The Second QALB Shared Task on Automatic Text Correction for Arabic

Alla Rozovskaya[1], **Houda Bouamor**[2], **Nizar Habash**[3],
Wajdi Zaghouani[2], **Ossama Obeid**[2] **and Behrang Mohit**[4]

[1]Center for Computational Learning Systems, Columbia University
[2]Carnegie Mellon University in Qatar
[3]New York University Abu Dhabi
[4]Ask.com

alla@ccls.columbia.edu,hbouamor@qatar.cmu.edu,nizar.habash@nyu.edu
wajdiz@qatar.cmu.edu,owo@qatar.cmu.edu,behrang@cmu.edu

Abstract

We present a summary of QALB-2015, the second shared task on automatic text correction of Arabic texts. The shared task extends QALB-2014, which focused on correcting errors in Arabic texts produced by native speakers of Arabic. The competition this year, in addition to native data, includes texts produced by learners of Arabic as a foreign language. The report includes an overview of the QALB corpus, which is the dataset used for training and evaluation, an overview of participating systems, results of the competition and an analysis of the results and systems.

1 Introduction

The task of text correction has recently been attracting a lot of attention in the Natural Language Processing (NLP) community, but most of the effort in this area concentrated on English, especially on errors made by learners of English as a Second Language. Four competitions devoted to error correction for non-native English writers took place recently: HOO (Dale and Kilgarriff, 2011; Dale et al., 2012) and CoNLL (Ng et al., 2013; Ng et al., 2014). Shared tasks of this kind are extremely important, as they bring together researchers and promote the development of relevant techniques and dissemination of key resources, such as benchmark data sets.

In the area of Arabic text correction, there has been a significant body of work, as well (Shaalan et al., 2003; Hassan et al., 2008). However, due to the lack of a common benchmark data set, making progress on this task has been difficult. The QALB shared task on automatic text correction of Arabic, organized within the framework of the Qatar Arabic Language Bank (QALB) project,[1] is the first effort aimed at constructing a benchmark data set, which will allow for development and evaluation of automatic correction systems for Arabic.

In this paper, we present a summary of the second edition of the QALB competition. The first one – QALB-2014 (Mohit et al., 2014) – took place in conjunction with the Arabic NLP workshop at EMNLP-2014 and focused on errors found in online commentaries produced by native speakers of Arabic. QALB-2014 attracted a lot of attention and resulted in nine systems being submitted with a variety of approaches that included rule-based frameworks, machine-learning classifiers, and statistical machine translation methods. This year's competition extends the first edition by adding another track that focuses on errors found in essays written by learners of Arabic.

Eight teams participated in the competition this year, including several participants from last year who submitted improved systems for the native track. The non-native (L2) track also allowed the participants to determine to what extent their approaches need to be modified to adapt to a new set of errors. Overall, QALB-2015 generated a diverse set of approaches for automatic text correction of Arabic.

The rest of the paper is organized as follows. In Section 2, we present the shared task framework. This is followed by an overview of the QALB corpus (Section 3). Section 4 describes the shared task data, and Section 5 presents the approaches adopted by the participating teams. Section 6 discusses the results of the competition. Section 7 concludes the paper.

[1]http://nlp.qatar.cmu.edu/qalb/

Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 26–35,
Beijing, China, July 26-31, 2015.

2 Task Description

The QALB-2015 shared task extends QALB-2014, the first shared task on Arabic text correction that was created as a forum for competition and collaboration on automatic error correction in Modern Standard Arabic and took place in conjunction with the Arabic NLP workshop at EMNLP-2014 (Mohit et al., 2014).

QALB-2014 addressed errors in online user comments written to Aljazeera articles by native Arabic speakers. This year's competition includes two tracks – native and non-native. In addition to the Aljazeera commentaries written by native speakers, it also includes texts produced by learners of Arabic as a foreign language (L2).

Both the native and the non-native data is written in Modern Standard Arabic and is part of the *QALB corpus* (see Section 3), a manually-corrected collection of Arabic texts. The Aljazeera section of the corpus is presented in Zaghouani et al. (2014). The L2 data is extracted from two learner corpora of Arabic – the Arabic Learners Written Corpus (ALWC) (Farwaneh and Tamimi, 2012) and the Arabic Learner Corpus (ALC) (Alfaifi and Atwell, 2012). For details about the L2 data, we refer the reader to Zaghouani et al. (2015a).

The shared task participants were provided with training and development data to build their systems, but were also free to make use of additional resources, including corpora, linguistic resources, and software, as long as these were publicly available.

