

Applying Weighted Finite State Transducers and Ripple Down Rules for Myanmar Name Romanization

1st Wint Theingi Zaw

*Department of Computer Engineering and Information
Technology,
Yangon Technological University,
Yangon, Myanmar,
wint.wtgz@gmail.com*

3rd Ye Kyaw Thu

*National Electronics and Computer Technology Center
(NECTEC),
Pathumthani, Thailand,
Language Understanding Lab.,
Myanmar,
yktnlp@gmail.com*

2nd Shwe Sin Moe

*Department of Computer Engineering and Information
Technology,
Yangon Technological University,
Yangon, Myanmar,
shwesinmoe.ytu25@gmail.com*

4th Nyein Nyein Oo

*Department of Computer Engineering and Information
Technology,
Yangon Technological University,
Yangon, Myanmar,
nno2005@gmail.com*

Abstract— Romanization known as Latinization that refers to the representation of names with Roman (Latin) alphabets. The Romanization process is not trivial especially for Myanmar Language (Burmese) due to different Roman variations of a single Myanmar proper noun. The Myanmar Name Romanization project has so far developed a 55K Romanized word pairs of Myanmar personal names. We apply Weighted Finite State Transducers (WFST) and Ripple Down Rules (RDR) based tagging approaches to the task of Myanmar name Romanization and vice versa. We perform experiments on one closed test data set (5,000 of training data) and three open test data sets: 5,000 Myanmar personal names, 571 city and town names of Myanmar, and 292 Myanmar food names. The result shows that RDR approach gives better performance than WFST for open test data sets.

Keywords— Myanmar Name Romanization, Weighted Finite State Transducers (WFST), Ripple Down Rules (RDR)

I. INTRODUCTION

Romanization is the process of translating from a non-Latin writing system into Latin script. It is an important task in natural language processing such as name translation in machine translation. There is no standard Romanization system for Myanmar Language. According to our knowledge, there are several Romanization systems such as the Myanmar Language Commission (MLC) transcription system [1], [2], the University of Foreign Language (UFL) pronunciation system [3] that is an extended version of MLC, and ALA-LC Romanization for Burmese that is used by the Library of Congress for cataloguing Burmese language book holdings [5], [4]. Moreover, Myanmar people used several Romanization methods such as transliteration, transcription and combined systems based on their English and Myanmar word spelling knowledge (local Romanization systems) [6]. For example, Romanization of Myanmar word “မော်တော်ကား” (motorcar in English) can be “mawtauka:” (with MLC), “mo.to.ka” (with UFL) and “mawtawcar” or “maw taw kar” (with local Romanizations). This paper investigates the automatic Romanization of Myanmar names and conversion of Romanized Myanmar names to original Myanmar graphemes based on Weighted Finite State Transducers (WFST) [8] and Ripple Down Rules (RDR) [7].

One more contribution is developing a 55K parallel corpus for Romanized Myanmar names.

The structure of the paper is as follows: in section.II, a brief review of Myanmar Romanization, WFST and RDR are presented. The nature of Myanmar Language is explained in section.III. Section.IV describes the parallel corpus building of Myanmar name Romanization for automatic conversion study. In section.V introduces the WFST and RDR methodologies. Section.VI presents information of the corpus and the experimental settings. Section.VII reports the experimental results with some discussions and Section.VIII presents the error analysis on the experimental results followed by a conclusion in section.IX.

II. RELATED WORK

A brief introduction of previous studies related to Myanmar Romanization and applying WFST and RDR for natural language processing tasks such as part-of-speech (POS) tagging, text to phonetic transcription are presented.

Lei Lei Win [9] presented the rule-based Pali word Romanization system for Myanmar language. In this paper, firstly, ten Romanization rules for Pali are defined, and checked Myanmar Pali words, Romanized these Pali words and finally, transformed Text to Speech. Even though the accuracy of Romanization is 89.6, this system is still lack of checking in Pali words.

Taguchi et al. [11] proposed a method for inducing Romanization systems directly from a bilingual alignment at the grapheme level. First, transliteration word pairs are aligned using a non-parametric Bayesian approach, and then for each grapheme sequence to be Romanized, a Romanization is selected according to a user specified criterion. The approach was applied to the task of transliteration mining and used Levenshtein distance as the selection criterion. The experimental results on three languages: Japanese, Russian and Chinese proved that the mining system built from the induced Romanization system is able to outperform existing baseline Romanization systems.

