QCRI@QALB-2015 Shared Task:Correction of Arabic Text for Native and Non-Native Speakers' Errors

Hamdy Mubarak, Kareem Darwish, Ahmed Abdelali
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
{hmubarak, kdarwish, aabdelali}@qf.org.qa

Abstract

This paper describes the error correction model that we used for the QALB-2015 Automatic Correction of Arabic Text shared task. We employed a case-specific correction approach that handles specific error types such as dialectal word substitution and word splits and merges with the aid of a language model. We also applied corrections that are specific to second language learners that handle erroneous preposition selection, definiteness, and gender-number agreement.

1 Introduction

In This paper, we provide a system description for our submissions to the Arabic error correction shared task (QALB-2015 Shared Task on Automatic Correction of Arabic) as part of the Arabic NLP workshop. The QALB-2015 shared task is an extension of the first QALB shared task (Mohit et al., 2014) which addressed errors in comments written to Aljazeera articles by native Arabic speakers (Zaghouani et al., 2014). The current competition includes two tracks, and, in addition to errors produced by native speakers, also includes correction of texts written by learners of Arabic as a foreign language (L2) (Zaghouani et al., 2015). The native track includes Alj-train-2014, Alj-dev-2014, Alj-test-2014 texts from QALB-2014. The L2 track includes L2-train-2015 and L2-dev-2015. This data was released for the development of the systems. The systems were scored on blind test sets Alj-test-2015 and L2-test-2015.

We submitted runs to the automatic correction of text generated by native speaker (L1) and non-native speakers (L2). For both L1 and L2, we employed a case-specific approach that is aided by a language model (LM) to handle specific error types such as dialectal word substitutions and word splits. We also constructed a list of corrections that we observed in the QALB-2014 data set and in the QALB-2015 training set. We made use of these corrections to generate alternative corrections for words. When dealing with L2 text, we noticed specific patterns of mistakes mainly related to gender-number agreement, phonetic spellings, and definiteness. As for punctuation recovery, we opted only to place periods at the end of sentences and to correct reversed question marks. We opted not to invest in punctuation recovery based on the mixed results we obtained for the QALB-2014 shared task (Mubarak and Darwish, 2014).

2 QALB L2 Corpus Error Analysis

The QALB corpus used for the task contains over two million words of manually corrected Arabic text. The corpus is composed of text that is produced by native speakers as well as non-native speakers (Habash et al., 2013). While annotating the corpus, Zaghouani et al. (2014) detailed various types of errors that were encountered and addressed - mainly L1. Additional proposed corrections for L2 errors were summarized with no details. Understanding the error types would shed light on their manifestations and help correct them properly. We inspected the training and development sets and noticed a number of potential issues that can be summarized as follows:

1. Syntax Errors due to first language influence: L2 learners may carry over rules from their native languages resulting in syntactic and morphological errors, such as:

(a) Definiteness: In Arabic syntax, a possessive case, idafa construct, which happens between two words, mostly requires that the first word be indefinite while the second be definite. Such as the case of "كتاب التلميذ" (ktAb Al-

tlmy* [1] – "The book of the student"). Note, the first Arabic word doesn't contain the definite article "Al" while the second does. Erroneous application, or not, of the definite article was common. For example, the student may say: "كتاب تلميذ" (ktAb tlmy*) or "الكتاب التلميذ" (AlktAb Altlmy*).

(b) Gender-number agreement: Gender-number agreement is another common error type. The inflectional morphology of Arabic may: embed gender-number markers in verbs as in "أعجبتني المدينة" (>Ejbtny Almdynp – I liked the city) and the learner may write "أعجبني المدينة" (>Ejbny Almdynp) without the feminine marker; and the use of feminine singular adjectives with masculine plural inanimate nouns as in "مدن عظيمة" (mdn EZymp – great cities) and the learner may write "مدن عظيمون" (mdn EZymwn) or "مدن عظيمات" (mdn EZymAt).

(c) Prepositions: Mixing the usage of prepositions is another typical challenge for L2 learners, as it requires good understanding of spacio-temporal aspects of language. Thus, L2 learners tend to mix between these prepositions as in "وصلت في المدينة" (wSlt fy Almdynp – I arrived in the city) instead of "وصلت إلى المدينة" (wSlt ¡lY Almdynp – I arrived to the city).

