when we process the last word of the sentence, which can influence or substitute the previously executed hypotheses.

5 Improvement Precision

In order to increase the robustness and performance of our main system shown in Figure 1 and reduce its response time. We have added a new compartment called collocation. In this section, we will present in detail the process of extraction of collocations in the system as well as the integration steps of two approaches. Collocations refer to the most widespread pair of lexemes (l_i ,l_{i+1}) commonly used in the spontaneous Arabic language. They are necessarily consecutive whose existence of a lexeme li at position X_i in a corpus T certainly requires the presence of the lexeme l_{i+1} at the position X_{i+1}. A collocation is expression of two words that corresponding to some conventional method of saying things. There is considerable overlap between the concept of collocation and notion like term, technical term and terminological. Collocation are crucial for several

domain: natural language generation, computational lexicography and corpus linguistic research. It comprise:

- Proper names : الولايات المتحدة (United State)
- Verbal expression : (I saw the light) أبصر النور
- Terminologies : (Hello) السلام عليكم

5.1 Conventional Approaches for Extraction of Collocations

The t Test. If two words occur together many times, then we expect the two words to co-occur a lot just by chance. The t- test has been widely used for collocation discovery. It looks at the difference between the observant and expected means. If t is large enough the w1 and w2 are associated, we compute the t static:

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{N}}}, \tag{6}$$

where $\bar{X}$ is the sample mean, N is the sample size, μ is the mean distribution and S^2 is the sample variance [20].

Likelihood Ratio. is further method for hypothesis testing. In applying this test to collocation discovery, we have the ability to distinguish the occurrence of both common and rare phenomenon [20]. This method gives two hypotheses and test, which one is most probably the two hypotheses $H_1$and H_2 are:

- H_1: independence between w_1 and w_2: $p(w_2|w_1) = p(w_2|\neg w_1) = p$,
- H_2: dependence between w_1 and w_2: $p(w_2|w_1) = p_1 \neq p_2 = p(w_2|\neg w_1)$.

The likelihood ratio is:

$$\lambda = \frac{L(H_1)}{L(H_2)}, \tag{7}$$

where L is the likelihood function, assuming a binominal distribution L is given by:

$$L(p; n, r) = r^p (1 - r)^{n-p}, \tag{8}$$

where n is the number of trials, r the number of successes, and p is the probability of success.

Mutual Information. is a measure of how much one tell us about the other. It allows to compare the probability of observing w_1 and w_2 independently $p(w_1)\ p(w_2)$ mutual information is calculated by:

$$I(w_1|w_2) = \log_2 \frac{p(w_1|w_2)}{p(w_1)\ p(w_2)}. \tag{9}$$

If mutual information is large then w_1 and w_2 are related else, it is too low then w_1 and w_2 are independent [20].

5.2 The Steps of Extraction of Collocations

To extract all the most common collocations of the arable language, we have combined the three methods recently mentioned, called the t-test, the Likelihood ratio and the mutual information. Thus for each candidate of the collocation w_1 w_2, these thee measures will be used to calculate the dependency between w_1 and w_2. Then, we calculate the average value of the three measures for each bigram. We consider a collocation all bigrams corpus having a mean higher than a very high empirical threshold. The preliminary step consists to segmenting the corpus by identifying the basic units forming the corpus. This means identifying the separators used to isolate the morphemes. We also define a stop list to omit the words that cannot form a collocation as:

- The particles of coordination: (ثم، أم، أو، أما، إما).
- The interrogative particles : (أيّ, كيف, أين, متى).
- The particles of Appeal: (يا، أيا ،أيها هيا).

Once the bigrams have been identified, the next step is the calculation, for each bigram, we calculate the average of three measures mentioned previously. If the value found is greater than a threshold, then the bigrams is considered collocation and we add it to the list of collocations.

Notations used are summarized in the following:

- T: Corpus size.
- L_i : lexeme i, 1≤ i ≤ T .
- B_i : Bigram i.
- SL : Stop List.
- E_i : a real which designates the calculated average of each bigrams

```
1. //Bigrams extraction and Measures computation
2. For all lexemes l_i, 1 ≤ i ≤ T − 1 Do
3. B_j = { l_i, l_{i+1} / l_i ∉ SL ∧ l_{i+1} ∉ SL } End.
4. //calculate the average of each bigram
5. Ei = average( Mutual inf (Bi), t test (Bi), Likehood.ratio(Bi)
6. // add the bigram to all the collocations If (E_i > thrushold) {
7. Col = Col ∪ { w_i, w_{i+1} } }. End
```

Figure 6 illustrates some of the accumulated collocations in our database collocation [20].