Ye Kyaw Thu et al. [10] described the comparison of Grapheme-to-Phoneme (G2P) conversion methods on a Myanmar pronunciation dictionary. This paper used 25,300 words of Myanmar Language Commission Dictionary data. In

this experiment, 25,000 words used for training and 300 words used for three open test sets (100 words for each open test set) for evaluation. Experiment results showed that CRF, PBSMT and WFST approaches were the best performance for G2P conversion on Myanmar Language according to phoneme error rate (PER) and manual checking.

Khin War War Htike et al. [12] presented the comparison of six POS tagging methods on 10K sentences Myanmar Language (Burmese) POS tagged corpus. There were 11,000 sentences (including various area). In this experiment, 10K sentences were used for training, 10% of training data were used as closed test set and 1K sentences were used as open test data for evaluation. The results showed that RDR approach can consistently achieved 97.05% on open data set and best among six POS tagging methods (CRF, HMM, MaxEnt, RDR, SVM, Two-Hours).

III. NATURE OF MYANMAR LANGUAGE

Myanmar language is the official language which is spoken by two-thirds of the population in Myanmar. Myanmar alphabet contains 33 consonants and 12 vowels, and is written from left to right. Syllable or words are composed of consonants combining with vowels. Vowels are placed above, below, before and/or after the consonant character. For example, “ka” consonant with a right vowel “ye char” ကာ (“protect” in English), “ka” consonant with a down vowel “ta chaung ngin or U” ကု (“treat” in English), “ka” consonant with a left vowel “tha way htoo or E”, down vowel “wa swe or Wa” and right vowel “witsa nha lone pauk or Visarga” combination ကွေး (“curved” in English). Some syllables can be composed of consonants without any vowels such as a single consonant word က (“dance” in English).

IV. BUILDING PARALLEL CORPUS

Currently there is no freely available Myanmar name Romanization corpus. Thus, we are developing a parallel corpus for Romanization of Myanmar names. We consider all possible Romanizations and especially focus on a real world Romanizations of Myanmar people (i.e. local Romanization). For example: “မြိုး” can be Romanized as “Phyo” and “Phyoe”, “သန္တာ” can be Romanized as “Thanda” and “Thandar”, and “ထွန်း” can be Romanized as “Tun”, “Htun” and “Htoon”. And vice versa, one English word “Thu” can be Romanized into three Myanmar syllables: “သု”, “သူ”, “သူး” in Myanmar. In details, our corpus has one-to-many association for a Myanmar word to Romanized words and many-to-one association for Romanized words to a Myanmar word. In this case, one Myanmar name syllable can be Romanized at most six Romanized words in my corpus. For example, the “မြိုးသန္တာထွန်း” Myanmar personal name can Romanize into several English words as shown below.

Many-to-One Association:

Myanmar English

ပိုင်သူ } Paing Thu
ပိုင်သူ }

One-to-Many Association:

Myanmar

English

မြိုးသန္တာထွန်း { Phyto Thanda Tun
Phyto Thanda Htun
Phyto Thanda Htoon
Phyto Thandar Tun
Phyto Thandar Htun
Phyto Thandar Htoon
Phyoe Thanda Tun
Phyoe Thanda Htun
Phyoe Thanda Htoon
Phyoe Thandar Tun
Phyoe Thandar Htun
Phyoe Thandar Htoon

V. METHODOLOGY

In this section, Weighted Finite State Transducers and Ripple Down Rules approaches are briefly explained.

A. WFST-based Romanization

A weighted finite state transducer (WFST) [15] is a finite automaton in which each transition has an input, output and a weight. Fig. 1 depicts a weighted finite state transducer: the initial state is labeled 0 and final state is 2 with final weight of 3.5. There is a transition from state 0 to 1 with input label (“ခင်”), output label (“Khin”), and weight 1. Another transition from state 0 to 1 is input label (“စု”), output label (“Su”), and weight 1.5. This machine transduces the words “ခင်တိုး” to “Khin Toe” with weight 7.0 which is the sum of the arc and final weight and the words “စုတိုး” to “Su Toe” with weight 7.5 [13], [20]. In these sorts of WFST graphs, common assumption is that higher weights are worse and search the lowest score path through the whole graph.

