Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 08, No. 04, December 2022.

COTA 2.0: AN AUTOMATIC CORRECTOR OF TUNISIAN ARABIC SOCIAL MEDIA TEXTS

Asma Mekki, Inès Zribi, Mariem Ellouze and Lamia Hadrich Belguith*

(Received: 17-Jun.-2022, Revised: 7-Sep.-2022 and 18-Oct.-2022, Accepted: 10-Nov.-2022)

ABSTRACT

In written text, orthographic noise is a common concern for NLP, especially when operating social-network comments and raw documents. This is mainly due to its orthographic conventions and morphological ambiguity. We propose to automatically normalize the social-media dialect corpora by following CODA-TA, the conventional Orthography for TA. The existing system developed for TA «COTA Orthography 1.0» is not able to handle all forms of TA. Therefore, we propose to extend its rules and lexicons to address the peculiarities of social media dialect. In certain words, the COTA Orthography 1.0 system provides the user with several correction possibilities. Therefore, in the new version, we incorporated a trigram language model to automatically select the right correction. Our results show that the system can reduce transcription errors by 95.72%.

KEYWORDS

Orthographic normalization, Tunisian Arabic, COTA Orthography system, CODA-TA.

1. INTRODUCTION

Dialectal Arabic is a linguistic variety that is historically related to classical Arabic and exists side-by-side with Modern Standard Arabic (MSA). In fact, MSA is the official written and spoken language used by the government, media and education. Dialectal Arabic is the spoken variety used in daily communication of the Arabic World and is not generally written [1]. Indeed, it has no standard orthographies.

Since the political Tunisian revolution in 2011, the Internet is taking an increasingly important role in Tunisians' lives with 7,447,000 Facebook users and 1,910,000 Instagram users in Tunisia in January 2019[1]. Generally, Tunisians use their dialect for expressing their opinions and emotions. Researchers have taken advantage of the high number of comments shared on social media, as well as the availability and ease of accessing these tools, to build large corpora for Tunisian Arabic [2]-[3]; [1]; [4]-[7].

Indeed, social-media dialect is characterized by its high level of orthographic heterogeneity, which made its processing a serious challenge for Natural Language Processing (NLP) tools. Despite the efforts of researchers to normalize the orthographic form of dialectal Arabic, most of the existing corpora are not standardized, where the same word is written in several forms (i.e., a word can have dozens of writing forms).

In this paper, we propose to automatically normalize social-media dialect corpora written with Arabic characters into CODA-TA Conventional Orthography for Tunisian Arabic (TA) [8]. We decided to use and expand CODA-TA, because there is already a semi-automatic tool (COTA Orthography [9]) that follows its linguistic guidelines. The process of orthographic normalization was made easier by this tool. Furthermore, many TA corpora have already been normalized using this convention. However, the existing system developed for TA [9] is not able to address all forms of TA (see sub-section 3.1). Therefore, we propose to extend its rules and lexicons in order to treat the particularities of social-media dialect. Hence, our contributions can be summarized as follows:

- We started by extending the CODA-TA spelling convention to include social-media dialect features.

[1] http://napoleoncat.com

Mekki, I. Zribi, M. Ellouze and L. H. Belguith are with ANLP Research Group, MIRACL, Uni. of Sfax. Emails: asma.elmekki.ec@gmail.com, ineszribi@gmail.com, mariem.ellouze@planet.tn and l.belguith@fsegs.rnu.tn

- The next step is to enrich the lexicon by the vocabulary used in social networks.
- We also proposed adding new patterns to treat onomatopoeia, accentuations of words, …etc.
- Thereafter, we trained an n-gram model on a large textual scale based on the three forms of dialect (intellectualized dialect, spontaneous dialect and social media dialect).
- Then, we integrated this model into the first version of the Conventionalized Tunisian Arabic Orthography (COTA Orthography system) to choose the right correction automatically.

We show in the evaluation section the effect of using an automatically normalized corpus on the diverse tools, such as sentence segmentation, POS tagger and parser. It ensures that the number of orthographic errors in a document decreases significantly, which is very helpful for NLP tools. Our system contributes to a significant improvement in this assessment. It can also ensure high quality without wasting time on manual orthographic normalization, because it is a fully automatic tool.

This paper is structured as follows: Section 2 is dedicated to review related work. Section 3 presents the Tunisian dialect forms as well as social-media dialect and COTA orthography automatic normalization challenges. Section 4 details our proposed method. Finally, we present and discuss, in Section 5, the evaluation results.

2. RELATED WORK

2.1 Orthographic Conventions

The orthographic normalization of Arabic dialects has been the subject of many studies, given its importance for many NLP tools. Indeed, the authors in [10] have proposed CODA (Conventional Orthography for Dialectal Arabic). The goal of this convention is to provide a set of rules standardizing the transcription of dialectal Arabic. The convention is based on the MSA spelling rules for their decisions. [10] defined CODA for the Egyptian dialect. Then, [8] made an extension of the spelling convention for TA. Also, [11] suggested an orthographic convention for Algerian dialect based on CODA. Many other extensions were proposed, such as [12] for the Palestinian dialect, [13] for Gulf Arabic and [14] for Moroccan and Yemeni Arabic. [15] proposed CODA* that presents a conventional orthography applicable on multiple Arabic dialects at the same time (i.e., from 28 Arab cities).

2.2 Orthographic Normalization Systems

In the literature, research on normalization of Arabic orthography can be classified based on the variety treated. Therefore, the remainder of this section shows the related work of both varieties MSA and dialectal Arabic.

2.2.1 Modern Standard Arabic

Several works exploited language modeling within the orthographic normalization of MSA. It is a technique based on contextual information in the decision. It uses the estimation of probabilities of sequences of n words (n-grams). [16] and [17] trained language models based on «n-gram» for error correction in inserting and deleting spaces. They also addressed error detection through two character-based trigram language models to classify words as valid and invalid.

[18] used a character-based 15-gram model to deal with merged word errors. In fact, authors in [18] statistically divided them by space, forming a network. In this network, they employed a heuristic evaluation, using an n-gram probability estimation, on each character to estimate the best path through it. Thus, the sequence of letters and spaces with the highest marginal probability, given by the language model, is selected.

[19] addressed the problem of automatically detecting real-word errors by using an n-gram ($n \in [1, 3]$) statistical language model and an SVM algorithm [20]. For the correction phase, the authors applied an n-gram language model to generate all error-word matches using Damerau-Levenshtein distance [21]. The test set is composed of 10K sentences from the KSU corpus[2] and artificially populates it with context errors using single edit distance and mixed-edit distance. The edit distance between two words

[2] http://ksucorpus.ksu.edu.sa/ar

is the minimum number of valid operations required to normalize the word (e.g. insertion, deletion or replacement of a single character). The overall F-measure value was 90.7%.

[22] built an Arabic error detection and correction system using a Bi-LSTM architecture. This classifier allows Boolean predictions rather than inferring error types. Therefore, the authors manually compiled a list of approximately 150 errors, including punctuation, spelling, morphological, syntax and named entity-recognition errors. For evaluation, they developed a corpus of 15M fully inflected Arabic words. The experimental results revealed an F-measure of 93.89%.

2.2.2 Dialectal Arabic

[22] proposed a system able to transform spontaneous orthography of the Egyptian dialect into the conventionalized form CODA. The authors start with a pre-processing step that eliminates letters repeated more than twice. For normalizing Egyptian dialect, [23] proposed two techniques: contextual and non-contextual. The first technique builds a unigram model that replaces every word in the spontaneous orthography with its most likely CODA form as seen in the training data based on the word level. In the second technique, a set of transformations is applied on a character level using the k-nearest neighbor algorithm (k-NN) [24]. It does not depend on the character context inside the word. In addition to the techniques discussed above, the authors have used a morphological tagger [25]. The best results come with the combination of the cited approaches with 68.1% of error reduction.