2. Spelling errors: Grasping sounds is another challenging issue particularly given:

(a) Letter that sound the same but written differently, such as "ت" (t) and "ة" (p), may lead to erroneous spellings like "مبارات" (mbArAt – game) instead of "مباراة" (mbArAp). Other example letter pairs are "ص" (S) and "س" (s) and "ط" (T) and "ت" (t)

(b) Letters that have similar shapes but a differ number of dots on or below them. We noticed that L2 learners often confuse letter such as: "ج" (j), "ح" (H), and خ (x); and "ص" (S) and "ض" (D). This may lead to errors such as "صبب الخادث" (Sbb AlxAdv) instead of "سبب الحادث" (sbb AlHAdv – the reason for the accident).

3 Word Error Correction

In this section we describe our case-specific error correction system that handles specific error types with the aid of a language model (LM) generated from an Aljazeera corpus. We built a word bigram LM from a set of 234,638 Aljazeera articles[2] that span 10 years. Mubarak et al. (2010) reported that spelling mistakes in Aljazeera articles are infrequent. We used this language model in all subsequent steps.

We attempted to address specific types of errors including dialectal words, word normalization errors, and words that were erroneously split or merged. Before applying any correction, we always consulted the LM. We handled the following cases in order (L2 specific corrections are noted):

• Switching from English punctuation marks to Arabic ones, namely changing: "?" → "؟" and "," → "،".

• Correcting errors in definite article (ال "Al") when it's preceded by the preposition (ل "l") ex: لالعمل "lAlEml" → للعمل "llEml".

• Handling common dialectal words and common word-level mistakes. To do so, we extracted all the errors and their corrections from the QALB-2014 (train, dev, and test) and the training split of the QALB-2015 data set. In all, we extracted 221,460 errors from this corpus. If an error had 1 seen correction and the correction was done at least 2 times, we used the correction as a deterministic correction. For example, the word (الاحداث "AlAHdAv" – the events) was found 86 times in this corpus, and in all cases it was corrected to (الأحداث $Al > HdAv''$). There were 10,255 such corrections. Further, we manually revised words for which a specific correction was made in 60% or more of the cases (2,076 words) to extract a list of valid alternatives for each word. For example, the word (الامور "AlAmwr") appeared 157 times and was corrected to (الأمور $Al > mwr''$) in 99% of the cases. We ignored the remaining seen corrections. An example dialectal word is (اللي "Ally" – "this" or "that")

[1] Buckwalter transliteration

[2] http://www.aljazeera.net

which could be mapped to (الذي "Al*y"), (التي "Alty"), or (الذين "Al*yn"). An example of a common mistake is (إنشاء الله " $> n\$A'Allh''$ – "God willing") which is corrected to (إن شاء الله ">n $A' Allh"). When performing correction, given a word appearing in our list, we either replaced it deterministically if it had one correction, or we consulted our LM to pick between the different alternatives. When dealing with L2 data, we added 297 more deterministic errors (ex: وثم "wvm" was always corrected to ثم "vm").

• Handling split conjunctions (و "w") that should be concatenated with the next word (ex: و هناك "w HnAk" → وهناك "wHnAk").

• Handling errors pertaining to the different forms of *alef*, *alef maqsoura* and *ya*, and *ta marbouta* and *ha* as described in Table 1 and Table 2. We used an approach similar to the open suggested by Moussa et al. (Moussa et al., 2012), and we also added the following cases, namely attempting to replace: ؤ "&" with ؤو "&w" or ئو "}w"; and ئ "}" with يء "y"' or vice versa (ex: مرؤس "mr&s" → مرؤوس "mr&ws", قاريء "qAry"' → قارئ "qAr}"). To generate the alternatives for words, we normalized all the unique words in the Aljazeera corpus, and we constructed a reverse look-up table that has the normalized form as the key and a list of seen alternatives that could have generated the normalized form. The look-up table contained 905k normalized word entries with corresponding denormalized forms. When correcting, a word is normalized and looked-up in the table to retrieve possible alternatives. We used the LM to pick the best alternative in context. Table 2 shows examples from the look-up table for normalized words and their alternative corrections.

• Removing repeated letters. Often people repeat letters, particularly long vowels, for emphasis as in (أخيييرااا ">xyyyyrAAA") (meaning "at last"). We corrected for elongation in a manner similar to that of Darwish et al. (Darwish et al., 2012). When a long vowel is repeated, we replaced it with a either the vowel (ex. أخيرا ">xyrA" – finally) or the vowel with one repetition (ex. سعودييين "sEwdyyn" – Saudis) and scored it using the LM. This was expanded to consonants also (ex. بكثيررررر "bkvyrrrr" → بكثير "bkvyr"). If a repeated *alef* appeared in the beginning of the word, we attempted to replace it with alef lam (ex. االحضارة "AAHDArp" → الحضارة "AlHDArp" – "civilization"). A trailing alef-hamza-alef sequence was replaced by alef-hamza (ex. سماءا "smA'A" → سماء "smA"' (meaning "sky")). Also, we replaced (لللل "lll") at the beginning of word by (لل "ll") (ex. للغة "lllgp" → لغة "llgp").