For evaluation, a standard framework developed for similar error correction competitions in English and that we also used last year has been adopted: system outputs are compared against gold annotations using *Precision*, *Recall* and F_1. Systems are ranked based on the F_1 scores obtained on the test sets.

3 The QALB Corpus

The QALB corpus was created as part of the QALB project. One of the goals of the QALB project is to develop a large manually corrected corpus for a variety of Arabic texts, including texts produced by native and non-native writers, as well as machine translation output. Within the framework of this project, comprehensive annotation guidelines and a specialized web-based annotation interface have been developed (Zaghouani et al., 2014; Obeid et al., 2013; Zaghouani et al., 2015a).

The texts are manually annotated for errors by native Arabic speakers. The annotation begins with an initial automatic pre-processing step. Next, the files are processed with the morphological analysis and disambiguation system MADAMIRA (Pasha et al., 2014) that corrects a common class of spelling errors. The files are then assigned to a team of trained human annotators who were instructed to correct all errors in the input.

The errors include spelling, punctuation, word choice, morphology, syntax, and dialectal usage. However, it should be stressed that the error classification was only used for guiding the annotation process; the annotators were not instructed to mark the type of error but only needed to specify an appropriate correction.

Once the annotation was complete, the corrections were automatically grouped into the following seven *action categories* based on the *action* required to correct the error: {*Edit, Add, Merge, Split, Delete, Move, Other*}.[2]

Table 1 presents a sample Arabic news comment along with its manually corrected form, its romanized transliteration,[3] and the English translation. The errors in the original and the corrected forms are underlined and co-indexed. Table 2 presents a subset of the errors for the example shown in Table 1 along with the error types and annotation actions. The Appendix at the end of the paper lists **all** annotation actions for that example.[4]

Essays written by L2 speakers differ from the native texts both because of the genre and the types of mistakes. For this reason, the general QALB L1 annotation guidelines were extended by adding new rules describing the error correction procedure in texts produced by L2 speakers (Zaghouani et al., 2015a). Because the genres are different, the writing styles exhibit different distributions of words, phrases, and structures. Further, while native texts mostly contain orthographic and punctuation mistakes, non-native writings also reveal lexical choice errors, missing and extraneous words (e.g. articles, prepositions), and mistakes in word

[2] In the shared task, we specified two *Add* categories: *add_before* and *add_after*. Most of the add errors fall into the first category, and we combine these here into a single *Add* category.

[3] Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007): (in alphabetical order) *AbtθjHxdðrzsšSDTĎς*γ*fqklmnhwy* and the additional symbols: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ئ , ى ħ ة, ý ى.

[4] Tables 1 and 2, and the appendix are reproduced from Mohit et al. (2014) to help explain the format of the files used in QALB-2014 and QALB-2015 shared task evaluations.

Original	Corrected
لا تتصوروا مدي سعادتي عند قرائة هذة التحليلات الرائعة و المحترمة لأاني شاب و كنت بتمني من الله ان أؤدي العمرة مرورا بالمسجد الاقصي و كان يبدوا ان هذا بعيد المنال فكل ما في حد يسمع الامنية كان بيقول انك ممكن تتمني ان أحفاد أحفادك يحققوهالأن امنيتك مستحيلة.	لا تتصوروا مدى سعادتي عند قراءة هذه التحليلات الرائعة والمحترمة. لأنني شاب وكنت أتمنى من الله أن أؤدي العمرة مرورا بالمسجد الأقصى، وكان يبدو أن هذا بعيد المنال، فكل واحد يسمع الأمنية كان يقول أنك ممكن أن تتمنى أن أحفاد أحفادك يحققوها لأن أمنيتك مستحيلة.
lA ttSwrwA mdy[1] sçAdty çnd qrAŷħ[2] hðħ[3] AltHlylAt AlrAŷçħ w AlmHtrmħ[4] lÂAny[6] šAb w knt[7] btmny[8] mn Allh An[9] Âŵdy Alçmrħ mrwrA bAlmsjd AlAqSy[10] w kAn[12] ybdwA[13] An[14] hðA bçyd AlmnAl fkl mA[16] fy[17] Hd[18] ysmç AlAmnyħ[19] kAn byqwl[20] Ank[21] mmkn ttmny[23] An[24] ÂHfAd ÂHfAdk yHqqwhAlÂn[25] Amnytk[26] mstHylħ.	lA ttSwrwA mdý[1] sçAdty çnd qrA'ħ[2] hðh[3] AltHlylAt AlrAŷçħ wAlmHtrmħ[4],[5] lÂnny[6] šAb wknt[7] Âtmný[8] mn Allh Ân[9] Âŵdy Alçmrħ mrwrA bAlmsjd AlÂqSý[10],[11] wkAn[12] ybdw[13] Ân[14] hðA bçyd AlmnAl,[15] fkl wAHd[18] ysmç AlÂmnyħ[19] kAn yqwl[20] Ânk[21] mmkn Ân[22] ttmný[23] Ân[24] ÂHfAd ÂHfAdk yHqqwhA lÂn[25] Âmnytk[26] mstHylħ.