[9] proposed a method similar to that proposed for Egyptian dialect. The authors have proposed a hybrid approach to normalize the spelling of the spontaneous Tunisian Arabic (TA) based on the spelling convention CODA-TA [8]. The first method using k-NN supervised algorithm corrects the attached proclitic with generally several types of errors for the same word. Then, the linguistic method is based on pre-defined patterns and a specific lexicon for each error form.

[26] created a dataset that consists of 185K Algerian texts. The authors began by automatically pre-processing the corpus by eliminating punctuation, emoticons and reducing the number of recurring letters to not more than two. Then, the dataset was manually normalized by experts. The parallel corpus contains 50,456 words and 26,199 unique words to be normalized. [26] introduced two deep-learning models for this task, with the CNN model achieving the best evaluation result with an overall F-score of 64.74%.

Despite the richness and relevance of research, we must point out that the only orthographic normalization tool developed for the TA does not support social-media dialect, which is the most widely available and easiest to collect dialect. Furthermore, COTA orthography [9] remains semi-automatic, requiring user intervention to normalize certain words.

3. TUNISIAN ARABIC

In this section, we discuss the characteristics of Tunisian Arabic (TA), its different forms as well as the orthographic errors that can be found in TA writings.

3.1 Brief Presentation

Tunisian Arabic (TA) is a North African dialect of Arabic that represents the native language spoken in Tunisia by almost 12 million people [27]. It differs from the Modern Standard Arabic (MSA) in different levels [28]; [11]: morphology, syntax, pronunciation and vocabulary. Its lexicon contains several words from different languages, such as Maltese, Berber, French, English, …etc. [28]; [8]. TA is classified into three different forms [27]; [3]: intellectualized dialect, spontaneous dialect and social media dialect according to the specificities of each one.

Intellectualized dialect [3] is mainly used by intellectuals. This form is a mixture of MSA and TA with a relatively high frequency of MSA words. Its syntactic structure is the closest to the MSA, which makes it the most regular form.

Spontaneous dialect is the form of communication dialect that contains the highest mass of TA words with its co-existence of multiple languages, such as Maltese, MSA, Italian and mostly French. It is characterized by the presence of disfluencies (e.g., incomplete words, repetition, filled pause, stuttering, …etc.). Several papers proposed transcribed corpora from several audio sources, such as

[1]-[2]; [29].

Textual content of social networks represents a combination between the two forms previously cited. However, content in social media generally contains more orthographic errors than the other forms of dialect (intellectualized dialect and spontaneous dialect). This is obvious, since, at the time when tens contribute to the writing of intellectualized-dialect and spontaneous-dialect corpora, each internet user contributed with a very limited number of comments in the corpus. Moreover, each time the number of writers increases, the heterogeneity of the written words increases. In addition, we notice the presence of non-standard abbreviations, onomatopoeia, emoji, accentuation, …etc. Social-media dialect is divided into two parts according to the alphabet character used whether Arabic or Latin ≪Arabizi≫. It is a term used to describe an encoding system that uses the Latin script and substitutes some Arabic letters with Arabic numbers instead. The Arabic numbers fill in for Arabic phonemes that are absent in the Latin language, but resemble Arabic letters and their forms, where each letter represents an Arabic phoneme that corresponds to it in pronunciation [30]. For example, the number 3 stands for the Arabic character (ع, E), the number 7 comes for (ح, H), …etc. In this paper, we only consider the correction of text with Arabic letters.

Table 1 shows some sentences for each form of dialect with their English translation and Arabic transliteration [31][3].

Table 1. Examples of sentences of the three forms of TA.

Sentence	Translation	Script	Form of TA
الحق في القراف مضمون AlHq fy AlqrAf mDmwn	*The right to strike is guaranteed*	Arabic	Intellectualized dialect
امم أنا étudiante في # Amm AnA étudiante fy #	*Amm I am a student in #*	Arabic & Latin	Spontaneous dialect
Bjr Hmd chna7welek enti	*Hello, it's OK; what's up?*	Latin	Social-media dialect
واو جو كبيير☺ wAw jwkbyyyr ☺	*Waw a lot of fun ☺*	Arabic	

3.2 Tunisian Orthographic Errors

Most of the textual resources available are not standardized. Therefore, words can be presented in several forms, which greatly increases the error rate of any NLP applications. Table 2 presents an example of the word (ثمة, there is) and some of its different writing forms in TA. For all of the provided mistake instances, we rely on the CODA-TA convention's rules [8].

Table 2. Examples of spelling errors in the study corpus.

CODA-TA spelling	Corpus	Transliteration
ثمة *vmp*	فمة	fmp
	ثمّة	vmp
	ثما	vmA
	فما	fmA
	ثمّا	vm A
	فمّا	fm A
	ثم	vm
	فم	fm
	ثاما	vAmA
	فاما	fAmA
	ثامت	vAmt

[9] detected and presented several types of errors for TA. Some of them are shared with MSA, such as writing errors of some letters (ى, Y; ا, A; ي, y, ة, p and هـ, h) and the presence of space between the coordination conjunction (و, w) and the following word, …etc. According to the CODA-TA orthographic convention, they also specified TA specific errors:

[3] We follow the Arabic transliteration convention: http://www.qamus.org/transliteration.htm

- **Space after the negation form:** according to the CODA-TA orthographic convention, they also specified TA specific errors due to the absence of space between the negation form (ما, mA) or (م, m) and the following verb, etymologically spelled, the neglecting of the silent Alif at the end of the third-person plural affix.
- **The attached proclitics:** the TA marks a set of proclitics, such as (هـ, h; ك, k; م, m; ع, E; ف , f) which must be attached to the following name. For example, the word (هلكرسي, hlkrsy) must be written according to CODA-TA as (هلكرسي, hAlkrsy, this chair).
- **Plural Waw:** the affix (وا, wA) is used to express the third-person plural, but the character (ا, A) written at the end of the word is often overlooked. For example, the normalized verb (خرجوا , xrjw, they came out) is often written as خرجو, xrjw).

For social-media dialect, we found other types of errors. Among the errors, we can cite:

- **Accentuation of words.** This phenomenon represents the repetition of a letter several times successively (e.g علاااش , ElAAA$, why). Sometimes, people intend these repetitions to show affirmation or intensification.
- **Interjection.** An interjection is a term that is grammatically independent of the rest of the sentence. It mainly expresses a short and sudden expression of emotion rather than meaning. Internet users write interjections with multiple forms, such as واااو, wAAAw instead of واو,wAw (i.e., they do not use the same number of characters).
- **Onomatopoeia.** To imitate or resemble the sound of an animal, objects or human sounds, Internet users write onomatopoeia with multiple forms. For example, they do not use the same number of characters while writing laughing sound.
- **Tatweel.** Arabic scripts present horizontal strokes. In contrast to white space that creates justification by expanding spaces between words, Tatweel increases the length of a text by elongating characters at certain points (e.g. بـاهـــي, b_Ah_y Good).

To summarize, several types of orthographic errors can be detected in the three forms of TA. These mistakes show the necessity for a spelling normalization tool to be implemented. In the following section, we'll go over this in more depth.

4. NORMALIZATION OF TUNISIAN SOCIAL MEDIA DIALECT

COTA orthography system [9] (Conventionalized Tunisian Arabic orthography) is the only system that semi- automatically corrects Tunisian Arabic (TA) spelling errors. What we propose in this paper is an extension of this system and the orthographic convention CODA-TA [8]. Our goal is to automate the task of standardizing the spelling of the social-media dialect (sub-section 3.1). Figure 1 illustrates the steps of the proposed method.

