INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Conditional Random Fields for the Tunisian Dialect Grapheme-to-Phoneme Conversion

Abir Masmoudi[1 2]*, Mariem Ellouze*[2]*, Fethi Bougares*[1]*, Yannick Esètve*[1] *and Lamia Belguith*[2]

[1]LIUM, University of Maine, France
[2]ANLP Research group, MIRACL Lab., University of Sfax, Tunisia

masmoudiabir@gmail.com, mariem.ellouze@planet.tn, fethi.bougares@lium.univ-lemans.fr,
yannick.esteve@lium.univ-lemans.fr and l.belguith@fsegs.rnu.tn

Abstract

Conditional Random Fields (CRFs) represent an effective approach for monotone string-to-string translation tasks. In this work, we apply the CRF model to perform grapheme-to-phoneme (G2P) conversion for the Tunisian Dialect. This choice is motivated by the fact that CRFs give a long term prediction and assume relaxed state independence conditions compared to HMMs [7]. The CRF model needs to be trained on a 1-to-1 alignement between graphemes and phonemes. Alignments are generated using Joint-Multigram Model (JMM) and GIZA++ toolkit. We trained CRF model for each generated alignment. We then compared our models to state-of-the-art G2P systems based on Sequitur G2P and Phonetisaurus toolkit. We also investigate the CRF prediction quality with different training size. Our results show that CRF perform slightly better using JMM alignment and outperform both Sequitur and Phonetisaurus systems with different training size. At the end, our system gets a phone error rate of 14.09%.

Index Terms: Tunisian Dialect, Grapheme-to-phoneme conversion, CRF models, JMM, GIZA

1. Introduction

Grapheme-to-phoneme (G2P) conversion can be defined as the task of determining the pronunciation of a word from its written form [11]. It is needed for several speech processing applications such as automatic speech synthesis and speech recognition. G2P conversion of the Tunisian Dialect is a complex task because the correspondence between the orthography (spelling) and its phonetic transcription is not completely consistent. This complexity stems from four factors. Firstly, multiple (sometimes one, more than one or zero) grapheme(s) can correspond to multiple (sometimes one, more than one or zero) phoneme(s). Secondly, the absence of vowelization (short vowels) generates ambiguity at the phonetic level and consequently at the lexical, syntactic and semantic levels. Thus, a non-vowelized Tunisian Dialect word can have multiple pronunciations. Thirdly, the presence of foreign words in the vocabulary of this language makes G2P conversion difficult [1]. Fourthly, the grapheme-phoneme relations are sometimes ambiguous.

Besides, the Tunisian Dialect is characterized by the existence of six major dialectal areas: the North-East area, the Northwest area, the coastal area, the area of Sfax, the South East area and the South West area [12]. These varieties of areas affect the phonological system of the Tunisian Dialect. Indeed, the pronunciation of a word varies from one region to another. In fact, certain letters are pronounced in a region can be not pronounced in another region. For example, the word /AlqyrwAn/ the citys name can be pronounced /AlqrwAn/ by removing the letter /y/. Consequently, they make the G2P conversion of the Tunisian Dialect difficult. Moreover, the Tunisian Dialect G2P conversion is marked by several problems. Among these problems, we can cite the presence of phenomenon of liaison, elision, metatheses and assimilation. Also, we noticed that the phonology of the Tunisian Dialect presents some intriguing aspects; for example, we can find a variation in the pronunciation of some consonants and vowels.

All these factors mentioned above make G2P conversion of the Tunisian Dialect a non-trivial and an interesting problem. In this regard, we need to choose an approach of the G2P conversion to take into account all these cases and generate all possible pronunciations for each word.

In the literature, most works of G2P conversion have used two approaches: the first one is a rule-based approach in which the conversion is done by applying phonetic rules such as [5] and [15]. The second is the data-driven approach which enables to learn the pronunciation generation using statistical models such as decision trees [9], HMM [18], Joint-Multigram Models (JMM)[11] and CRF [7].