Translation

You cannot imagine the extent of my happiness when I read these wonderful and respectful analyses because I am a young man and I wish from God to perform Umrah passing through the Al-Aqsa Mosque; and it seemed that this was elusive that when anyone heard the wish, he would say that you can wish that your great grandchildren may achieve it because your wish is impossible.

Table 1: A sample of an original (erroneous) text along with its manual correction and English translation. The indices in the table are linked with those in Table 2 and the Appendix.

#	Error	Correction	Error Type	Correction Action
#1	مدي mdy	مدى mdý	Spelling	Edit
#6	لأاني lÂAny	لأنني lÂnny	Spelling	Edit
#8	بتمني btmny	أتمنى Âtmný	Dialectal	Edit
#11	*Missing Comma*	،	Punctuation	Add_before
#12	و كان w kAn	وكان wkAn	Spelling	Merge
#13	يبدوا ybdwA	يبدو ybdw	Morphology	Edit
#25	يحققوهالأن yHqqwhAlÂn	يحققوها لأن yHqqwhA lÂn	Spelling	Split

Table 2: Error type and correction action for seven examples extracted from the sentence pair in Table 1. The indices are linked to those in Table 1 and the Appendix.

order, as shown in Table 3. Finally, even when a sentence written by a non-native writer does not contain obvious mistakes, it often still does not sound fluent to a native speaker.

4 Shared Task Data

To develop their systems, participants were provided with training and development data three months prior to the release of the blind test sets. For the native (*Aljazeera*) track, the participants used the data sets from QALB-2014. We refer to these data sets as *Alj-train-2014*, *Alj-dev-2014*, and *Alj-test-2014*. The L2 track includes *L2-train-2015* and *L2-dev-2015*. The systems were evaluated on blind test sets *Alj-test-2015* and *L2-test-2015*.

Both for the native and L2 data, we ensured that sentences from the same comment or essay belonged to the same set, i.e. training, development, or test. Furthermore, Aljazeera comments belonging to the same article were included only in one of the shared task subsets (i.e., training, development or test). The commentaries were also split by the annotation time.

Similar to QALB-2014, the data was made available to the participants in three versions:

Error	المدينة **المدينة منوره جميل جدا جوه** *Almdynħ* ***Almdynħ mnwrh*** *jmyl jdA* ***jwh***
Edit	المدينة **المنورة جميل جدا جوها** *Almdynħ* ***Almnwrħ*** *jmyl jdA* ***jwhA***
English	The Madinah Munawwarah's atmosphere is very beautiful

Table 3: Example of three errors shown in bold and described in order. The word **المدينة** *Almdynħ* is repeated and should be removed. The word **منوره** *mnwrh* is missing the definite article ال *Al* at the beginning of the word and the Ta-Marbuta ة *ħ* is confused with the letter Ha ه *h*. The correct word should be **المنورة** *Almnwrħ*. Finally, there is a possessive pronoun agreement error in the word **جوه** *jwh* and it should be spelled جوها *jwhA* instead.

Data	**Error type (%)**						
	Edit	**Add**	**Merge**	**Split**	**Delete**	**Move**	**Other**
Alj-train-2014	55.3	32.4	5.9	3.5	2.2	0.1	0.5
Alj-dev-2014	53.5	34.2	5.0	3.7	2.0	0.1	0.5
Alj-test-2014	51.9	34.7	5.9	3.5	3.3	0.2	0.5
Alj-test-2015	51.9	34.7	5.9	3.5	3.3	0.2	0.5
L2-train-2015	60.7	27.2	5.0	1.9	4.4	<1	-
L2-dev-2015	60.8	26.9	5.2	1.5	4.4	1.4	-
L2-test-2015	60.3	27.5	5.2	1.5	4.6	1.1	-

Table 5: Distribution of annotations by type in the shared task data. Error types denotes the action required in order to correct the error.