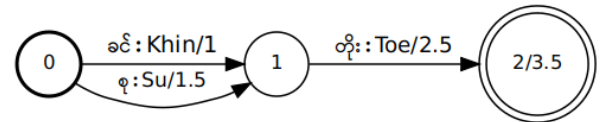


Fig. 1. An Example of a Weighted Finite State Transducer.

WFST-based Myanmar name Romanization needs the following stages:

- Alignment: This step produces one-to-many and many-to-one alignments based on EM (Expectation Maximization Algorithm).
- Building Joint Sequence n -gram Model: The aligned corpus is the input at this step. It builds an n -gram model over sequences of aligned a Myanmar name to Romanized syllables (like grapheme-phoneme symbols aligning in G2P conversion). We build a language model using KenLM [14] and the language model is converted into a WFST using OpenFST [13] for decoding.
- Decoding: The decoding computes the best shortest path in the following equation:

$$H_{best} = ShortestPath(Project_o(w \circ C)) \quad (1)$$

where (H_{best}) refers to the lowest cost path, ($Project_o$) refers to projecting the output labels, (w) refers to test words, (C) refers to the output of model and \circ indicates composition.

B. Ripple Down Rules-Based Tagging

Ripple Down Rules (RDR) [21], [16] is an approach building knowledge-based systems. RDRPOSTagger provides an error-driven approach to automatically restructure transformation rules in the form of a Single Classification Ripple Down Rules (SCRDR) tree. A SCRDR can be written as a triple $\langle \text{rule}, X, Y \rangle$, where X and Y are the exception RDR and the succeeding RDR (i.e. if-not rules) respectively.

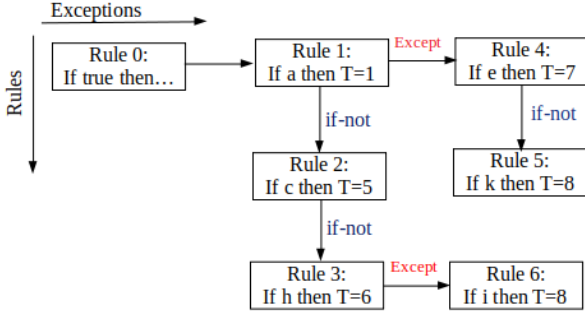


Fig. 2. A binary tree of Single Classification Ripple Down Rules.

For example, the SCRDR tree in the Fig. 2, given a case “i/T=8 a/T=1 k/T=8” where “a/T=1” is a current Myanmar syllable and tag pair (here, a tag represents a Romanized Myanmar syllable), the case satisfies the condition of the Rules (0) and (1), it then is continued to the Rule (4) using except link. As the case does not satisfy the condition of the Rule (4), it will be transferred to Rule (5) using if-not link. Therefore, the tag for “a” is concluded as “T=8”.

The case in SCRDR is evaluated by passing a case to the root node (Rule 0) in Fig. 2. Rule (1) is the exception rule of the default node (Rule 0). Rule (2) is the if-not child node of the Rule (1) and an exception rule of the Rule (0). And then, Rule (4) and (5) are exception rules of Rule (1) etc. [17].

VI. EXPERIMENT SETUP

A. Corpus Statistics

Both Myanmar names and respective English names were collected from student affairs in our university and social media such as Facebook by manually. We used 55,111 Myanmar personal names which include not only Burmese names but also ethnic names in Myanmar such as “Chin”, “Kayin”, “Rakhine” and “Shan”. In this experiment, 50,111 names (more than 90% of total personal names) are trained.

B. Closed and Open Test Sets

There are one closed test data and three domains open test data sets. Closed test data set contains 5,000 personal names that are extracted from our training data (around 10% of training data). And three domains open test data sets: one is “Myanmar names” that are extracted from our corpus which contains 5,000 names (around 10% of the Myanmar personal name corpus). The numbers of other two domains open test data sets are 571 city, town names for “the names of city and town in Myanmar” and 292 food names for “Myanmar

traditional food” respectively. These two different domains data were manually collected from the Wikipedia.