In order to perform a G2P conversion of the Tunisian Dialect, we focus on a data-driven G2P approach based on Conditional Random Fields (CRF) [10]. CRFs are undirected graphical models in which each vertex represents a random variable whose distribution is to be inferred [7]. We have chosen CRF to perform this task for various reasons. Firstly, in the literature CRF provided good results for several NLP tasks, especially tasks with monotonic alignments. Indeed, this is a significant feature of the G2P conversion task. Secondly, CRF have been used for G2P conversion for many languages such as English and French and it provides good results. Finally, the advantages of CRFs are relaxed independence conditions compared to HMMs, a global inference algorithm, and discriminative training [7].

The main contributions of this article are as follows:

- Describing the CRFs based approach for Tunisian G2P conversion;
- Investigating the different aspects of the proposed approach on several test sets: evaluating several features, analyzing GIZA and JMM alignments;
- Studing n-best predictions pronunciations
- Our final results show that our system is comparable to state-of-the-art systems on a large pronunciation dictionary.

The remainder of this paper is organized as follows: Section 2 explains the theoretical foundations of the CRF-based G2P

http://dx.doi.org/10.21437/Interspeech.2016-1320

approach. In section 3 and 4, we report our experimental results and their analysis with respect to implementation aspects and a comparison with state-of-the-art approaches. Section 5 conclude the paper and present an insight on the future work.

2. CRF for G2P conversion

A CRF models the conditional probability distribution of a label sequence given an observation sequence [4]. In the context of G2P conversion, a grapheme sequence is considered as observation sequence and a phoneme sequence is the label sequence to be inferred. One constraint for the use of CRFs is that a 1-to-1 alignment between graphemes and phonemes which is necessary to train the model. However, in the Tunisian Dialect the phonemes and graphemes sequences of words are often of different length. This difference of length is due to several factors; we will present some in following:

- Usually, written texts of the Tunisian Dialect are not vowelized. Therefore, short vowels are absent in the grapheme part. However, at oral, phonemes of these short vowels are pronounced naturally.
- In several examples of Tunisian Dialect word, a double graphemes represented by a single phoneme as the case of Waw jame3a [plural in Arabic] that is composed of Waw and Alif [double graphemes] but in phonemic part is presented by simple phoneme UW.
- On the other side, two phonemes may correspond to one grapheme. For example, when the Alif and lem followed by a solar consonant, so during pronunciation there are a doubling of phoneme of this consonant. Thus, there is a consonant in the part of graphemes and two phonemes that correspond to this consonant.
- In another situation, we can find the problem of some graphemes which are quiescent. For example, the tamarbouta at the end of the word is always not pronounced.

These various cases show that there is a possible length difference between the grapheme sequence and the phoneme sequence of words in the Tunisian Dialect.

The alignment grapheme-to-phoneme is usually provided by an external model and can easily be transferred to a 1-to-1 alignment. The CRF based G2P conversion of the Tunisian Dialect is done in two steps: first a grapheme-to-phoneme alignment is generated for all the words of the training dictionary. Then, this dictionary is used to train the CRF-based models. Finally, these models are evaluated using a held-out test set.

2.1. Alignment step

In order to produce the graphemes-to-phonemes alignments required to train CRF model, we used two external alignments tools: GIZA ++ [6] and JMM [11].

2.1.1. GIZA++ based alignment

GIZA++ [6] is a statistical machine translation toolkit for word alignment between two languages: source and target language. In this work, we employ the GIZA++ toolkit to get alignments between a sequence of graphemes and a sequence of phonemes. Indeed, it treats the set of sequence of grapheme as a source language and the set of sequence phoneme as a target language.

After runing GIZA++, forced alignment between graphemes and phonemes of all the words of the training corpus is performed. The format of the alignment obtained is different to that admitted by the CRF. For this, we apply some pre-processing to extract the associations between one grapheme and one phoneme as it is necessary for CRF. For this, before each grapheme we put its phoneme. In the case of absence of grapheme or phoneme, we simply put an epsilon ϵ.