• Handling grammar errors in verb suffixes to restore missing alef (ex. افعلو "AfElw" → افعلوا "AfElwA" – do (plural); سيفعلوا "syfElwA" → سيفعلون "syfElwn" – they will do; لتحفظون "ltHfZwn" → لتحفظوا "ltHfZwA" – that you may memorize/protect).

• Handling merges and splits. Often words are concatenated erroneously. Thus, we attempted to split all words that were at least 5 letters long after letters that don't change their shapes when they are connected to the letters following them, namely different alef forms, د "d", ذ "*", ر "r", ز "z", و "w", ة "p", and ى "Y" (ex: ياربنا "yArbnA" → يا ربنا "yA rbnA"). If the bigram was observed in the LM and the LM score was higher (in context) than when they were concatenated, then the word was split. Conversely, some words were split in the middle. We attempted to merge every two words in sequence. If the LM score was higher (in context) after the merge, then the two words would be merged (ex: انتصار ات "AntSAr At" → انتصارات "AntSArAt").

• Correcting out-of-vocabulary (OOV) words. For words that were not observed in the LM, we attempted replacing phonetically or visually similar letters and deleting/replacing letters that appear in dialectal words as shown in Table 3. Generated suggestions are scored in context using the LM. Many of these errors are common in the L2 data set.

• For L2 data only, as we mentioned earlier we observed errors pertaining to definiteness and gender-number agreement. We generated possible corrections as follows: words that start with definite article, we scored the word with and with-

out a definite article. We did the same with words ending with ta marbouta (p). We also added other alternatives for the word by adding the definite article and/or that ta marbouta (for words without one or the other or neither). In all cases, we used the LM to select the most probable alternative in contexts.

Letter	Norm.	Example
أ، إ، آ >, <, |	ا A	أحمد ← احمد $AHmd \leftarrow > Hmd$ إقناع ← اقناع $AqnAE \leftarrow < qnAE$ آمن ← امن $Amn \leftarrow \|mn$
ى Y	ي y	قصوى ← قصوي $qSwy \leftarrow qSwY$
ة p	ه h	قيادة ← قياده $qyAdh \leftarrow qyAdp$
ؤ، ئ &, }	ء '	مسؤول ← مسءول ms&wl ← ms'wl
diacritics	*null*	مُثَقَف ← مثقف $mvqf \leftarrow muvaq\ fK$
kashida	*null*	كبيـــــر ← كبير $kbyr \leftarrow kby\_\_r$

Table 1: Word Normalization.

4 Official Shared Task Experiments and Results

We submitted 1 run for L1 errors (QCRI-1-ALJ), and 2 runs for L2 errors (QCRI-1-L2, QCRI-2-L2) as follows:

1. QCRI-1-ALJ: case-based correction for L1 test.
2. QCRI-2-L1: case-based correction for L2 test file and also by adding alternatives for possible errors in the definite article ال "Al" and feminine mark ة "p" as described in section 3.
3. QCRI-1-L2: case-based correction for L2 test file with handling the definiteness or feminine marker.

Table 4 and Table 5 report the officially submitted results against the development set and test set in order, and Table 6 reports the results of the new system against the development set and test set of QALB 2014 shared task.

Word	Alternatives and Frequencies
اعلام AElAm	5 اعلام ,632 أعلام ,20352 إعلام $< ElAm$ 20352, $> ElAm$ 632, $AElAm$ 5
حضاره HDArh	1 حضاره ,1271 حضارة $HDArp$ 1271, $HDArh$ 1

Table 2: Word Alternatives.

Case	Example
ض، ظ Z, D	ظابط ← ضابط DAbT ← ZAbT
د، ذ d, *	الدهب ← الذهب Al*hb ← *Aldhb*
ب+ b+	بيلعب ← يلعب ylEb ← bylEb
د+ d+	ديلعب ← يلعب ylEb ← dylEb
ح+، ه+ H+, h+	حيكتب ← سيكتب syktb ← Hyktb
هال+ hAl+	هالبنت ← هذه البنت h*h Albnt ← hAlbnt
عال+ EAl+	عالأرض ← على الأرض ElY AlArD ← EAlArD
ت، ط t, T	اللاطينية ← اللاتينية AllAtynyp ← AllATynyp
ج، ح، خ j, H, x	التحصص ← التخصص AltxSS ← AltHSS
ق، ك q, k	دقتوراه ← دكتوراه dktwrAh ← dqtwrAh
ال+ ... +ي Al+ ... +y	التخرجي ← التخرج Altxrj ← Altxrjy

