The First QALB Shared Task on Automatic Text Correction for Arabic

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

[1]Carnegie Mellon University in Qatar
[2]Center for Computational Learning Systems, Columbia University
[3]New York University Abu Dhabi
behrang@cmu.edu, alla@ccls.columbia.edu, nizar.habash@nyu.edu
wajdiz@qatar.cmu.edu,owo@qatar.cmu.edu

Abstract

We present a summary of the first shared task on automatic text correction for Arabic text. The shared task received 18 systems submissions from nine teams in six countries and represented a diversity of approaches. Our report includes an overview of the QALB corpus which was the source of the datasets 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 gained a lot of attention in the Natural Language Processing (NLP) community. 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 who focus on this problem and promote development and dissemination of key resources, such as benchmark datasets.

Recently, there have been several efforts aimed at creating data resources related to the correction of Arabic text. Those include human annotated corpora (Zaghouani et al., 2014; Alfaifi and Atwell, 2012), spell-checking lexicon (Attia et al., 2012) and unannotated language learner corpora (Farwaneh and Tamimi, 2012). A natural extension to these resource production efforts is the creation of robust automatic systems for error correction.

In this paper, we present a summary of the QALB shared task on automatic text correction for Arabic. The Qatar Arabic Language Bank (QALB) project[1] is one of the first large scale data and system development efforts for automatic correction of Arabic which has resulted in annotation of the QALB corpus. In conjunction with the EMNLP Arabic NLP workshop, the QALB shared task is the first community effort for construction and evaluation of automatic correction systems for Arabic.

The results of the competition indicate that the shared task attracted a lot of interest and generated a diverse set of approaches from the participating teams.

In the next section, 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. Finally, in Section 7, we offer a brief analysis and present preliminary experiments on system combination.

2 Task Description

The QALB shared task was created as a forum for competition and collaboration on automatic error correction in Modern Standard Arabic. The shared task makes use of the QALB corpus (Zaghouani et al., 2014), which is a manually-corrected collection of Arabic texts. 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 devel-

* These authors contributed equally to this work.

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

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 39–47,
October 25, 2014, Doha, Qatar.

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	mdy مدي	mdý مدى	Ya/Alif-Maqsura Spelling	Edit
#9	An ان	Ân أن	Alif-Hamza Spelling	Edit
#11	*Missing Comma*	،	Punctuation	Add_before
#12	w kAn و كان	wkAn وكان	Extra Space	Merge
#13	ybdwA يبدوا	ybdw يبدو	Morphology	Edit
#20	byqwl بيقول	yqwl يقول	Dialectal	Edit
#25	yHqqwhAlÂn يحققوهالأن	yHqqwhA lÂn يحققوها لأن	Missing Space	Split

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

oped for similar error correction competitions is 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 set.

After the initial registration, the participants were provided with training and development sets and evaluation scripts. During the test period, the teams were given test data on which they needed to run their systems. Following the announcement of system results, the answer key to the test set was released. Participants authored description papers which will be presented in the Arabic NLP workshop.

3 The QALB Corpus

One of the goals of the QALB project is to create a large manually corrected corpus of errors for a variety of Arabic texts, including user comments on news web sites, native and non-native speaker essays, and 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).

The annotation process includes an initial automatic pre-processing step followed by an automatic correction of common spelling errors by the

morphological analysis and disambiguation system MADA (v3.2) (Habash and Rambow, 2005; Habash et al., 2009). The pre-processed files are then assigned to a team of expert (human) annotators.

For a given sentence, the annotators were instructed to correct all errors; these include spelling, punctuation, word choice, morphology, syntax, and dialectal usage. It should be noted 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 erroneous 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 from 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.

To ensure high annotation quality, the annotators went through multiple phases of training; the inter-annotator agreement was reviewed routinely. Zaghouani et al. (2014) report an average Word Error Rate (WER) of 3.8% for all words (excluding punctuation), which is quite high. When punctuation was included, the WER rose to 11.3%. The high level of agreement indicates that the annotations are reliable and the guidelines are useful in producing homogeneous and consistent data. Punctuation, however, remains a challenge.

4 Shared Task Data

The shared task data comes from the QALB corpus and consists of user comments from the Aljazeera News webpage written in Modern Standard Arabic.

Comments belonging to the same article were included only in one of the shared task subsets (i.e., training, development or test). Furthermore, we split the data by the annotation time. Consequently, the training data is comprised of comments annotated between June and December, 2013; the development data includes texts annotated in December 2013; the test data includes documents annotated in the Spring of 2014.

Statistics	Train.	Dev.	Test
Number of docs.	19,411	1,017	968
Number of words	1M	54K	51K
Number of errors	306K	16K	16K

Table 3: Statistics on the shared task data.

We refer to each comment in the shared task data as *document* and assign it a special ID that indicates the ID of the article to which the comment refers and the comment's number.

The data was made available to the participants in three versions: (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 produces forms that correct 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).

Table 3 presents statistics on the shared task data. The training data contains over one million words of text; the development and test data contain slightly over 50,000 words each. Table 4 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 several words to be merged together (about a third of all errors).

[2] In the shared task, we specified two *Add* categories: *addBefore* and *addAfter*. 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) *AbtθjHxdðrzsšSDTĎʕγfqklmnhwy* and the additional symbols: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ, ŷ ئ, ħ ة, ý ى.