Data set	**# of words**	**# of corrections**
Alj-train-2014	1M	306K
Alj-dev-2014	54K	16K
Alj-test-2014	51K	16K
Alj-test-2015	49K	13K
L2-train-2015	43K	13.2K
L2-dev-2015	25K	7.3K
L2-test-2015	23K	6.6K

Table 4: Statistics on the shared task data.

(1) plain text, one document per line; (2) text with annotations specifying errors and the corresponding corrections; (3) feature files specifying morphological information obtained by running MADAMIRA, a tool for morphological analysis and disambiguation of Modern Standard Arabic (Pasha et al., 2014). MADAMIRA performs morphological analysis and contextual disambiguation. Using the output of MADAMIRA, we generated for each word thirty-three features. The features specify various properties: the part-of-speech (POS), lemma, aspect, person, gender, number, and so on.

Among its features, MADAMIRA generates normalization forms and as a result corrects a large subset of a special class of spelling mistakes in words containing the letters *Alif* and final *Ya*. These letters are a source of the most common spelling types of spelling errors in Arabic and involve *Hamzated Alifs* and *Alif-Maqsura/Ya* confusion (Habash, 2010; El Kholy and Habash, 2012). We refer to these errors as *Alif/Ya* errors (see also Section 6). Several participants this year and in QALB-2014 (e.g. Rozovskaya et al. (2014)) used MADAMIRA predictions as part of their systems. We show the performance of the MADAMIRA baseline in Sec. 6.

Table 4 presents statistics on the shared task data for native and non-native tracks separately. Table 5 shows the distribution of annotations by the action type. The majority of corrections (over 50%) belong to the type *Edit*. This is followed by mistakes that require an insertion of missing word or punctuation (about a third of all errors). With respect to the differences between Aljazeera and L2 data, note that the L2 data has a higher percentage of corrections of type *Edit* but fewer additions of missing words. This could be explained by the fact that a large percentage of Aljazeera errors (over 40%) involve missing punctuation. In addition to this difference, there are almost twice as many deletions and five time more moves in the L2 data, which could be due to grammatical errors that are not typical for native speakers.

Team Name		Affiliation
ARIB	(AlShenaifi et al., 2015)	King Saud University (Saudi Arabia)
CUFE	(Nawar, 2015)	Cairo University (Egypt)
GWU	(Attia et al., 2015)	George Washington University (USA)
QCMUQ	(Bouamor et al., 2015)	Carnegie Mellon University in Qatar (Qatar) and Qatar Computing Research Institute (Qatar)
QCRI	(Mubarak et al., 2015)	Qatar Computing Research Institute (Qatar)
SAHSOH	(Zaghouani et al., 2015b)	Bouira University (Algeria) and Carnegie Mellon University in Qatar (Qatar)
TECH	(Mostefa et al., 2015)	Techlimed.com (France)
UMMU	(Bougares and Bouamor, 2015)	Laboratoire d'Informatique de l'Université du Maine (France) and Carnegie Mellon University in Qatar (Qatar)

Table 6: List of teams that participated in the shared task.

Team	Approach	External Resources
ARIB	Corrections proposed by MADAMIRA; rules; levenshtein distance for spelling correction; Probabilistic-Based Spelling Correction; autocorrect Ghaltawi; Punctuation module	KSU corpus of classical Arabic; Open Source Arabic Corpora; Al Sulaiti Corpus; KACST Arabic Corpus; KHAWAS tool; autocorrect Ghaltawi
CUFE	Rules extracted from the Buckwalter morphological analyser; their probabilities are learned using the training data	Buckwalter morphological analyzer Version 2.0 (Buckwalter, 2004)
GWU	A CRF model for punctuation errors; a dictionary, probabilistic candidate generation, and a language model for spelling and grammar errors; regular expressions and normalization rules	AraComLex Extended dictionary (Attia et al., 2012); Arabic Gigaword Fourth Edition (Parker et al., 2009)
QCMUQ	Rule-based techniques; MADAMIRA corrections; SMT; language models; finite-state automata	AraComLex dictionary (Attia et al., 2012);Arabic Gigaword Fourth Edition (Parker et al., 2009); news commentary corpus
QCRI	Case-specific correction module; language model	Aljazeera articles
TECH	(1) Rule-based system using Hunspell (2) Hybrid system: Statistical MT with Madamira and rules	Newspaper articles from Open Source Arabic Corpora; other corpora collected online; Hunspell
SAHSOH	Rules, regular expressions, Ghaltawi	Arabic word list; JRC-Names; Alfaifi L1 and L2 corpus; Hunspell; Ayaspell dictionary; Ghalatawi; AkhtaBot script
UMMU	MADAMIRA corrections; word-level SMT and character-level SMT systems	Native Arabic data

Table 7: Approaches adopted by the participating teams.