2.1.2. Joint-Multigram-Model-based alignment

JMM have been used generally to perform the G2P conversion task directly [11]. The fundamental idea of JMM is based on the concept of a *graphone u*, denoted $u = (\tilde{g}, \tilde{q})$ where $\tilde{g}$ represents a pair of a grapheme sequence and $\tilde{q}$ represents a phoneme sequence. Hence, graphones U defines the joint probability of spelling G and pronunciation Q [4].

$$P(G|Q) = \sum_{U;G(U)=G;Q(U)=Q} P(U) \quad (1)$$

$$= \sum_{U;G(U)=G;Q(U)=Q} P(u_1, u_2, ...u_K) \quad (2)$$

where *G(U)* and *Q(U)* denote the grapheme and phoneme component of U, respectively. In our case, to perform the alignment between phoneme and grapheme sequences we employ the JMM. So, the probability *P(U)* becomes as graphone n-gram model:

$$P(U) = \prod_{j=1}^{|U|} P(u_j|h_j) \quad (3)$$

where h_j is the graphone history of u_j.

In our experiment, we chose an 8-gram model to perform grapheme/phoneme alignment. Moreover, we used 0-1 graphones for alignment, meaning that either one or zero phoneme is allowed to be aligned to one or zero grapheme.

2.2. CRF model training

CRF are probabilistic models for computing the conditional probability of a possible output given an input sequence also called the observation sequence. In order to train G2P associations and some predefined feature sets, CRF learns a set of weights w. Learning the parameter set w is usually done by maximum likelihood learning for $p(\bar{x}|\bar{y}; w)$:

$$p(\bar{x}|\bar{y}; w) = \frac{1}{z(\bar{x}|w)} exp \sum_j w_j F_j(\bar{x}, \bar{y}) \quad (4)$$

$$p(\bar{x}|\bar{y}; w) = \sum_{i=1}^{n} f(\bar{y}_{i-1}, \bar{y}_i, \bar{x}, i) \quad (5)$$

According to these equations, $\bar{x}$ represents the sequence of graphemes (observation), $\bar{y}$ represents the sequence of phonemes, and w represents the weights. f_j corresponds to a feature function. This function depends on the sequence of word letters, the current phoneme, the previous phoneme and the current position in the word. While Equation (4) expresses the unigram features, Equation (5) expresses bigram features. Unigram features mean that only current phonemes will be taken into account whereas bigram features use the current and the previous phoneme.

3. Experimental setup

3.1. The Tunisian Dialect Phonetic Dictionary: TunDPDic

To measure the performance of our proposed approach, the TunDPDic (The Tunisian Dialect Phonetic Dictionary) pronunciation dictionary has been used. TunDPDic is a Tunisian Dialect phonetic dictionary generated by our internal rule-based tool [3]. This dictionary is verified and reviewed by experts to correct errors if they exist. The principle of a rule-based approach consists in using a set of phonetic rules and a lexicon of exceptions. This lexicon is consulted before the rules are used. If the word is among the exceptions, it is encoded directly in phonetic form. Otherwise, we must apply a set of rules to generate its phonetic form. Our rules are provided for each letter in the Tunisian Dialect. Each rule tries to match certain conditions relative to the context of the letter and to provide a phonetisation. Each rule is read from right to left and follows this format [2] :

- **Graph** : current letter in the word;
- **Right-Condition** : context before the current position;
- **Left-Condition** : context after the current position;
- **Phonetisation** : is either a phoneme or more of a phoneme or a vacuum (*) if the graph is omitted in pronunciation.

The TunDPDic consists of about 18K words. Tunisian Dialect contains 32 letters and a phone set of 39 phonemes. We divided randomly this corpus into disjoint training (75%), development (5%), and test (20%) sets.