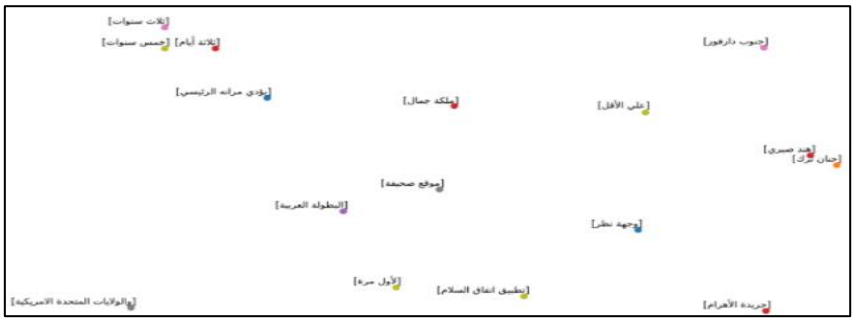

Fig. 6. Colocation Group in a Two-Dimensional Space.

5.3 The Integration of Collocations into the Main System

At this level, we have completed the construction of a collocation base. However, the obvious question is about the contribution of integrating the concept of collocations into our system?

Let S=w_0,w_1,…,w_{n-1} be the context at a given instance. S represents the words pronounced by the speaker. At this point, our system has completed the verification of the whole sequence in order word after word with success. Let w_n be the word that will be treated. If the word w_{n-1} does not belong to the stop list, our heuristic checks if the previous word w_{n-1} is part of one of the collocations previously collected. We recall that a colocation is composed of two lexemes (l_i,l_{i+1}). If w_{n-1} exists in the collocation base (w_{n-1}=l_i), then it is sufficient to apply the acoustic comparison of w_n with the second lexeme l_{i+1}.

That means that the steps for creating the search space provided by the RNNLM language model and word2vec be canceled, the general heuristic has two paths, if the last word processed by the system is part of the collocations, then we just perform the acoustic test. If this test is positive then this is the word to look for. If not we execute as usual our initial approach (SyMAT). The integration of the colocation approach into the SyMAT system is very beneficial to the level of confidentiality and accuracy of the final result.

Indeed, if the word belongs to the list of collocations stored, and the acoustic test established is positive. Doubtless, we are confident that this is the exact word uttered by the speaker.

6 Experimentation

To construct the language model, we have used an Arabic text corpus of 100M words collected from corpus available on the used. This same corpus served to the construction of the model based on label. As for the testing of our system, we recorded a caustic corpus of 40 hours. We set up our SyMAT system at the exit of two known SPAP namely Sphinx [21].

Table 2. Results of the system.

ASR	Precision	F-mesure
Sphinx	51,38	56,41
Sphinx + SyMAT	56,52	62,05
HTK	46,24	50,77
HTK + SyMAT	52,72	57,88

The obtained results show that the proposed approach effectively contributed to improving ASR. We may also note that our method is more efficient for the HTK system than for the Sphinx system. This is justified by:

- The high clean error rate of the HTK system as compared to the sphinx system [21].
- The acoustic models trained by Sphinx were much better than that of HTK.

Table 3. Samples of Collocation Candidates.

$w_1 w_2$	M.I	T.T.	L.R.	E_i
مليون دولار (Million dollars)	0.9999	0.9999	0.9999	0.9999
أشراط الساعة (Signs of the Hour)	0.9987	0.9750	0.8774	0.9503
الطبعة الاولى (First Edition)	0.6487	0.2548	0.3458	0.4164
اتفاق السلام (Peace Agreement)	0.7814	0.7895	0.8569	0.809

The obtained results show that if the sum of the three values exceeds a threshold equal to 0.8, then the bigram is considered collocation.

7 Conclusion

On this paper, we propose heuristics with the aim of improving the transcription generated by an ASR for Arabic. This method exploits semantic, phonetic levels and collocation's concept in order to evaluate the output of the ASR system and to propose the most likely word in case there is an error. To enforce this approach, we resorted to the techniques of word similarity, t test, mutual information, likelihood ration and to the RNNLM language model to establish a search space based on the history of a transcription $W_1...W_{n-1}$. After that, we carried out a phonetic pruning to choose the most probable word. We also resorted to the techniques of t test, mutual information and likelihood ration to extract collocations in order to increase the exactitude of final result. As a future work, we hope to promote our system from a model allowing taking account of the historic of applied corrections and assuring adaptation of the correction process to a particular user.