5 Participants and Approaches

Eight teams participated in the shared task. Table 6 presents the list of participating institutions and their names in the shared task. Each team was allowed to submit up to three outputs. Overall, we received 12 outputs for the native track and 10 outputs for the non-native track (one of the teams – TECH – did not participate in the non-native track).

The submitted systems included a diverse set of approaches that incorporated rule-based frameworks, statistical machine translation and machine-learning models, as well as hybrid systems. The teams that scored at the top employed hybrid methods by combining a variety of techniques. For example, the CUFE system extracted rules from the morphological analyzer and learned their probabilities using the training data, while the UMMU system combined statistical machine-translation with MADAMIRA corrections. Table 7 summarizes the approaches adopted by each team.

6 Results

In this section, we present the results of the competition. As was done in QALB-2014, we adopted the standard Precision (P), Recall (R), and F_1 metric. This metric was also used in recent shared tasks on grammatical error correction in English: HOO competitions (Dale and Kilgarriff, 2011; Dale et al., 2012) and CoNLL (Ng et al., 2013). The results are computed using the M2 scorer (Dahlmeier and Ng, 2012) that was also used in the CoNLL shared tasks.

Tables 8 and 9 present the official results of the evaluation on the test sets for the Aljazeera data and the L2 data, respectively. The results are sorted according to the F_1 scores obtained by the

Rank	Team	P	R	F_1
1	CUFE	88.85	61.76	**72.87**
2	UMMU-1	70.28	71.93	71.10
3	GWU	74.69	67.51	70.92
4	UMMU-2	72.69	67.52	70.01
5	QCRI	84.74	58.10	68.94
6	QCMUQ	71.39	65.13	68.12
7	TECH-2	71.20	64.94	67.93
8	TECH-1	71.08	64.74	67.76
9	TECH-3	69.99	60.41	64.85
10	ARIB-1	64.50	56.50	60.23
11	ARIB-2	67.56	51.61	58.52
12	SAHSOH	81.88	40.24	53.97
	MADAMIRA	80.32	39.98	53.39

Table 8: Official results on the test set (Alj-test-2015). Column 1 shows the system rank according to the F_1 score. MADAMIRA refers to the *baseline* of applying corrections proposed by MADAMIRA.

Rank	Team	P	R	F_1
1	UMMU-1	54.12	33.26	**41.20**
2	QCMUQ	50.37	31.68	38.90
3	UMMU-2	55.83	29.47	38.58
4	CUFE	70.92	23.85	35.69
5	GWU	55.66	23.32	32.87
6	ARIB-3	48.79	24.57	32.68
7	ARIB-2	50.08	22.30	30.86
8	QCRI-1	45.86	20.16	28.01
9	QCRI-2	54.87	17.63	26.69
10	SAHSOH	59.75	15.90	25.12
	MADAMIRA	45.24	13.09	20.30

Table 9: Official results on the test set (L2-test-2015). Column 1 shows the system rank according to the F_1 score. Column 1 shows the system rank according to the F_1 score. MADAMIRA refers to the *baseline* of applying corrections proposed by MADAMIRA.

systems. The range of the scores is quite wide – from 53 to 72 F_1 on the native data and from 25 to 41 on non-native. Observe that the performance on the non-native data is substantially lower for all of the teams. This is to be expected as non-native writers exhibit a variety of errors – spelling, grammar, word choice. In contrast, the native data contains many punctuation and spelling mistakes that can be handled by MADAMIRA and are much easier to address (see also analysis below). In fact, we used MADAMIRA as a baseline system (last row in the tables). As the results show, MADAMIRA provides quite a competitive baseline, especially on the native data. But all of the teams managed to beat this baseline, in many cases by a large margin. This suggests that even though MADAMIRA is a sophisticated system, it cannot handle all of the errors, and the participating teams developed approaches that are complementary to it.