C. Syllable Segmentation

In experiment, Regular Expression based Myanmar Syllable segmentation tool named “sylbreak” (<http://github.com/ye-kyaw-thu/sylbreak>) is used for syllable segmentation of Myanmar names. Example of Myanmar names syllable segmentation are as following:

သဇင်မြတ် | သ ဇင် မြတ်

ခင်စန္ဒာစိုး | ခင် စန္ဒာ စိုး

ခင်ကြည်သာခိုင် | ခင် ကြည် သာ ခိုင်

D. Software for Building WFST and RDR Models

- Phonetisaurus: A WFST-driven G2P converter [15] (source code link: <https://github.com/AdolfVonKleist/Phonetisaurus>).
- RDRPOSTagger: RDRPOSTagger (Version 1.2.4) is a rule-based Part-of-Speech and morphological tagging toolkit [16], (source code link: <https://github.com/datquocnguyen/RDRPOSTagger>). It is a robust, easy-to-use and language-independent toolkit.

E. Evaluation

The performance of our system was measured by accuracy using the following equation:

$$Accuracy = \frac{\# \text{ of correct conversion}}{\# \text{ of total syllables in test set}} \quad (2)$$

F. Word Error Rate (WER)

Romanized output is analyzed by Word Error Rate [18] and the SCTK (the NIST Scoring Toolkit) version 2.4.10 [19] for making dynamic programming-based alignments between reference (ref) and hypothesis (hyp) and calculation. The formula can be stated as (3).

$$WER = \frac{(I + D + S) \times 100}{N} = \frac{(I + D + S) \times 100}{S + D + C} \quad (3)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words and N is the number of words in the reference ($N = S + D + C$).

VII. RESULT AND DISCUSSION

The accuracies of two models (WFST and RDR) on closed test data set and three open test data sets are shown in Table. I, Table. II and Table. III respectively. Bold numbers show the highest scores of these approaches.

The experimental results show that WFST gives slightly better accuracy than RDR with closed test data set in Table. I. On the other hand, RDR achieves better accuracy in Myanmar to Romanization for open test data sets in Table. II that are 0.64 with personal name data (5,000 names and 10% of the corpus), 0.38 with town and city name data (571 names) and 0.31 with food name (292 names). And RDR also achieves better accuracy in Romanized name to Myanmar name for

open test data sets in Table. III that are 0.97 with personal names data, 0.34 with town and city names data and 0.30 with food names.

To compare Myanmar to Romanization Table. II with Romanization to Myanmar Table. III, it can be seen clearly that the accuracy of Romanized name to Myanmar name conversion is significantly higher than Myanmar name to Romanized name conversion with the same domain (i.e. personal name) data. However, the accuracies of Myanmar name to Romanized name conversion are higher for the different domain test data sets (i.e. town, city names and food names). This is because one Romanized Myanmar syllable can be translated into six Myanmar syllables (e.g. “tone” ==> “တုန်” or “တုံ” or “တုန်” or “တုန်း” or “တုံး” or “တုန်း”). Similarly, one Myanmar syllable can have many possible Romanizations (e.g. “ထွန်း” ==> “tun” or “htun” or “htoon”).

TABLE I. ACCURACIES OF WFST AND RDR MODELS FOR CLOSED TEST DATA SET

Model	Myanmar-Roman	Roman-Myanmar
WFST	0.71	0.99
RDR	0.70	0.98

TABLE II. ACCURACIES OF WFST AND RDR MODELS FOR MYANMAR NAMES ROMANIZATION WITH THREE OPEN TEST DATA SETS

Model	Personal Name	City, Town Name	Food Name
WFST	0.59	0.35	0.24
RDR	0.64	0.38	0.31

TABLE III. ACCURACIES OF WFST AND RDR MODELS FOR ROMANIZED WORDS TO MYANMAR WORD CONVERSION WITH THREE OPEN TEST DATA SETS

Model	Personal Name	City, Town Name	Food Name
WFST	0.97	0.31	0.21
RDR	0.97	0.34	0.30

VIII. ERROR ANALYSIS

WER calculation (refer 3) for both WFST and RDR outputs was done with “SCLITE” command of the SCKT toolkit [19]. The WER results of WFST and RDR models with closed test data for Myanmar name Romanization and conversion of Romanized name to Myanmar name can be seen in Table. IV.