4.1 CODA-TA Extension

CODA-TA [8] is a convention primarily based on MSA spelling rules (see Table 3). It provides an extension of the Arabic dialect orthographic convention CODA (Conventional Orthography for Dialectal Arabic) [32]. Both CODA-TA and CODA* are based on the convention proposed by [32]. They both share the same objectives and guiding principles. There are a few small variations between the two standards, such as the use of (برشة, br$p, a lot) in CODA-TA *versus* (برشا, br$A) in CODA*. Taking the example of numbers, both conventions add the character (ن, n) at the end of the word. However, CODA* adds the letter (ع, E) to numbers, such as (ثمانتعشن, vmAntE$n, eighteen) despite the fact that Tunisians do not pronounce this character. The same example is written (ثمنطاشن, vmnTA$n) according to CODA-TA. However, the main distinction is that CODA* is useful in multi-dialect processing cases, while CODA-TA was designed specifically for the Tunisian dialect. There is already a semi-automatic tool (COTA Orthography) that follows its linguistic guidelines. Furthermore, many TA corpora have already been normalized using this convention.

The CODA-TA extension rules are described in the remaining paragraphs of this sub-section, along with several examples that help to make them clear.

- This convention is mainly based on the consonants and vowels of the Arabic language. For example, non-Arabic phonemes (/V/, /G/ and /P/) are converted into (ف for /f/, ق for /q/ and ب for /b/). Indeed, they generally keep the same MSA spelling rules to choose the right form to use.

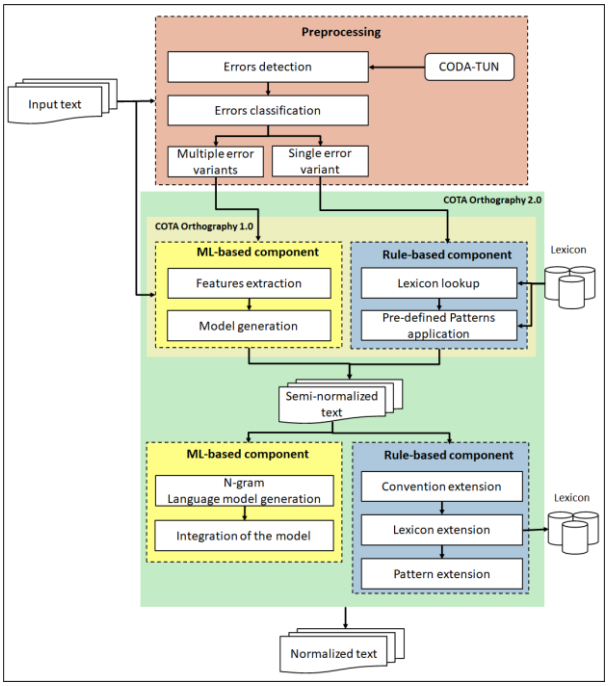

Figure 1. COTA orthography architecture.

Table 3. Most frequent CODA-TA examples.

Category	Tunisian	CODA-TA	Translation
Letter	ڤ /G/	ق/q/	
	ڤـ /V/	ف/f/	
	پ /P/	ب/p/	
Word	يقولو , yqwlw	يقول له, yqwl lh	*he tells him*
	يقول لو , yqwl lw		
	يقوله , yqwlh		
	يقل لو , yql lw		
	يقلو, yqlw		
Number	اسبعطاش باب, AsbETA$ bAb	سبعطاشن باب , sbETA$n bAb	*seventeen doors*
	سبعطاش باب , sbETA$ bAb		
Enclitic	ع , E	ع , E	*On*
	م , m	م , m	*From*
	خ , x	خ , x	*Let*
Proclitic	ش , $	ش , $	*Not*
	شي, $y	شي , $y	*Question*
	ش , $		

- Long vowels are written in a long form like in MSA. For example, the word يقول له, yqwl lh (he tells him) can be written in TA as (يقوله, yqwlh).
- The letter (ن, n) is added after some numerical structures in CODA-TA (e.g. سبعطاشن كرهبة, sbETA$n krhbp, seventeen cars).

- TA shares most of the attached clitics with MSA, such as definite article (ال, Al), coordinating conjunction (و, w), …etc. Other attached clitics are specific to TA such as interrogation (شي, $y) and negation (ش, $) enclitics, proclitics (ع, E). Thus, the verb (خلي, xly, let) is sometimes used as a prefix and attached to the following word (e.g. خنمشيوا, xnm$ywA, let's go).

This spelling convention is more practical for word-processing corpora, especially when it comes to a method of adapting the system from MSA to TA. However, it does not take into account the phenomena of social-media dialect. For example, the sentence extracted from social-media comments presented in Table 5 as «raw text». We notice that three words of the sentence have not been corrected (مممم , mmmm), (باارشا, bAAAr$A) and (تعب, t_E_b) (see «before extension» in Table 5). Therefore, we propose a set of rules that will be the subject of an extension of the orthographic convention CODA-TA. Table 4 lists examples for the different proposed patterns.

- **Accentuation of words** - All repeated characters in a word will be eliminated and we keep only one character of them (e.g. علاش, ElA$ instead of علااش, ElAAA$).
- **Interjections and onomatopoeia** - We unified all interjections and onomatopoeia written by a single repeated character into three characters. If they are composed of more than one character, where each character or a set is repeated sequentially, we keep only two appearances of each letter. Table 4 presents three examples of orthographic normalization of onomatopoeia.
- **Tatweel** - We proceed to eliminate all instances of Tatweel from words. For example, (باهي, b_Ah_y Good) turns into باهي.

After adding these rules to CODA-TA spelling convention, the comment presented above becomes correctly normalized as shown in Table 5.

4.2 First Version of COTA Orthography System

4.2.1 Method Overview

In order to normalize the spelling of Tunisian spontaneous dialect, Boujelbane et al. started by spotting transcription mistakes using CODA-TA convention. Indeed, they established two categories of errors. The first characterizes errors having several variants for the same word, such as (هالناس, hAlnAs, *these people*) which has several variants, such as ها الناس, hA AlnAs; هلناس, hlnAs; هل الناس, hl AlnAs, …etc. The second category includes errors having only one variant, such as the word (ثقافة, vkAfp, *culture*) that can be written (ثقافه, vqAfh). COTA orthography is a correction system based on the characteristics of each type. The machine learning-based component was defined to correct the first category of errors. It is inspired by [23]. Indeed, the same list of classes was used. The k- NN correction model was created after establishing the feature set. It determines whether a character should be replaced, removed or added to another character. The linguistic method is mainly composed of two techniques: the application of a set of pre-defined patterns and the lexicon lookup. For the first technique, standardization patterns were assigned for each form of word agglutination. For the lexicon lookup, [9] proposed six sub-lexicons which include 6,063 words for two sub-lexicons of similar spelling errors, 1,632 for etymologically spelled consonants, 1,066 words for the sub-lexicon of the third-person singular pronoun, 111 CODA-TA word list and 5,674 for waw of plurality. [9] tested COTA orthography system with a part of the non-normalized version of STAC [1] that contains 10,236 words (2,640 wrong words). The accuracy result achieved was 86.6%.

Table 4. List of new CODA-TA normalization examples.

Spelling error	Error	CODA-TA	Translation
Accentuation	brrr$p **بررشة**	br$p **برشة**	*a lot*
Onomatopoeia	hhhh **هههه**	hhh **هههه**	
	hhxxx **هههخخ**	hhxx **هههخ**	
	hEhEhE **هههههع**	hEhE **هههع**	
Tatweel	t_wn_s **تـونـس**	twns **تونس**	*Tunisia*
Attached clitics	Al mdrsp **المدرسة**	Almdrsp **المدرسة**	*the school*

Expression containing the name «Allah»	n$AAllh نشاالله	An $A All h ان شا الله	*Allah willing*
Word Each	kAlwAHd كلواحد	kl wAHd كل واحد	*Each one*

Table 5. Example of TA comment normalized according to CODA-TA before and after the extension.

	Example	**Translation**
Raw text	ممم الخدمه فيها باالرشا تعب ربي يعينو mmmm Alxdmh fyhA bAAAr$A t_E_b rby yEynw	*Mmm this work is very tiring, may God help him*
Before extension	ممم الخدمة فيها باالرشا تعب ربي يعينه mmmm Alxdmp fyhA bAAAr$A t_E_b rby yEynh	
After extension	ممم الخدمة فيها برشة تعب ربي يعينه mmm Alxdmp fyhA bAr$p tEb rby yEynh	

4.2.2 COTA Orthography System Errors

Although the COTA orthography system achieves an encouraging result for the Tunisian spontaneous dialect, we notice several failure cases that were not considered during the system's implementation.