It is interesting to compare the obtained results to those obtained on similar shared tasks on English as a Second Language (ESL) writings. While the performance on native MSA data in Table 8is better than on ESL, performance on L2 writings is quite similar. For instance, the highest score in the HOO-2011 shared task (Dale and Kilgarriff, 2011) that addressed all errors was 21.1 (Rozovskaya et al., 2011); the highest performance in the CoNLL-2013 shared task that also used the same evaluation metric was 31.20 (Rozovskaya et al., 2013).[5]

In addition to providing the official rankings, we also analyze system performance for different types of mistakes by automatically assigning errors to one of the following categories: punctuation errors; errors involving *Alif* and *Ya*; and all other errors. Punctuation errors account for 39% of all errors in the Aljazeera data.[6] Tables 6 and 6 show the performance of the teams in three settings: with punctuation errors removed; with *Alif*/*Ya* errors removed; and when both punctuation and *Alif*/*Ya* errors are removed. In general, both for the native and the non-native data, performance drops when the *Alif*/*Ya* errors are removed, which indicates that these errors may be easier. When the punctuation errors are removed, the performance on the native data improves slightly, but goes down a little on the non-native data. Overall, it can be concluded that the punctuation mistakes do not significantly affect the performance and are of the same difficulty level as the remaining of the errors.

Finally, the majority of the teams participated last year and relied on the findings from the previous round. Overall, it can be said that the participants were able to make progress and to im-

[5]This is not a fair comparison, though, since the CoNLL-2013 shared task only evaluated on 5 types of errors and ignored about 50% of all mistakes in the data. In CoNLL-2014 that evaluated on all errors the top teams scored 35-37 points but the evaluation favored precision twice as much as recall.

[6]For example, there many sentences with missing final periods; we speculate that this may be due to the fact that the data was collected online.

Team	No punc. errors			No Alif/Ya errors			No punc. No Alif/Ya errors		
	P	R	F_1	P	R	F_1	P	R	F_1
ARIB-1	73.57	59.86	66.01	49.87	44.87	47.24	54.53	38.47	45.11
CUFE	85.80	77.98	81.70	84.25	43.29	57.19	80.12	58.24	**67.45**
GWU	81.12	76.60	78.79	61.15	52.32	56.39	67.80	54.86	60.65
QCMUQ	75.89	76.29	76.09	56.45	48.73	52.31	59.05	54.77	56.83
QCRI	81.28	75.62	78.35	75.90	36.52	49.31	69.78	51.68	59.38
SAHSOH	83.85	55.65	66.90	71.44	24.78	36.79	79.86	41.45	54.57
TECH-2	81.90	70.74	75.91	54.82	46.40	50.26	65.77	39.53	49.38
UMMU-1	82.98	80.98	**81.97**	56.46	58.09	**57.26**	73.09	61.44	66.76

Table 10: **Alj-test-2015**: Results on the test set in different settings: with punctuation errors removed from evaluation; normalization errors removed; and when both punctuation and normalization errors are removed. Only the best output from each team is shown.

Team	No punc. errors			No Alif/Ya errors			No punc. No Alif/Ya errors		
	P	R	F_1	P	R	F_1	P	R	F_1
ARIB-3	50.13	20.28	28.88	41.38	18.46	25.53	36.80	10.11	15.86
CUFE	65.05	28.43	39.57	65.68	16.46	26.32	58.28	17.83	27.31
GWU	54.39	22.60	31.93	45.27	15.63	23.24	38.28	10.76	16.79
QCMUQ	55.17	27.60	36.79	43.25	24.53	31.31	44.74	15.85	23.40
QCRI-1	42.71	25.82	32.18	32.88	11.46	17.00	28.51	13.51	18.34
SAHSOH	58.95	21.70	31.72	48.82	09.37	15.73	49.23	12.69	20.18
UMMU-1	57.32	30.49	**39.81**	47.45	26.15	**33.72**	48.79	18.98	**27.32**

Table 11: **L2-test-2015**: Results on the test set in different settings: with punctuation errors removed from evaluation; normalization errors removed; and when both punctuation and normalization errors are removed. Only the best output from each team is shown.

prove their systems since last year. Although direct comparison is not possible since the test sets are not the same and the test data from last year was used for development, we observe that four teams scored more than 70 F_1 points on the native data this year, while last year the best result that was obtained by the CLMB system (Rozovskaya et al., 2014) was 67.91 points. We refer the reader to the system description papers for more detail on how the respective systems have been improved.

7 Conclusion

This paper presented a report on QALB-2015, the second shared task on text correction of Arabic. QALB-2015 extended QALB-2014 that took place last year and focused on correcting texts written by native Arabic speakers. This year, we added a second track, on non-native data. We received 12 system submissions from eight teams. We are pleased with the extent of participation, the quality of results and the diversity of approaches. Many participants continued from last year and improved and extended their systems. We feel motivated to conduct new research competitions in the near future.