TABLE IV. WER OF WFST AND RDR MODELS FOR CLOSED TEST DATA SET (HERE, LOWER SCORES ARE BETTER)

Model	Myanmar-Roman	Roman-Myanmar
WFST	28.7%	0.5%
RDR	29.4%	1.4%

The WER results of WFST and RDR models for Myanmar name Romanization based on three open test data can be seen in Table. V.

TABLE V. WER OF WFST AND RDR MODELS FOR MYANMAR NAME ROMANIZATION (HERE, LOWER SCORES ARE BETTER)

Model	Personal Name	City, Town Name	Food Name
WFST	40.9%	67.5%	74.1%
RDR	35.7%	61.0%	68.7%

The WER results of WFST and RDR models for Romanized words to Myanmar word conversion based on three open test data can be seen in Table. VI.

TABLE VI. WER OF WFST AND RDR MODELS FOR ROMANIZED WORDS TO MYANMAR WORD CONVERSION (HERE, LOWER SCORES ARE BETTER)

Model	Personal Name	City, Town Name	Food Name
WFST	2.3%	71.3%	77.1%
RDR	2.5%	65.5%	69.4%

The followings are some example calculations. For example, translated for Myanmar name “ကျော့ကေခိုင်လင်း” into Romanized words which compare to a reference sentence, the output of the SCLITE program is as follows:

Scores: (#C #S #D #I) 2 2 0 0

REF: Kyawt Kay Khaing Lin

HYP: Kyawt Kay Khine Lynn

Eval: S S

In this case, two substitutions (Khaing ==> Khine) and (Lin ==> Lynn) happened, that is C=2, S=2, D=0, I=0, N=4 and its WER is 50%.

The following example is for Romanization to Myanmar translation and all translated syllables are correct, C=4, S=0, D=0, I=0, N=4 and its WER is 0%.

Scores: (#C #S #D #I) 4 0 0 0

REF: ဝတ် ရည် လင်း ထက်

HYP: ဝတ် ရည် လင်း ထက်

Eval:

The next one is WER calculation using RDR approach. For example, translated for Myanmar name “အိပိုမောင်” into Romanized tag pairs which compare to a reference sentence, the output of the SCLITE program is as follows:

Scores: (#C #S #D #I) 3 1 0 0

REF: Ei Po Po Maung

HYP: Ei Po Po Mg

Eval: S

In this case, one substitution (Maung ==> Mg) happened, that is C=3, S=1, D=0, I=0, N=4 and its WER is 0.25%.

The following example is for Romanization to Myanmar translation and all translated syllables are correct, C=4, S=0, D=0, I=0, N=4 and its WER is 0%.

Scores: (#C #S #D #I) 4 0 0 0

REF: ခုန့် နိုင် ပြည့် ထွန်း

HYP: ခုန့် နိုင် ပြည့် ထွန်း

Eval:

The top 10 confusion pairs of WFST and RDR for Myanmar name Romanization can be seen at Table. VII. After analyzing the errors of WFST and RDR in details, as we expected the errors are caused by different Romanization system or many-to-many mapping. And thus, although hypothesis and reference are not exactly matched, all top 10 confusion pairs are correct Romanization for native Myanmar speakers.

TABLE VII. TOP 10 CONFUSION PAIRS OF WFST AND RDR MODELS FOR CONVERSION OF MYANMAR NAME TO ROMANIZED NAME WITH CLOSED TEST DATA SET

WFST (Ref-Hyp)	Freq	RDR (Ref-Hyp)	Freq
Htet → Htat	91	Win → Winn	158
Htat → Htet	82	Htat → Htet	148
Khaing → Khine	75	Wynn → Winn	126
Htun → Htoon	72	Linn → Lynn	117
Win → Wynn	70	Tun → Htoon	105
Win → Winn	69	Aeint → Eaint	101
Winn → Wynn	69	Htun → Htoon	93
Tun → Htoon	66	Lin → Lynn	93
Yie → Yi	65	Phyoe → Phyoe	88
Wynn → Winn	64	Yi → Yee	88

IX. CONCLUSION

In this paper, Myanmar name to Romanization and Romanized name to Myanmar name conversion results are presented applying WFST and RDR approaches. We used over 55K Myanmar-Romanized name parallel corpus. According to the experimental results with one closed test data set (personal name), WFST achieves slightly better accuracy than RDR. However, from the experimental results with three open test data sets (personal name, city, town name and food name) RDR model gives better performance than WFST model. The WER scores and detail error are caused by variations of Myanmar language Romanization. In the near future, we plan to investigate Myanmar name Romanization with other sequence learning approaches Conditional Random Fields and Neural Conditional Random Fields.