The social-media dialect represents a valuable source of data for researchers. Despite this, the system is unable to detect and correct a variety of errors. For example, the different types of onomatopoeia (e.g. ههه, hhh), accentuation (e.g. كبييير, kbyyyr, big) and Tatweel signs (e.g. نـــوح, n_wH, Noah) are not taken into account and the system does not correct them. Moreover, social-media dialect corpora contain words with the character (ا, A) added at the beginning of some words, such as the verb (دخلتوا, dxltwA, you got in) that can be found as (ادخلتوا, AdxltwA). Furthermore, some of the suggested patterns only called for a specific grammatical category. However, several words, generally used in social-media dialect, cannot be detected and corrected by the system (e.g. لايكيو, lAykyw, like) is an incorrectly written word which is generally used in social-media dialect corpora.

Sometimes, Internet users forget to add a space between the comment's words, which implies two or more words attached. Even native speakers of the dialect are often unable to read sentences that do not contain any spaces to delimit the words (e.g. the two attached words تصيرساعات, tSyrsAEAt which means it happens sometimes). COTA orthography system does not take this type of error into account.

The term (كل, kl, each) is frequently used as quantity noun in TA. Nevertheless, COTA orthography normalizes the different types of errors in the (ك, k) enclitic, which may result in the modification of expressions containing this quantity nouns, creating text distortion. For instance, the rate of comments involving this form of error did not exceed 19% in our study corpus (80% of TAD [6]).

Another form of error caused by [9]'s system arises when changing words containing consonants with multiple pronunciations. For example, the verb (صمن, Smn, solidified) becomes (سمن, smn, gained weight). However, Tunisians use both terms. As a result, we cannot judge whether the Internet user is indicating « solidified » or « has gained weight ».

Moreover, COTA orthography system [9] grants several alternatives for some words separated by «/». These terms require more than a correction regardless of the subject. In other words, each option can be valid in a given context and wrong otherwise. For instance, words ending with (و, w) may mean the third-person singular pronoun (هـ, h) or the affix of the third-person plural (وا, wA). It depends on the context of the word in the sentence. Take the example of the word (فهمتو, fhmtw) in TA, which does not follow any spelling convention. When it is considered as the third-person singular pronoun, the word is corrected by: (فهمته, fhmth), which means (I explained to him) or (I get it). However, if it is considered as the affix of the third-person plural, the word will be corrected by (فهمتوا, fhmtwA, you understood). Therefore, COTA orthography system gives the two writing

choices separated by a slash «/» which requires a manual decision between the two propositions depending on the context.

4.3 Second Version of COTA Orthography System

In this sub-section, we detail our proposed method for the orthographic normalization of TA errors. We focus on texts from social-media dialect. We start by extending the CODA-TA spelling convention [8]. Subsequently, we propose to automatically extract a lexicon from the Tunisian Treebank «TTB» [33] and a list of predefined patterns. Thus, we create a language model to process multiple-choice words.

4.3.1 Linguistic Techniques' Extensions

To extend the linguistic techniques, we have used three social-media dialect corpora to fix orthographic errors. The first one is TAD (Tunisian Arabic Dialect) [6]. They extracted 73,024 messages of which over 72% are in Latin letters. These messages went through three steps of processing: spam filtering, message division (Arabic or Latin characters) and message classification (dialectal or non-dialectal). TAD is composed of 7,145 messages (151,598 words) written in Arabic letters. The messages are collected from Facebook comments, messages from mobile phones, …etc. The corpus contains only dialectal texts [6]. This dataset is available by email request to the first author.

The second corpus is of Masmoudi et al.'s corpus [34]. It is a collection of 21,917 words extracted from Tunisian blogs treating various fields (politics, sports, culture, science, …etc.). Two experts who are native Tunisian Arabic (TA) speakers have validated the comments as TA. They manually translated 3,500 Arabizi words (530 sentences) into Arabic script. We get access to this corpus by emailing the first author.

The TSAC[4] (Tunisian Sentiment Analysis Corpus) [5] contains 17,000 comments that are classified to positive (63,874 words) and negative (49,322 words) polarities. This data was collected from Facebook comments that are written on the official pages of Tunisian radio and television channels (Mosaïque FM, JawhraFM, HiwarElttounsi TV, …etc.). The version proposed in GitHub is licensed under the GNU-v3.0.

We used 80% of TAD corpus [6] for the extraction of patterns and the enrichment of sub-lexicons. The remaining part was used for the test. Furthermore, we randomly selected 43,247 words from TSAC [5] and Masmoudi et al.'s corpus [34]. Indeed, two native speakers manually normalized them according to CODA-TA [8]. We calculated the inter-annotator agreement to measure how well our two experts can make the same normalization. The obtained kappa value is 0.896, indicating a high degree of concordance. Table 6 presents the size of the corpus and its error rate according to CODA-TA [8].

Table 6. Details about the test set for system evaluation.

Corpus	Number of words	Error rate
TAD (6)	22,740	22.62%
Corpus of (34)	10,371	20.16%
TSAC (5)	10,136	24.85%
Total	**43,247**	**22.55%**

4.3.1.2 Extension of the Lexicon: [9] collected a set of six sub-lexicons for each error form. It can help in detecting spelling errors (i.e., if the word is not recognized by the system, it will not be corrected). Moreover, for the correction phase, two contributions are possible: the parallel sub-lexicon containing the incorrect word and its normalized equivalent can be used for substituting the wrong detected form by the correct one. Also, it is also used to call certain patterns.

Based on our study corpus, we semi-automatically enriched these lexicons with new words. For example, we added several verbs, such as (برتاجي, brtAjy, share), (عدّل, Edˇl, adjust), …etc. Thus, we

[4] https://github.com/fbougares/TSAC

noticed that many Internet users frequently add the character (ا, A) to the beginning of words. Therefore, we proposed to add a new sub-lexicon to correct this type of error. This lexicon is collected from various sources (STAC corpus [1], the Tunisian constitution [7], Boujelbane's corpus [35] and Younes's corpus [6]).

Table 7 details the size of the sub-lexicons before and after the update. We extracted 767 verbs, 118 nouns and 44 pronouns automatically from TTB [33]. Then, all the words were validated by native speakers.

Table 7. The new size of sub-lexicons.

Lexicon	**Initial size**	**Final size**
List of verbs	5,435	6,202
List of nouns	914	1,032
The third-person singular pronoun	1,066	1,110
Etymologically spelled consonants	1,632	1,514
List of words that start with *A*	-	7,219

Otherwise, the TA has many consonants with multiple pronunciations. The letter س s can be pronounced as ص S, which inducts multiple spellings. Therefore, [9] built two sub-lexicons for these consonants. For example, the first sub-lexicon is composed of a list of words containing the consonant س and their equivalents in the incorrect writing. By studying these sub-lexicons, we found a set of polysemous examples (i.e., the word proposed as a mistake is correct in another context). For example, the word (سورة, swrp, Surah) can be written incorrectly into (صورة, Swrp, picture) but most likely the writer means Surah[5]. Therefore, to solve this problem in this stage, we propose to eliminate these words from the sub-lexicons. We removed 75 words from the sub-lexicon ص to س, where the total number of words becomes 901 and 43 words from the sub-lexicon that transform the letter س to ص with a new total of 393 words.

4.3.1.2 Normalization Patterns: To improve the performance of COTA orthography system, we propose a set of manually implemented patterns to correct social-media dialect spelling errors (presented in Section 3) and to correct others TA errors not covered by [9] (see Table 8).

Attached Clitics: [9] generated a set of models that are able to correct errors related to attached clitics. These models are not able to correct spelling mistakes of the following clitics: الـ, Al; للـ, ll; لـ, l; and ب , b. Therefore, we defined a pattern that deletes the space between one of these clitics and the following word (see pattern 1 in Table 8).