References

1. Dua, M., Aggarwal, R.K., Virender, K., Dua, S.: Punjabi automatic speech recognition using htk (2012)
2. Aggarwal, R.K., Dave, M.: Acoustic modeling problem for automatic speech recognition system: advances and refinements (part II). I. J. Speech Technology 14(4):309 (2011)
3. Zolnay, A., Schlter, R., Ney, H.: Robust speech recognition using a voiced-unvoiced feature. In: IN PROC
4. Siegler, M.A., Stern, R.M.: On the effects of speech rate in large vocabulary speech recognition systems. In: 1995 International Conference on Acoustics, Speech and Signal Processing, ICASSP '95. Detroit, Michigan, USA, May 08-12, 1995, pp. 612–615 (1995)
5. Zhao, S.Y., Ravuri, S.V., Morgan, N.: Multi-stream to many-stream: using spectro-temporal features for ASR. In: INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association. Brighton, United Kingdom, September 6-10, 2009, pp. 2951–2954 (2009)
6. Rohit, P, Rohit, K, Sankaranarayanan, A, Chen, W., Hewavitharana, S., Roy, M.E., Choi, F., Challenner, A., Kan, E., Neelakantan, A., Natarajan, P.: Active error detection and resolution for speech-to-speech translation. In: 2012 International Workshop on Spoken Language Translation, IWSLT 2012. Hong Kong, December 6-7, 2012, pp. 150–157 (2012)
7. Alex, M., Tom, K., Mari, O., Luke, S.: Using syntactic and confusion network structure for out of vocabulary word detection. In: 2012 IEEE Spoken Language Technology Workshop (SLT). Miami, FL, USA, December 2-5, 2012, pp. 159–164 (2012)
8. Lecouteux, B., Linares, G., Oger, S.: Integrating imperfect transcripts into speech recognition systems for building highquality corpora. Computer Speech & Language 26(2):67–89 (2012)
9. Lecouteux, B., Nocera, P., Linares, G.: Semantic cache model driven speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010. Sheraton Dallas Hotel, Dallas, Texas, USA, pp. 4386–4389 (2010)
10. Favre, B., Rouvier, M., Béchet, F.: Reranked aligners for interactive transcript correction. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014. Florence, Italy, May 4-9, 2014, pp. 146–150 (2014)
11. Tor, H, Torleiv, K, Levenshtein, V.I.: Error correction capability of binary linear codes. IEEE Trans. Information Theory 51(4):1408–1423 (2005)
12. Antoine, L., Sylvain, M., Téva, M., Paul, D.: Computer-assisted transcription of speech based on confusion network reordering. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011. Prague Congress Center, Prague, Czech Republic, pp. 4884–4887 (2011)
13. Bougares, F., Esteve, Y., Deléglise, P., Linares, G.: Bag of n-gram driven decoding for LVCSR system harnessing. In: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011. Waikoloa, HI, USA, December 11-15, 2011, pp. 278–282 (2011)
14. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech & Language Processing 20(1):30–42 (2012)
15. Ebru, A., Tara, N., Brian, K., Bhuvana, R.: Deep neural network language models. In: Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, WLM '12, pp. 20–28. Stroudsburg, PA, USA, 2012. Association for Computational Linguistics (2012)
16. Tomas, M., Martin, K., Lukas, B., Jan, C., Sanjeev, K.: Recurrent neural network based language model. In: INTERSPEECH 2010, 11th Annual Conference of the International

Speech Communication Association. Makuhari, Chiba, Japan, September 26-30, 2010, pp. 1045–1048 (2010)
17. Jeffrey, P., Richard, S., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014. Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1532–1543 (2014)
18. Manning, C., Hinrich, S.: Foundations of statistical natural language processing. Cambridge, Mass. MIT Press (1999)
19. Frédéric, B., Benoit, F.: ASR error segment localization for spoken recovery strategy. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013. Vancouver, BC, Canada, May 26-31, 2013, pp. 6837–6841 (2013)
20. Ben Mohamed, M.A., Mallat, S., Nahdi, M.A., Zrigui, M.: Exploring the potential of schemes in building NLP tools for arabic language. Int. Arab J. Inf. Technol. 12(6):566–573 (2015)
21. Satori, H., Hiyassat, H., Harti, M., Chenfour, N.: Investigation Arabic Speech Recognition Using CMU Sphinx System. The International Arab Journal of Information Technology, Vol. 6, No. 2, April (2009)