Table 3: Phonetic, Dialectal, and L2 Errors

5 Conclusion

In this paper, we presented an automatic approach for correcting Arabic text based on handling specific error types. We handled common dialectal words, some dialectal morphological features, letter normalization errors (ex. alef, ta marbouta, etc.), and word splitting and merging. For the L2 corpus, we also corrected letters that L2 learners often confuse because of similarity in shape or sound, and we attempted to correct errors pertaining to definiteness and gender-number agreement. For punctuation recovery, we opted to put periods at the end of sentences. Preliminary experiments using fuzzy match using a character-based mod-

Run	P	R	F1
QCRI-1-ALJ (Alj-dev-2015)	84.2	49.8	62.6
QCRI-1-L2 (L2-dev-2015)	46.3	19.2	27.1
QCRI-2-L2 (L2-dev-2015)	57.6	16.3	25.4

Table 4: Official Results for Dev. Data

Run	P	R	F1
QCRI-1-ALJ (Alj-test-2015)	84.74	58.10	68.94
QCRI-1-L2 (L2-test-2015)	45.86	20.16	28.01
QCRI-2-L2 (L2-test-2015)	54.87	17.63	26.69

Table 5: Official Results for Test Data

els showed promising results(Sajjad et al., 2012; Durrani et al., 2014; Darwish et al., 2014). We intend to incorporate this development among others in our on-going research. The fuzzy match algorithm will correct cases like: (الأبعاء، يستخدنونها, Al>bEA' , ystxdnwnhA) to (الأربعاء، يستخدمونها, Al<rbEA' , ystxdmwnhA).

L2 learners present new spelling error types. Such types may not typical spelling errors as they may produce valid words that are erroneous in context. Hence employing a methodology to detect such cases will be of great help.Also, we plan to handle more grammar errors for cases like: numbers, case endings, gender-number agreement, irregular (broken) plurals, and Tanween errors (الممنوع من الصرف).

Run	P	R	F1
Alj-dev-2014	65.42	62.96	64.17
Alj-test-2014	65.79	61.94	63.81

Table 6: Results for QALB 2014 Data Sets

References

Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for arabic microblog retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2427–2430. ACM.

Kareem Darwish, Ahmed M. Ali, and Ahmed Abdelali. 2014. Query term expansion by automatic learning of morphological equivalence patterns from wikipedia. In *SIGIR 2014 Workshop on Semantic Matching in Information Retrieval (SMIR)*, volume 1204, pages 24–29. CEUR-WS.

Nadir Durrani, Hassan Sajjad, Hieu Hoang, and Philipp Koehn. 2014. Integrating an unsupervised transliteration model into statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 148–153, Gothenburg, Sweden, April. Association for Computational Linguistics.

Nizar Habash, Behrang Mohit, Ossama Obeid, Kemal Oflazer, Nadi Tomeh, and Wajdi Zaghouani. 2013. Qalb: Qatar arabic language bank. In *Proceedings of Qatar Annual Research Conference (ARC-2013)*, Doha, Qatar.

Behrang Mohit, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, and Ossama Obeid. 2014. The First QALB Shared Task on Automatic Text Correction for Arabic. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing*, Doha, Qatar, October.

Mohammed Moussa, Mohamed Waleed Fakhr, and Kareem Darwish. 2012. Statistical denormalization for arabic text. In *In Empirical Methods in Natural Language Processing*.

Hamdy Mubarak and Kareem Darwish. 2014. Automatic correction of arabic text: a cascaded approach. *Arabic NLP 2014 Workshop*.

Hamdy Mubarak, Mostafa Ramadan, and Ahmed Metwali. 2010. Spelling mistakes in arabic newspapers. In *Arabic Language and Scientific Researches conference*, Faculty of Arts, Ain Shams University, Cairo, Egypt.

Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2012. A statistical model for unsupervised and semi-supervised transliteration mining. In *Proceedings of the Association for Computational Linguistics*, ACL '12, pages 469–477, Jeju, Korea.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale arabic error annotation: Guidelines and framework. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May.

Wajdi Zaghouani, Nizar Habash, Houda Bouamor, Alla Rozovskaya, Behrang Mohit, Abeer Heider, and Kemal Oflazer. 2015. Correction annotation for non-native arabic texts: Guidelines and corpus. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 129–139, Denver, Colorado, USA, June. Association for Computational Linguistics.