8 Acknowledgments

We would like to thank the organizing committee of ACL 2015 and its Arabic NLP workshop and also the shared task participants for their ideas and support. We thank Al Jazeera News (and especially, Khalid Judia) for providing the user comments portion of the QALB corpus. We also thank the QALB project annotators: Hoda Fathy, Dhoha Abid, Mariem Fekih, Anissa Jrad, Hoda Ibrahim, Noor Alzeer, Samah Lakhal, Jihene Wefi, Elsherif Mahmoud and Hossam El-Husseini. This publication was made possible by grant NPRP-4-1058-1-168 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Appendix A: Sample annotation file

The sequence of manual corrections for the example in Table 1 is shown below.

#1	مدي	A 2 3|||Edit|||مدى|||REQUIRED|||-NONE-|||0
#2	قرائة	A 5 6|||Edit|||قراءة|||REQUIRED|||-NONE-|||0
#3	هذة	A 6 7|||Edit|||هذه|||REQUIRED|||-NONE-|||0
#4	و المحترمة	A 9 11|||Merge|||والمحترمة|||REQUIRED|||-NONE-|||0
#5		A 11 11|||Add_before|||.|||REQUIRED|||-NONE-|||0
#6	لأاني	A 11 12|||Edit|||لأنني|||REQUIRED|||-NONE-|||0
#7	و كنت	A 13 15|||Merge|||وكنت|||REQUIRED|||-NONE-|||0
#8	بتمني	A 15 16|||Edit|||أتمنى|||REQUIRED|||-NONE-|||0
#9	ان	A 18 19|||Edit|||أن|||REQUIRED|||-NONE-|||0
#10	الاقصي	A 23 24|||Edit|||الأقصى|||REQUIRED|||-NONE-|||0
#11		A 24 24|||Add_before|||،|||REQUIRED|||-NONE-|||0
#12	و كان	A 24 26|||Merge|||وكان|||REQUIRED|||-NONE-|||0
#13	يبدوا	A 26 27|||Edit|||يبدو|||REQUIRED|||-NONE-|||0
#14	ان	A 27 28|||Edit|||أن|||REQUIRED|||-NONE-|||0
#15		A 31 31|||Add_before|||،|||REQUIRED|||-NONE-|||0
#16	ما	A 32 33|||Delete||||||REQUIRED|||-NONE-|||0
#17	في	A 33 34|||Delete||||||REQUIRED|||-NONE-|||0
#18	حد	A 34 35|||Edit|||واحد|||REQUIRED|||-NONE-|||0
#19	الامنية	A 36 37|||Edit|||الأمنية|||REQUIRED|||-NONE-|||0
#20	بيقول	A 38 39|||Edit|||يقول|||REQUIRED|||-NONE-|||0
#21	انك	A 39 40|||Edit|||أنك|||REQUIRED|||-NONE-|||0
#22		A 41 41|||Add_before|||أن|||REQUIRED|||-NONE-|||0
#23	تتمني	A 41 42|||Edit|||تتمنى|||REQUIRED|||-NONE-|||0
#24	ان	A 42 43|||Edit|||أن|||REQUIRED|||-NONE-|||0
#25	يحققوهالأن	A 45 46|||Split|||يحققوها لأن|||REQUIRED|||-NONE-|||0
#26	امنيتك	A 46 47|||Edit|||أمنيتك|||REQUIRED|||-NONE-|||0

References

A. Alfaifi and E. Atwell. 2012. Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors. In *The 8th International Computing Conference in Arabic*.

N. AlShenaifi, R. AlNefie, M. Al-Yahya, and H. Al-Khalifa. 2015. ARIB@QALB-2015 Shared Task: A Hybrid Cascade Model for Arabic Spelling Error Detection and Correction . In *Proceedings of ACL Workshop on Arabic Natural Language Processing*, Beijing, China, July.

M. Attia, P. Pecina, Y. Samih, K. Shaalan, and J. van Genabith. 2012. Improved Spelling Error Detection and Correction for Arabic. In *Proceedings of COLING*.

M. Attia, M. Al-Badrashiny, and M. Diab. 2015. GWU-HASP-2015: Priming Spelling Candidates with Probability . In *Proceedings of ACL Workshop on Arabic Natural Language Processing*, Beijing, China, July.

H. Bouamor, H. Sajjad, N. Durrani, and K. Oflazer. 2015. QCMUQ@QALB-2015 Shared Task: Combining Character level MT and Error-tolerant Finite-State Recognition for Arabic Spelling Correction. In *Proceedings of ACL Workshop on Arabic Natural Language Processing*, Beijing, China, July.