Expressions Containing the Name «Allah»: Each person writes the expressions containing the name God in his own way (e.g. (الحمدلله, AlHmdll h) turns into (الحمد لله, AlHmd l lh, Thank God)). Thus, we created a pattern that detects wrong expressions and corrects them according to the convention CODA-TA (see pattern 2 in Table 8).

Word كل kl: The numerical approach proposed by [9] deals automatically with quantity nouns كل, kl as an error of writing (i.e., as the attached clitic ك, k). In the study corpus, we remark that the «k model» of [9] increases the error rate by 83.18%. Therefore, we developed a pattern that covers and avoids all changes of the quantity nouns (كل, kl, each) (see pattern 3 in Table 8).

Table 8. Examples of patterns.

Number	**Pattern**
1 - Attached clitics	**IF** len(word) == 1 **AND** word **IS** valid_clitic **THEN** Remove space after word
2 - Expression containing the name "Allah"	**IF** "Allah" **IN** word **THEN** Replace word with normalized form

[5] The Quran is divided into Surahs (chapters). Source: https://en.wikipedia.org/wiki/List_of_chapters_in_the_Quran

3 - Word كل kl	**IF** word **IS** quantity_nouns **THEN** Do nothing
4 - Onomatopoeia 1	**IF** word **CONTAINS_ONLY** letter **AND** len(word) > 3 **THEN** Remove extra letters
5 - Onomatopoeia 2	**IF** word **CONTAINS_ONLY** (letter_1 **AND** letter_2) **AND** len(word) > 4 **THEN** Remove extra letters

Table 9. List of new normalization pattern examples.

Pattern	**Before normalization**	**After normalization**	**Translation**
Accentuation	mzyAAAn مزياان	mzyAn مزيان	*beautiful*
Interjection	Ammm اممم	Amm امم	
Nomatopoeia	hhhAA هههاا	hhAA هههاا	
	Hyhyhy هيهيهي	hyhy هيهي	
Tatweel	x_wy_A خـــويـــا	xwyA خويا	*my brother*
Attached clitics	Al klyp ال كلية	Alklyp الكلية	*the university*
Expression containing the name «Allah»	m$Allh مشالله	mA $A Allh ما شا الله	*machallah*
Word each	kAlEbd كالعبد	kl Ebd كل عبد	*each person*

Interjections, Onomatopoeia and Accentuation: Interjections, onomatopoeia and accentuation are often used in social media dialect. We developed a set of patterns that detect these terms and correct them by removing repeated characters (see patterns 4 and 5 in Table 8).

Tatweel: We proposed a pattern that eliminates all Tatweel forms in the input. Table 9 shows some examples of the new normalization patterns.

4.3.2 Language Model

COTA orthography system provides a semi-automatic normalization for terms that imply more than one correction, independently of the context. Therefore, we introduce in this sub-section our method for automating this task. This method relies on the comparison of two language models to fix errors semi-processed by the system while taking advantage of the textual resources already created in favor of TA [3]; [1]; [34]; [5]-[6]; [36] and MSA [37]. Table 10 presents the size of each corpus.

We started with the first step of preparing the collected textual resources that consists in checking the normalization of the TA corpus using the COTA orthography system [9] and manually selecting the correct option from the choices given by COTA orthography system. Our result corpus consists of 7,571 multi-choice words. We have used diverse datasets from different fields and topics to generate the language models. The total size of the corpus in TA is 379,063 words.

Table 10. Size of corpora used for language-model generation.

Corpus	**Size**
[3]'s corpus	37,964
STAC [1]	42,388
TAD [6]	151,598
TSAC [5]	113,196
Normalized Tunisian constitution [36]	12,000
[34]'s corpus	21,917
KACST corpus [37]	2,207,469
Total	**2,586,532**

- The first corpus [3] is a transcription of 5 hours and 20 minutes of recordings mainly from a Tunisian television channel. This corpus contains 37,964 words, where 12,207 words come from a TV news program and 25,757 words from programs of political debates. This corpus is available by email request to the first author.
- The second is the STAC[6] (Spoken Tunisian Arabic Corpus) [1] containing 42,388 words (4 hours and 50 minutes of recordings). It is a transcription and annotation of spontaneous TA spoken in various TV and radio channels. It has 97.20% words in TA, 0.37 % in MSA and 2.43 % in French. STAC includes disfluencies. It is licensed under the GNU-v3.0.
- TA constitution [7] is an intellectualized dialect that consists of 12,000 words, normalized by [36]. It is available by emailing the first author.
- TAD, TSAC and Masmoudi et al. datasets [6]; [5]; [34] (see description in sub-section 4.3.1).
- KACST (King Abdulaziz City for Science and Technology) (37) MSA corpus is made up of 2,207,469 words, carefully sampled and its content is classified according to different parameters, such as time, country, field, subject, etc. KACST is licensed under the GNU-v3.0.

We have divided the corpus into training, development and test sets (80/10/10, respectively) according to the number of multiple-choice words.

N-gram-based Language Model

N-gram models are among the most commonly used language models of spell checking, due to their flexibility and utility. In this type of model, the probability of a word is calculated as a function of its history (the previous n-1 words). These probabilities are determined depending on the count of each sequence detected. The n-gram model was trained on our learning corpus. To obtain an efficient language model, we have configured the basic model generated using the development corpus by suggesting all the possible hypotheses (i.e., offered by the COTA orthography system) for a given sentence. Then, the model assigns a perplexity value to each proposition. As a result, the sentence admitting the lowest perplexity is held as correct.

In the following paragraphs, we describe the process of setting up our model:

- **Fixation of the n-gram.** We trained 6 models with different n-grams to select the most adequate n-grams. The best generated model reached 65.46% using the trigram model.
- **Adjustment of the learning corpus.** We automatically checked the normalization training data using COTA orthography system. Therefore, the training set contains orthographic errors that increase the perplexity of the model. Thus, the elimination of repetitive sentences resulted in 5.72% improvement with an accuracy equal to 71.18%. In the literature, orthographic normalization works using language models are based on large corpora [17]; [16]; [18]. We thus tried to extend the size of our training corpus. Due to the lack of TA textual resources, we decided to add an MSA corpus to our textual base. Indeed, COTA Orthography system relies on MSA spelling rules to correct errors [9]. In addition, MSA supports universally known spelling rules. Therefore, its corpora often do not contain spelling mistakes (i.e., especially written by journalists - case of KACST corpus). Adding KACST [37] to the TA corpus without duplications, raised the accuracy to 72.34%.
- **Choice of options**. Among the available options, -unk was the unique alternative to mark an improvement in the language model. Using the default configuration, Out-Of-Vocabulary (OOV) words are deleted. When using this option, the language model keeps unknown words and treats them as normal words (not OOV). As a matter of fact, it allows keeping the vocabulary open. This addition marked an increase in the accuracy by 9% to reach 81.34%.

LSTM-based Language Model

GluonNLP [38] is a natural-language processing deep learning-based toolkit. Several models have been supplied by this toolkit for natural-language processing tasks, such as word embedding, language

[6] https://sites.google.com/site/ineszribi/ressources/corpus

modeling, machine translation, etc. In this paper, we used GluonNLP to implement a typical LSTM language model architecture. Then, we trained the language model on our training dataset (see Table 10). Several experiments were carried out to improve the LSTM- based language model using the development corpus. We chose the best alternative according to the perplexity result.

Grid search was applied to fine-tune the parameters of the LSTM-based language model. The optimal configuration is based on a batch size of 64, Adam optimizer and Softmax function. The number of epochs is set to 100. At this stage, we get an accuracy of 80.17% using the development set. As for the N-gram-based language model, we tested the effect of adding the MSA corpus to the training set, which improved the result by 1.01%. The best obtained language model reached an accuracy of 81.18%.