3.2. Performance metrics

All G2P conversion models presented in this paper are evaluated using the phoneme error rate (PER) and the word error rate (WER) metrics.

3.3. Used software

3.3.1. The CRF++ software

CRF++[1] is a customizable and open source implementation of CRF for segmenting and labeling sequential data. It is written in C++, uses fast training based on gradient descent and generates n-best candidates.

3.3.2. The Phonetisaurus software

Phonetisaurus [2] is used for sake of comparison with state-of-the-art G2P system. Phonetisaurus is a WFST[3] -driven G2P framework suitable for rapid development of high quality G2P or P2G systems. This software includes a fast, EM-driven, WFST-based multiple-to-multiple alignment program, model conversion tools, a fast WFST-based decoder, and a Lattice Minimum Bayes-Risk decoder implementing a novel length-normalized loss function for computing N-gram factors. A specialized test distribution implementing N-best rescoring with Recurrent Neural Network Language Models via RNNLM is also included.

In this paper, Phonetisaurus model is trained used with 8-gram back-off LM. The model is optimized using a development set applied to generatge the pronunciations of the test lexicon entries.

[1] crfpp.sourceforge.net

[2] https://code.google.com/p/phonetisaurus/

[3] WFST: Weighted Finite-State Transducer

3.3.3. The Sequitur G2P software (JMM)

Joint-Multigram Model (JMM) approach is also used as a state-of-the-art approach. JMM is trained using the Sequitur [4] G2P software. The key idea of JMM is to determine the optimal set of joint sequences, where each sequence is in fact composed of a sequence of graphemes and its associated sequence of phonemes.

4. Experimental results

4.1. 1-best prediction

4.1.1. Impact of training, development and test set size

In this section, we study the influence of the training, development and test set size for CRF prediction. For this we defined different data size and we divided this data to a train, dev and test set as presentaed in table 1.

Table 1: *Training, development and test size per set : 1 to 5.*

	Train	Dev	Test
Set 1 (5K)	3.75K	0.25K	1K
Set 2 (7K)	5.25K	0.35K	1.4K
Set 3 (8K)	6K	0.4K	1.6K
Set 4 (10K)	7.5K	0.5K	2K
Set 5 (18K)	13.5K	0.9K	3.6K

Using table 1 sets different G2P systems are trained using CRF [5], JMM prediction and Phonetisaurus. The results of each G2P system for different set is presented in the following table.

Table 2: *PER of G2P conversion for CRF, JMM and Phonetisaurus using sets 1 to 5.*

	CRF	**JMM**	**Phonetisaurus**
Set 1	22.87%	23.57%	28.46%
Set 2	21.54%	22.13%	25.28%
Set 3	20.74%	21.51%	23.21%
Set 4	19.74%	20.41%	21.54%
Set 5	14.31%	16.32%	17.83%

According to table 2, even though we used only about half of our corpus, the performance of the CRF system is not so bad. For example, when we employed the test set of **Set 3**, we obtained a 20.74% of (PER) with CRF, a 21.51% of (PER) with JMM and a 23.21% (PER) using Phonetisaurus. Furthermore, and as expected, the best result of CRF prediction (PER of 14.31%) is obtained when we employed the bigger training data (**Set 5**). Finally, we can see also that the CRF model outperform significantly JMM and Phonetisaurus for all data sizes. This result (set 5 in table 2) presents a 2.01% absolute improvement in comparison with JMM generated pronunciations and 3.52% absolute improvement in comparison with Phonetisaurus generated pronunciations.

4.1.2. Alignments impact on CRF prediction

In a second experiment, we wanted to investigate the effect of the alignment on the performance of the CRF model. We tested two different external models to produce alignments: GIZA++

[4] www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

[5] CRF is trained using 2-gram and GIZA alignment

and JMM toolkit. For each of the resulting alignments, a CRF was trained with exactly the same features. Thus, only the influence of the alignment could be observed. Table 3 contains the results for the CRF prediction using both alignment. Comparing these results, we noticed a slight performance decrease with the JMM 0.22% (PER) compared to the GIZA alignment. In this experiment, we used **Set 5** (18K) from table 1.