ACKNOWLEDGMENT

We would like to thank student affairs department at Yangon Technological University for all the help of data collection process.

REFERENCES

- [1] "MLC Transcription System", https://en.wikipedia.org/wiki/MLC_Transcription_System
- [2] Myanmar Language Commission, Ministry of Education, Myanmar: "Myanmar-English Dictionary", 9th Edition, 2008
- [3] University of Foreign Language, Yangon, Myanmar: "An introductory course in Myanmar language", 2005
- [4] "Romanization of Burmese", https://en.wikipedia.org/wiki/Romanization_of_Burmese
- [5] "Library of Congress table for romanization of Burmese", <https://www.loc.gov/catdir/cpsd/romanization/burmese.pdf>
- [6] John Okell, "A guide to the romanization of Burmese", Volume 27 of Publications, Royal Asiatic Society of Great Britain and Ireland James G. Forlong Fund, Luzac for The Royal Asiatic Society of Great Britain and Ireland, 1971
- [7] Richards, Deborah, "Two decades of Ripple Down Rules research", Knowledge Eng. Review. 24, 2009, pp. 159-184
- [8] Roche, Emmanuel; Yves Schabes, "Finite-state language processing", MIT Press, 1997, pp. 1-65
- [9] Lei Lei Win, "Rule-Based Pali Romanization System for Myanmar Language", International Journal of Scientific and Research Publications, Volume 8, Issue 9, September 2018 ISSN 22503153.
- [10] Ye Kyaw Thu, Win Pa Pa, Yoshinori Sagisaka, Naoto Iwahashi, "Comparison of Grapheme-to-Phoneme Conversion Methods on a Myanmar Pronunciation Dictionary", In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), COLING2016, December 11-17, 2016, Osaka, Japan, pp. 11-22.
- [11] Keiko Taguchi, Andrew Finch, Seiichi Yamamoto, Eiichiro Sumita, "Automatic Induction of Romanization Systems from Bilingual Corpora", IEICE Transactions on Information and Systems, Volume E98.D, Number 2, 2015, pp. 381-393
- [12] Khin War War Htike, Ye Kyaw Thu, Zuping Zhang, Win Pa Pa, Yoshinori Sagisaka and Naoto Iwahashi, "Comparison of Six POS Tagging Methods on 10K Sentences Myanmar POS Tagged Corpus", at 7th Workshop on Natural Language and Speech Processing, ICCA2017, Yangon, Myanmar (16th Feb 2017).
- [13] C. Allauzen and M. Riley and J. Schalkwyk and W. Skut and M. Mohri, "OpenFST: A General and Efficient Weighted Finite State Transducer Library", Proc. CIAA 2007, pp. 11-23.
- [14] Heafield, Kenneth, "KenLM: Faster and Smaller Language Model Queries", In Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT 11, Edinburgh, Scotland, 2011, pp. 187-197.
- [15] Josef R. Novak, Nobuaki Minematsu, and Keikichi Hirose. 2012. "WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding", In Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing, FSMNLP 2012, Donostia-San Sebastián, Spain, July 23-25, 2012, pp. 45-49.
- [16] D. Q. Nguyen, D. Q. Nguyen, D. D. Pham, and S. B. Pham. "RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger". In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 17-20, 2014.
- [17] Nguyen, Dat Quoc, Dai Quoc Nguyen, Dang Duc Pham and Son Bao Pham 2016. "A robust transformation-based learning approach using ripple down rules for part-of-speech tagging". AI Communications, 29(3):409-422.
- [18] Wikipedia of Word Error Rate: https://en.wikipedia.org/wiki/Word_error_rate
- [19] (NIST) The National Institute of Standards and Technology. Speech recognition scoring toolkit (sctk), version: 2.4.10, 2015.
- [20] Chetifi Ei-Hadi and Guerti Mhania. "Phonetisaurus-based letter-to-sound transcription for Standard Arabic", In Proceeding of Conference: 2017 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B).
- [21] Scheffer, Tobias 1996. "Algebraic Foundation and Improved Methods of Induction of Ripple Down Rules". In pages 23-25.