We can conclude that the N-gram technique is just 0.16% better. Therefore, we conducted non-parametric tests on the dataset. The p-value for the non-parametric independent Wilcoxon test [39] is 0.031. Since the p-value is less than the threshold of 0.05, we can conclude that the values are statically significant.

5. EVALUATION RESULTS

In this part, we provide the language model's experimental results as well as the new version of the COTA orthography system (Conventionalized Tunisian Arabic orthography) while describing the qualitative analysis of these results.

5.1 Experimental Results

5.1.1 Language Model Evaluation

We tested our final trigram model using the test corpus (10% of the collected corpus). By entering the input text, COTA orthography system begins by applying the linguistic techniques. This processing gives us a semi-automatic result. The language model is then used to select the best alternative to keep in sentences with several choices. For example, by entering a sentence admitting two choices, we enter two alternatives to the model (i.e., the first sentence contains the first option, whereas the second sentence contains the second option). The model grants perplexity to each sentence and we keep the one with the lowest value of perplexity. This process is fully automatic. Afterwords, we created a reference version of the test corpus to validate the model's output. The accuracy result obtained using our language model is equal to 79.38%.

5.1.2 COTA Orthography System 2.0 Evaluation

In this sub-section, we seek to examine how the complementary patterns and sub-lexicons of social-media dialect can generate additional gains in Tunisian Arabic (TA) automatic normalization.

We present in Table 11 the errors that we treated with their accuracy, their frequency and their percentage in the erroneous part of the test corpus (i.e., the overall results are presented in Table 12). The best accuracy value found achieves 100% for the 6 patterns of interjections, onomatopoeia, word JS, attached clitics, expression containing the name «Allah» and «Tatweel». However, words starting with the character A give the lowest value with 68%.

For the consonants, we detected an accuracy of 72.73% by testing the corpus with the basic system. However, filtering the sub-lexicon increases the results by 50%. Similarly, all interjections, onomatopoeia, accentuation of words or Tatweel errors have not been corrected with the system of [9].

Table 11. System performance for each spelling error with the test corpus.

Orthographic errors	**Frequency**	**Percentage**	**Accuracy**
Accentuation	430	8.4%	96.98%
Interjections and onomatopoeia	329	8.4%	100%
Nouns list	212	4.12%	88.68%
Verbs list	172	3.34%	80.81%
Words starting with *A*	94	1.83%	67.86%

Word *kl*	89	1.73%	100%
The third-person singular pronoun	59	1.15%	86.44%
Clitics attached	57	1.1%	100%
Expression of the name of «Allah»	36	0.7%	100%
Tatweel	32	0.62%	100%
Consonants	22	0.43%	72.73%
Total	**1,532**	**31.82%**	**90.32%**

Table 12. Evaluation 1: Results of the normalization system based on the TAD corpus.

Measurement	**COTA 1.0**	**COTA 2.0**	**Improvement**
Number of wrong words	5,143		-
Number of properly corrected words	3,399	5,031	1,632
Number of uncorrected words	1,927	136	1,791
Recall	66.09%	97.28%	+ 31.19%
Precision	63.82%	94.2%	+ 30.38%
F-measure	64.94%	95.72%	+ 30.78%

For the evaluation of our orthographic normalization system, we calculate the measures of recall, precision and F-measure based on the number of properly corrected words.

We tested the new version of the system with the same corpus of spontaneous dialect used for the test [1]. The result showed a 3% improvement over the old version of the system. We obtained 91% of recall, 89% of precision and 90% of F-measure.

We evaluated the system with the corpus of [6] (test part). Our results are presented in Table 12. All of these results are significantly better than those of the old version of COTA orthography system. We conducted additional evaluations with alternative corpora ([34] and [5]). These results are shown in Table 13.

Table 13. Evaluation 2: Results of the normalization system based on Masmoudi et al.'s and TSAC corpora.

Measurement	**Masmoudi et al.**	**TSAC**
Number of wrong words	2,091	2,519
Number of properly corrected words	1,933	1,991
Number of uncorrected words	249	617
Recall	92.44%	79.04%
Precision	88.59%	76.34%
F-measure	90.47%	77.67%

5.2 Discussion

In the following part, we discuss the results achieved for the trigram-language model as well as version 2.0 of COTA orthography.

5.2.1 Language Model Evaluation

The language model manages to correctly choose the right suggestion. For example, the sentence (يعينوا/ الله يعينه, All h yEynh/yEynwA) is corrected as follows: (الله يعينه, All h yEynh, may God help him). However, the following sentence: (علاش نشاركو, ElA$ n$Arkw, Why are we participating?) is poorly normalized (علاش نشاركه, ElA$ n$Arkh). In fact, if the wrong option is made, it is mainly for two reasons. First, the correct choice is not included in the corpus. Second, the probability of appearance of the incorrect choice in the learning corpus is higher than that of the correct choice. Moreover, we notice that sometimes the candidates for standardization get the same perplexity. In fact, this is due to the balance of the probabilities of appearance of the two choices in the learning corpus. The test set includes 6.25% of these sentences. Thus, we set the first option as default.

To reduce the invalid selection of the alternatives, we need to add more of the sentences with

demonstrative pronouns (i.e., feminine (هاكي, hAky) and masculine (هاكه, hAkh)) as well as words ending with (هـ, h) and (وا, wA) in order to select more precisely the most appropriate alternative for our context.

5.2.2 COTA Orthography System 2.0 Evaluation

In general, the results are encouraging. However, among the errors we have detected, we can cite the example of word خدمتو that is corrected as (خدمته , xdmt, his job). The modification of this word can be true in a defined context, but false in another, depending on the grammatical category of the word. In fact, the word خدمته can indicate the verb «they worked» that should turn into (خدمتوا , xdmtwA). It can also represent the term «his job» which is correctly written.

The error analysis shows that some words are not corrected or wrongly modified due to our lexicon that does not cover some words. For example, the system eliminates the character I from the beginning of the word (ازين, Azyn), while the correct form is (يزين, yzyn, decorate).

The first evaluation using the TAD test corpus shows the best F-measure result with 95.72%. Our system has been improved by over 30% compared to the COTA orthography system. The results obtained are encouraging. Thus, for the second evaluation, the results are higher than 90%. Clearly, the construction of extra-patterns improved the system performance with additional and higher quality sub-lexicons.

Up to the third evaluation, we tested the system with more difficult cases. TSAC corpus [5] is a corpus dedicated to the analysis of feelings. It is misspelled, which explains that it gives the highest error rate among the three corpora. We shall note that the performance is significantly lower than the corresponding results of the two other corpora, which explains the degradation of the value of F-measure by more than 15%. Therefore, we tried to analyze the failure cases to understand their causes.

First, the most common mistake we have encountered comes from the attached words. It represents more than 20% of uncorrected errors in the corpus TSAC. Internet users can forget to type the space between words. We even found a 100-character sentence without any space to delimit the words. This sentence is not legible even for native speakers of TA. Thus, we can catch other errors, such as missing letters (e.g. الاحترا, AlAHtrA instead of الاحترام, AlAHtrAm, respect), added letters (e.g. تونسيلة, twnsylp instead of تونسية, twnsyp, Tunisian), wrong letters (e.g. غلاش glA$ instead of علاش, ElA$, why), …etc.

Furthermore, the use of a lexicon cannot cover all the words of the TA. Hence, some wrong words in the corpus do not admit any change, since they do not exist in the lexicon.

Table 14. Extrinsic evaluation of the second version of COTA orthography system.

Task	Corpus	Recall	Precision	F-measure
Segmentation	Raw text	69.83%	72.33%	71.06%
	Automatically normalized text	**76.9%**	**83.09%**	**79.88%**
POS Tagging	Raw text	71.93%	74.43%	73.16%
	Automatically normalized text	**78.92%**	**81.49%**	**80.18%**
Parsing	Raw text	47.78%	49.4%	48.58%
	Automatically normalized text	**70.31%**	**68.24%**	**69.26%**

5.3 Extrinsic Evaluation

We performed an extrinsic evaluation of our COTA orthography 2.0 system by evaluating the impact of its use on the TA segmenter [40], POS tagger and parser [33].