F. Bougares and H. Bouamor. 2015. UMMU@QALB-2015 Shared Task: Character and Word level SMT pipeline for Automatic Error Correction of Arabic Text. In *Proceedings of ACL Workshop on Arabic Natural Language Processing*, Beijing, China, July.

T. Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0.

D. Dahlmeier and H. T. Ng. 2012. Better Evaluation for Grammatical Error Correction. In *Proceedings of NAACL*.

R. Dale and A. Kilgarriff. 2011. Helping Our Own: The HOO 2011 Pilot Shared Task. In *Proceedings of the 13th European Workshop on Natural Language Generation*.

R. Dale, I. Anisimoff, and G. Narroway. 2012. A Report on the Preposition and Determiner Error Correction Shared Task. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*.

A. El Kholy and N. Habash. 2012. Orthographic and morphological processing for English–Arabic statistical machine translation. *Machine Translation*, 26(1-2).

S. Farwaneh and M. Tamimi. 2012. Arabic Learners Written Corpus: A Resource for Research and Learning. *The Center for Educational Resources in Culture, Language and Literacy.*

N. Habash, A. Soudi, and T. Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

N. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

A. Hassan, S. Noeman, and H. Hassan. 2008. Language Independent Text Correction using Finite State Automata. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 913–918, Hyderabad, India.

B. Mohit, A. Rozovskaya, N. Habash, W. Zaghouani, and O. 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.

D. Mostefa, J. Abualasal, O. Asbayou, M. Gzawi, and R. Abbeès. 2015. TECHLIMED@QALB-Shared Task 2015: a hybrid Arabic Error Correction System. In *Proceedings of ACL Workshop on Arabic Natural Language Processing*, Beijing, China, July.

H. Mubarak, K. Darwish, and A. Abdelali. 2015. QCRI@QALB-2015 Shared Task: Correction of Arabic Text for Native and Non-Native Speakers' Errors . In *Proceedings of ACL Workshop on Arabic Natural Language Processing*, Beijing, China, July.

M. Nawar. 2015. QALB 2015 Shared Task: CUFE Arabic Error Correction System. In *Proceedings of ACL Workshop on Arabic Natural Language Processing*, Beijing, China, July.

H. T. Ng, S. M. Wu, Y. Wu, Ch. Hadiwinoto, and J. Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of CoNLL: Shared Task*.

H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of CoNLL: Shared Task*.

O. Obeid, W. Zaghouani, B. Mohit, N. Habash, K. Oflazer, and N. Tomeh. 2013. A Web-based Annotation Framework For Large-Scale Text Correction. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*. Asian Federation of Natural Language Processing.

R. Parker, D. Graff, K. Chen, J. Kong, and K. Maeda. 2009. Arabic Gigaword Fourth Edition. LDC Catalog No.: LDC2009T30, ISBN: 1-58563-532-4.

A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

A. Rozovskaya, M. Sammons, J. Gioja, and D. Roth. 2011. University of Illinois System in HOO Text Correction Shared Task. In *Proceedings of the European Workshop on Natural Language Generation (ENLG)*.

A. Rozovskaya, K.-W. Chang, M. Sammons, and D. Roth. 2013. The University of Illinois System in the CoNLL-2013 Shared Task. In *Proceedings of CoNLL Shared Task*.

A. Rozovskaya, N. Habash, R. Eskander, N. Farra, and W. Salloum. 2014. The Columbia System in the QALB-2014 Shared Task on Arabic Error Correction. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing: QALB Shared Task*.

K. Shaalan, A. Allam, and A. Gomah. 2003. Towards Automatic Spell Checking for Arabic. In *Proceedings of the 4th Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE)*, Cairo, Egypt.

W. Zaghouani, B. Mohit, N. Habash, O. Obeid, N. Tomeh, A. Rozovskaya, N. Farra, S. Alkuhlani, and K. 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.

W. Zaghouani, N. Habash, H. Bouamor, A. Rozovskaya, Behrang B. Mohit, A. Heider, and K. Oflazer. 2015a. 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.

W. Zaghouani, T. Zerrouki, and A. Balla. 2015b. SAHSOH@QALB-2015 Shared Task: A Rule-Based Correction Method of Common Arabic Native and Non-Native Speakers' Errors. In *Proceedings of ACL Workshop on Arabic Natural Language Processing*, Beijing, China, July.