Table 3: *Impact of alignments on CRF G2P conversion*

	PER% GIZA++ alignm	PER% JMM alignm	WER% GIZA++ alignm	WER% JMM alignm
CRF	14.31%	14.09%	21.35%	20.48%

4.1.3. Effect of unigram and bigram features

In this third experiment, we studied the effect of n-gram order feature of the CRF. In addition, we examined different widths of grapheme contexts that the features may cover. To explain each of these features:

- Unigram features take into account only the current phoneme while bigram features use the current and previous phonemes.
- Widths of grapheme contexts mean the features may cover n graphemes preceding and following the current position. For example, (± 2) means that the features may cover two graphemes preceding and following the current position.

Using **Set 5**, table 4 present the results with JMM alignment using different n-gram order and variate grapheme contexts width. The best PER is obtained using bigram feature with a grapheme contexts of (± 1).

Table 4: *Impact of n-gram order and grapheme contexts width on JMM G2P conversion*

	PER% JMM alignm (unigr)	PER% JMM alignm (bigr)
CRF (± 1)	14.36%	14.09%
CRF(± 2)	14.51%	14.33%
CRF(± 3)	15.07%	15.10%
CRF(± 4)	15.34%	15.27%

Given the results obtained with JMM, we fixed the n-gram features to 2 and we train several CRF models with different size of grapheme contexts using JMM and GIZA++ alignment. As presented in table 5, the JMM alignment gives lower PER for all grapheme contexts width.

4.2. N-best prediction

In this section, we study the n-best outputs of G2P. For each word, we generate the n-best list pronunciations. We evaluated the n-best quality for $n = 4$ using recall and precision measure. The recall (R) represents the number of correct pronunciation variants generated divided by the total number of reference pronunciation variants. Whereas, the precision (P) is the number of correct pronunciation variants divided by the total number of generated pronunciation variants. In our experiments, only probability of generated pronunciation variants is greater than a threshold T ($T = 0.35$) that are retained.

Table 5: *Impact of grapheme contexts width and alignments model on G2P conversion tasks*

	PER% GIZA++ alignm	PER% JMM alignm
CRF (± 1)	14.31%	14.09%
CRF(± 2)	14.83%	14.33%
CRF(± 3)	15.67%	15.10%
CRF(± 4)	15.91%	15.27%

Recall and precision of JMM and CRF G2P conversion are presented on table 6. Theses results shows that better results are obtained with 4-best pronunciations of each word whatever the size of training data.

Table 6: Recall and precision for CRF and JMM G2P conversion using n-best pronunciations for each word

	CRF		
	Set3	Set4	Set5
Recall	88.51%	90.04%	91.41%
Precision	84.45%	86.73%	87.13%
	JMM		
	Set3	Set4	Set5
Recall	84.45%	85.75%	87.15%
Precision	80.42%	82.93%	83.46%

The best result of n-best pronunciation is obtained with CRF model using Set 5 : 91.41% of recall and 87.13% of precision. Overall, the CRF G2P conversion of the Tunisian Dialect gives a PER and recall high compared to other languages such as French and English [7]. This is due to the absence of short vowels in the training and test corpus.

5. Conclusion

In this paper, we proposed an approach to G2P conversion of the Tunisian Dialect based on a probabilistic method: CRF. In order to generate 1-to-1 alignment training exemples needed to train CRF, we employed GIZA++ and JMM. To measure the performance of the approach proposed in this article, different parameters were studied: training set size, effect of various alignments for CRF prediction, effect of unigram and bigram features and the multiple pronunciation generation. The best CRF configuration was identified and compared to the state-of-the-art JMM system and the Phonetisaurus system. Our results shows better precision and recall performance using n-best pronunciation. In the future, we plan to improve several aspects of our models, particularly the integration of others features into our G2P system such as POS of the Tunisian Dialect. We plan also to integrate and investigate the behavior of our G2P system when integrated into a Speech recognition system.