[40] examined three different methods (deep learning, CRF and SVM) for segmenting TA sentences. Several experiments were carried out in order to enhance the proposed models. Subsequently, the evaluation using a test set of 26.036 words from [1]; [36]; [41]; [6]; [40] revealed that the CRF model produced the highest performance (F-measure = 84,37%), with a 21.47% improvement over deep learning, 18.9% increase over SVM and 23% compared to STAr-TUN system [42].

[33] suggested a semi-automatic annotation method of treebank annotation for the social-media dialect as well as the generation of a parsing model that covers all forms of TA. To enrich the TTB treebank,

the authors annotated a part of the TAD corpus [6]. [33] experimented with different combinations of corpora to generate the best parsing model. This system can be used for POS tagging and parsing tasks. The model is available by email request to the first author.

We prepared automatically segmented, POS tagged and parsed two versions of our test corpus composed of 43K words (see Table 6): the raw corpus that does not follow any spelling convention and the automatically normalized corpus using COTA orthography 2.0. Moreover, two native TA speakers prepared manually POS tagged and parsed versions as a reference for the extrinsic evaluation tasks. Then, both versions were automatically segmented by the TA segmenter. Afterwards, the output provided was compared to the manually segmented version.

According to [40]'s statistics, 33% of the written audio laughter (هههه, hhh) occurs in the first word of the sentence, while 48% of it is in the end of the sentence. Therefore, the normalization of these terms has a significant impact on the segmentation result. The result of [40]'s system improved by 8.82% using the automatically normalized version of the corpus. For POS tagging and parsing, we evaluated the system by the raw (non-normalized) test corpus. The obtained results were 73.16% for POS tagging and 48.58% for parsing. Using the automatically normalized corpus, we achieved better results (80.18% for POS tagging and 69.26% for parsing) with an improvement between 7% and 20% (see Table 14). These evaluations show that our COTA orthography 2.0 system will contribute to the improvement of TA tools.

We may draw the conclusion that by reducing the orthographic heterogeneity, COTA orthography solves the spelling problems and simplifies the experience. One main benefit of using the proposed system is its accuracy in normalizing texts from all the forms of TA. Running a spell checker ensures that the number of orthographic errors in a document decreases significantly, which is very helpful for several NLP tools, such as text segmentation, POS tagging, parsing, etc. (see Table 14). Since it is an entirely automatic tool, it presents good practice to assure high quality without losing time for manual spell checking.

6. CONCLUSION AND FUTURE WORKS

In this paper, we presented an automatic system for orthographic normalization of the Tunisian Arabic. To begin, we expanded the CODA-TA spelling convention [8]. We also extended an existing tool for TA, COTA orthography system [9], by adding new patterns and lexicon. The lexicon was automatically extracted from the Tunisian Treebank «TTB» [33]. Then, a set of patterns was defined to correct social-media dialect errors, such as accentuation, interjections, onomatopoeia, attached clitics, Tatweel, …etc. We also added a language model that is able to choose the appropriate correction automatically. We experimented with the effect of using several options and different corpora combinations to improve the model. Our experiments show that we can improve the overall system performance by 30.78% and 3%, respectively for social-media dialect and spontaneous dialect. Moreover, the use of our system resulted in an increase of about 9% in the outcomes of automated TA segmentation.

In future works, we plan to correct the problem of attached words. Moreover, we consider experimenting with the impact of applying deep learning techniques. We take into account a comparison of the trigram-language model with other neural techniques and apply a tie-breaking method. We will also investigate incorporating this tool in a platform that contains all the linguistic tools of TA.

REFERENCES

[1] I. Zribi, M. Ellouze, L. H. Belguith and P. Blache, "Spoken Tunisian Arabic Corpus STAC: Transcription and Annotation," Research in Computing Science, vol. 90, pp. 123-135, 2015.

[2] A. Masmoudi, M. Ellouze Khmekhem, Y. Esteve, L. Hadrich Belguith and N. Habash, "A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition," Proc. of the 9th Int. Conf. on Language Resources and Evaluation, vol. 3, no. 1, pp. 306–310, 2014.

[3] R. Boujelbane, M. Ellouze, F. Béchet and L. Belguith, "De l'arabe Standard *vers* l'arabe Dialectal: Projection de Corpus et Ressources Linguistiques en vue du Traitement Automatique de l'oral dans les Médias Tunisiens," TAL. 2. Traitement Automatique du Langage Parlé, vol. 55, pp. 73–96, 2014.

[4] A. Masmoudi, N. Habash, M. Ellouze, Y. Estève and L. H. Belguith, "Arabic Transliteration of Romanized Tunisian Dialect Text: A Preliminary Investigation," Proc. of the 16th Int. Conf. on Computat. Linguistics and Intelligent Text Process. (CICLing 2015), pp. 608–619, Cairo, Egypt, 2015.

[5] S. Mdhaffar, F. Bougares, Y. Eve and L. Hadrich-Belguith, "Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments," Proc. of the 3rd Arabic Natural Language Processing Workshop (WANLP), pp. 55–61, Valencia, Spain, 2017.

[6] J. Younes, H. Achour and E. Souissi, "Constructing Linguistic Resources for the Tunisian Dialect Using Textual User-generated Contents on the Social Web," Proc. of the 15th Int. Conf. on Current Trends in Web Engineering, ICWE 2015 Rotterdam, pp. 3–14, The Netherlands, 2015.

[7] S. El Klibi, S. El Hamzaoui, H. Ben Abda, C. Kaddes, F. El Horcheni and A. Maalla, La Constitution en Dialecte Tunisien. Tunisie: Association Tunisienne de Droit Constitutionnel, 2014.

[8] I. Zribi, R. Boujelbane, A. Masmoudi, M. Ellouze, L. H. Belguith and N. Habash, "A Conventional Orthography for Tunisian Arabic," Proc. of the 9th Int. Conf. on Language Resources and Evaluation, European Language Resources Association (ELRA), pp. 2355–2361, Reykjavik, Iceland, May 2014.

[9] R. Boujelbane, I. Zribi, S. Kharroubi and M. Ellouze, "An Automatic Process for Tunisian Arabic Orthography Normalization," Proc. of the 10th International Conference on Natural Language Processing (HrTAL2016), Dubrovnik, Croatia, 2016.

[10] N. Habash, M. T. Diab and O. Rambow, "Conventional Orthography for Dialectal Arabic," Proc. of the 8th Int. Conf. on Language Resources and Evaluation, European Language Resources Association (ELRA), pp. 711–718, Istanbul, Turkey, May 23-25, 2012.

[11] H. Saadane and N. Habash, "A Conventional Orthography for Algerian Arabic," Proc. of the 2nd Workshop on Arabic Natural Language Processing, pp. 69–79, [Online], Available: http://www.aclweb.org/anthology/W15-3208, Beijing, China, July 2015.

[12] M. Jarrar, N. Habash, F. Alrimawi, D. Akra and N. Zalmout, "Curras: An Annotated Corpus for the Palestinian Arabic Dialect," Language Resources and Evaluation, vol. 51, pp. 745–775, 2016.

[13] S. Khalifa, N. Habash, D. Abdulrahim and S. Hassan, "A Large Scale Corpus of Gulf Arabic," Proc. of the 10th Int. Conf. on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 2016.

[14] F. Al-Shargi, A. Kaplan, R. Eskander, N. Habash and O. Rambow, "Morphologically Annotated Corpora and Morphological Analyzers for Moroccan and Sanaani Yemeni Arabic," Proc. of the 10th Int. Conf. on Language Resources and Evaluation (LREC 2016), pp. 1300–1306, Portorož, Slovenia, 2016.

[15] N. Habash, F. Eryani, S. Khalifa et al., "Unified Guidelines and Resources for Arabic Dialect Orthography," Proc. of the 11th Int. Conf. on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan, 2018.

[16] M. Attia, M. Al-Badrashiny and M. Diab, "Gwu-hasp-2015@ qalb-2015 Shared Task: Priming Spelling Candidates with Probability," Proc. of the 2nd Workshop on Arabic Natural Language Processing, pp. 138–143, Beijing, China, 2015.

[17] M. Attia, P. Pecina, Y. Samih, K. Shaalan and J. Van Genabith, "Arabic Spelling Error Detection and Correction," Natural Language Engineering, vol. 22, no. 5, p. 751, 2016.

[18] M. I. Alkanhal, M. A. Al-Badrashiny, M. M. Alghamdi and A. O. Al-Qabbany, "Automatic Stochastic Arabic Spelling Correction with Emphasis on Space Insertions and Deletions," IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 7, pp. 2111–2122, 2012.

[19] A. M. Azmi, M. N. Almutery and H. A. Aboalsamh, "Real-word Errors in Arabic Texts: A Better Algorithm for Detection and Correction," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 27, no. 8, pp. 1308–1320, 2019.

[20] V. N. Vapnik, The Nature of Statistical Learning Theory, New York, NY, USA: Springer-Verlag, New York, Inc., 1995.

[21] F. J. Damerau, "A Technique for Computer Detection and Correction of Spelling Errors," Communications of the ACM, vol. 7, no. 3, pp. 171–176, 1964.

[22] M. Alkhatib, A. A. Monem and K. Shaalan, "Deep Learning for Arabic Error Detection and Correction," ACM Transactions on Asian and Low-resource Language Information Processing (TALLIP), vol. 19, no. 5, pp. 1–13, 2020.

[23] R. Eskander, N. Habash, O. Rambow and N. Tomeh, "Processing Spontaneous Orthography," Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 585–595, Atlanta, Georgia, June 2013.

[24] J. Wang and J.-D. Zucker, "Solving the multiple-instance problem: A Lazy Learning Approach," Proc. of the 17th Int. Conf. on Machine Learning (ser. ICML'00), pp. 1119–1126, San Francisco, USA, 2000.

[25] N. Habash, R. Roth, O. Rambow, R. Eskander and N. Tomeh, "Morphological Analysis and Disambiguation for Dialectal Arabic," Proc. of the Human Language Technologies: Conf. of the North American Chapter of the Association of Computational Linguistics, pp. 426–432, Atlanta, USA, 2013.

[26] W. Adouane, J.-P. Bernardy and S. Dobnik, "Normalizing Non-standardized Orthography in Algerian Code-switched User-generated Data," Proc. of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pp. 131–140, Hong Kong, China, 2019.

[27] A. Mekki, I. Zribi, M. Ellouze Khmekhem and L. Hadrich Belguith, "Critical Description of TA Linguistic Resources," Proc. of the 4th Int. Conf. on Arabic Computational Linguistics (ACLing 2018) & Procedia Computer Science, Dubai, United Arab Emirates, 2018.

[28] S. Mejri, M. Said and I. Sfar, "Pluringuisme et Diglossie en Tunisie," Synergies Tunisie, vol. 1, pp. 53–74, 2009.

[29] M. Graja, M. Jaoua and L. H. Belguith, "Discriminative Framework for Spoken Tunisian Dialect Understanding," Proc. of the 1st Int. Conf. on Statistical Language and Speech Processing (SLSP 2013), vol. 7978, pp. 102–110, Tarragona, Spain, July 29-31, 2013.

[30] W. H. Allehaiby, "Arabizi: An Analysis of the Romanization of the Arabic Script from a Sociolinguistic Perspective," Arab World English Journal, vol. 4, no. 3, 2013.

[31] T. Buckwalter, "Arabic Transliteration," Available: http://www.qamus.org/transliteration.htm, 2002.

[32] N. Habash, M. T. Diab and O. Rambow, "Conventional Orthography for Dialectal Arabic," Proc. of the 8th Int. Conf. on Lang. Resour. and Evaluation (LREC'12), pp. 711-718, Istanbul, Turkey, May 2012.

[33] A. Mekki, I. Zribi, M. Ellouze and L. Hadrich Belguith, "Treebank Creation and Parser Generation for Tunisian Social Media Text," Proc. of the 17th ACS/IEEE Int. Conf. on Computer Systems and Applications (AICCSA), DOI: 10.1109/AICCSA50499.2020.9316462 Antalya, Turkey, 2020.

[34] A. Masmoudi and F. Bougares, "Automatic Speech Recognition System for Tunisian Dialect," Language Resources and Evaluation, vol. 52, no. 1, pp. 249–267, 2017.

[35] R. Boujelbane, Traitements Linguistiques Pour la Reconnaissance Automatique de la Parole Appliquée à la Langue Arabe: de L'arabe Standard *vers* L'arabe Dialectal, Thése de doctorat, Faculté des Sciences Économiques et de Gestion de Sfax, 2016.

[36] A. Mekki, I. Zribi, M. E. Khemakhem and L. H. Belguith, "Syntactic Analysis of the Tunisian Arabic," Proc. of the Int. Workshop on Language Processing and Knowledge Management, September 2017.

[37] A. Al-Thubaity, M. Khan, M. Al-Mazrua and M. Al-Mousa, "New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool," Proc. of the Int. Conf. on Asian Language Processing, pp. 67–70, Urumqi, China, 2013.

[38] J. Guo, H. He, T. He et al., "Gluoncv and Gluonnlp: Deep Learning in Computer Vision and Natural Language Processing," J. of Machine Learning Research, vol. 21, no. 23, pp. 1–7, 2020.

[39] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Datasets," J. of Machine Learning Research, vol. 7, pp. 1–30, 2006.

[40] A. Mekki, I. Zribi, M. E. Khemakhem and L. H. Belguith, "Sentence Boundary Detection of Various Forms of Tunisian Arabic," Language Resources and Evaluation, vol. 56, pp. 357-385, 2022.

[41] R. Boujelbane, M. Mallek, M. Ellouze and L. H. Belguith, "Fine-grained POS Tagging of Spoken Tunisian Dialect Corpora," Proc. of the Int. Conf. on Applications of Natural Language to Data Bases/Information Systems (NLDB 2014), vol. 8455, pp. 59–62, 2014.

[42] I. Zribi, I. Kammoun, M. Ellouze, L. H. Belguith and P. Blache, "Sentence Boundary Detection for Transcribed Tunisian Arabic," Proc. of the 12th Workshop on Natural Language Processing (KONVENS 2016), pp. 323-331, Bochum, Germany, September 2016.

ملخص البحث:

يعّد الخطأ الإملائي مسألةً مُقلقة في مجال معالجة اللّغات الطبيعية، خصوصاً عند التّعامل مع التّعليقات والنّصوص الخام المأخوذة من وسائل التواصل الاجتماعي. وذلك بسبب استخدام الصّيغ الصّرفية الاصطلاحية وعدم الالتزام بالقواعد.

نقترح في هذا البحث نظاماً أوتوماتيكياً لتطبيع النّصوص المكتوبة باللّهجة العامية باتّباع طريقة لتصحيح الأخطاء في العربية التونسية. والجدير بالذّكر أنّ النّظام المعروف باسم (COTA1) لتصحيح الأخطاء الإملائية ليس قادراً على التّعامل مع جميع الصّيغ المعروفة للعربية التونسية. لذا نقترح توسيع قواعد ذلك النّظام ومعاجمه لمعالجة الخصوصيات التي تتميّز بها نصوص وسائل التّواصل الاجتماعي. وبعبارة أخرى، فإنّ نظام (COTA1) يزوّد المستخدم بإمكانيات متعدّدة للتّصحيح. من هُنا، فإنّ النّسخة المقترحة في هذا البحث (COTA2) مزوّدة بنظام أوتوماتيكي يعمل على إرشاد المستخدم الى التّصحيح المناسب. ويشير تقييم النّظام المقترح الى أنّه يعمل على تقليل الأخطاء المتعلّقة بالتّرجمة بنسبة 95.72%.