6. References

[1] A. Masmoudi, M. Ellouze Khmekhem, Y. Estève, F. Bougares and L. Hadrich Belguith, Phonetic tool for the Tunisian Arabic, In the 4th International Workshop on

Spoken Language Technologies for Under-resourced Languages, 2014.

[2] A. Masmoudi, M. Ellouze Khmekhem, Y. Estève, F. Bougares, L. Hadrich Belguith Habash and N. Habash A corpus and a phonetic dictionary for Tunisian Arabic speech recognition, In 19th edition of the Language Resources and Evaluation Conference, Iceland, 2014.

[3] A. Masmoudi, M. Ellouze Khmekhem, Y. Estève, F. Bougares, S. Dabbar and L. Hadrich Belguith, Phonètisation automatique du dialecte Tunisien. In 30e Journes dtudes sur la parole, 2014.

[4] D. Wang and S. King, Letter-to-sound Pronunciation Prediction Using Conditional Random Fields", IEEE Signal Processing Letters, pp.122-125, 2011.

[5] F. Bèchet, LIA_PHON: un système complet de phonètisation de textes, Traitement Automatique des Langues - TAL - vol.42, no.1, pp.47-67, 2001.

[6] F. Casacuberta and E. Vidal, GIZA ++ : Training of statistical translation models, 2007.

[7] I. Illina, D. Fohr and D. Jouvet, Grapheme-to-Phoneme Conversion using Conditional Random Fields, Interspeech'2011, 2011.

[8] I. Zribi, R. Boujelbane, A. Masmoudi, M. Ellouze, L. Belguith and N. Habash,A Conventional Orthography for Tunisian Arabic, Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC-2014, pp.2355–2361, 2014.

[9] J. Hakkinen, J. Suontausta, S. Riis, and K. J. Jensen, Assessing text to- phoneme mapping strategies in speaker independent isolated word recognition. Speech Communication, vol.41, no.2, pp. 455467, 2003.

[10] J. Lafferty, A. McCallum and F.Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proc. ICML, 282-289, 2001.

[11] M. Bisani and H. Ney, Joint-Sequence Models for Grapheme-to- Phoneme Conversion", Speech Communication, vol.50, pp.434-451, Elsevier, 2008.

[12] M. Graja, M. Jaoua and L. Belguith, Statistical Framework with Knowledge Base Integration for Robust Speech Understanding of the Tunisian Dialect, IEEE/ACM Trans. Audio, Speech & Language Processing, Vol.23, no.12, pp. 2311–2321, 2015.

[13] M. Maamouri, T. Buckwalter and C. Cieri, Dialectal Arabic Telephone Speech Corpus: Principles, Tool Design, and Transcription Conventions, In: NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, September, pp. 22-23, 2004.

[14] S. F. Chen, Conditional and joint models for grapheme-to-phoneme conversion, in Proc. Eurospeech03, Geneva, Switzerland, pp. 20332036, 2003.

[15] S. Harrat, K. Meftouh, M, Abbas and K. Smali, Grapheme to Phoneme Conversion-An Arabic Dialect Case. In Spoken Language Technologies for Under-resourced Languages, 2014.

[16] S. Lawson and I. Sachdev, Code Switching in Tunisia: attitudinal and behavioral dimensions, In Journal of Pragmatics, vol. 32, no. 9, pp.1343-61, 2000.

[17] S. Mejri, M. Said, M and I. Sfar, Pluringuisme et diglossie en Tunisie, In: Synergies Tunisie, vol.1 , pp.53-74, 2009.

[18] P. Taylor, Hidden Markov Models for Grapheme to Phoneme conversion, Proc. Interspeech'2005, pp.1973-1976, 2005.