Data	Error type (%)						
	Edit	**Add**	**Merge**	**Split**	**Delete**	**Move**	**Other**
Train.	55.34	32.36	5.95	3.48	2.21	0.14	0.50
Dev.	53.51	34.24	5.97	3.67	2.03	0.08	0.49
Test	51.94	34.73	5.89	3.48	3.32	0.15	0.49

Table 4: Distribution of annotations by type in the shared task data. Error types denotes the action required in order to correct the error.

Team Name	Affiliation
CLMB (Rozovskaya et al., 2014)	Columbia University (USA)
CMUQ (Jeblee et al., 2014)	Carnegie Mellon University in Qatar (Qatar)
CP13 (Tomeh et al., 2014)	Université Paris 13 (France)
CUFE (Nawar and Ragheb, 2014)	Computer Engineering Department, Cairo University (Egypt)
GLTW (Zerrouki et al., 2014)	Bouira University (Algeria), The National Computer Science Engineering School (Algeria), and Tabuk University (KSA)
GWU (Attia et al., 2014)	George Washington University (USA)
QCRI (Mubarak and Darwish, 2014)	Qatar Computing Research Institute (Qatar)
TECH (Mostefa et al., 2014)	Techlimed.com (France)
YAM (Hassan et al., 2014)	Faculty of Engineering, Cairo University (Egypt)

Table 5: List of the nine teams that participated in the shared task.

Team	Approach	External Resources
CLMB	Corrections proposed by MADAMIRA; a Maximum Likelihood model trained on the training data; regular expressions; a decision-tree classifier for punctuation errors trained on the training data; an SVM character-level error correction model; a Naïve Bayes classifier trained on the training data and the Arabic Gigaword corpus	Arabic Gigaword Fourth Edition (Parker et al., 2009)
CMUQ	A pipeline consisting of rules, corrections proposed by MADAMIRA, a language model for spelling mistakes, and a statistical machine-translation system	AraComLex dictionary (Attia et al., 2012)
CP13	A pipeline that consists of an error detection SVM classifier that uses MADAMIRA features and language model scores; a character-level back-off correction model implemented as a weighted finite-state transducer; a statistical machine-translation system; a discriminative re-ranker; a decision tree classifier for inserting missing punctuation	None
CUFE	Rules extracted from the Buckwalter morphological analyser; their probabilities are learned using the training data	Buckwalter morphological analyzer Version 2.0 (Buckwalter, 2004)
GLTW	Regular expressions and word lists	AraComLex dictionary (Attia et al., 2012); in-house resources; Ayaspell dictionary
GWU	A CRF model for punctuation errors; a dictionary and a language model for spelling errors; normalization rules	AraComLex Extended dictionary (Attia et al., 2012); Arabic Gigaword Fourth Edition (Parker et al., 2009)
QCRI	Word errors: a language model trained on Arabic Wikipedia and Aljazeera data; punctuation mistakes: a CRF model and a frequency-based model trained on the shared task data	Arabic Wikipedia; Aljazeera articles
TECH	Off-the-shelf spell checkers and a statistical machine-translation model	Newspaper articles from Open Source Arabic Corpora; other corpora collected online; Hunspell
YAM	*Edit* errors: a Naïve Bayes classifier that uses the following features: a character confusion matrix based on the training data; a collocation model that uses the target lemma and the surrounding POS tags; a co-occurrence model that uses lemmata of the surrounding words; *Split* and *Merge* errors: a language model trained on the training data; *Add* errors: a classifier	AraComLex dictionary (Attia et al., 2012); Buckwalter Analyzer Version 1.0 (Buckwalter, 2002); Arabic stoplists

Table 6: Approaches adopted by the participating teams.

5 Participants and Approaches

Nine teams from six countries participated in the shared task. Table 5 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 eighteen outputs. 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 a variety of techniques and attempted to classify the errors in some way using that classification in developing their systems: the CLMB system combined machine-learning modules with rules and MADAMIRA corrections; the CUFE system extracted rules from the morphological analyzer and learned their probabilities using the training data; and the CMUQ system combined statistical machine-translation with a language model, rules, and MADAMIRA corrections. Table 6 summarizes the approaches adopted by each team.

6 Results

In this section, we present the results of the competition. For evaluation, we adopted the standard Precision (P), Recall (R), and F_1 metric that was 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.

Table 7 presents the official results of the evaluation on the test set. The results are sorted according to the F_1 scores obtained by the systems. The range of the scores is quite wide – from 20 to 67 F_1 – but the the majority of the systems stay in the 50-60 range.

It is interesting to note that these results are considerably higher than those shown on the similar shared tasks on English non-native data. For instance, the highest performance in the CoNLL-2013 shared task that also used the same evaluation metric was 31.20 (Rozovskaya et al., 2013).[4] The highest score in the HOO-2011 shared task (Dale and Kilgarriff, 2011) that addressed all errors was 21.1 (Rozovskaya et al., 2011). Of course, the setting was different,as we are dealing with texts written by native speakers. But, in addition to that, we hypothesize that our data contains a set of language-specific errors that may be "easier", e.g Alif/Ya errors.

We also asked the participants for the outputs of their systems on the development set. We show the results in Table 8. While these results are not used for ranking, since the development set was used for tuning the parameters of the systems, it is interesting to see how much the performance differs from the results obtained on the test. In general, we do not observe substantial differences in the performance and the rankings, with a few exceptions. In particular, CP13 submissions did much better on the development set, as well as the CUFE system: the CUFE system suffers a major drop in precision on the test set, while the CP13 systems lose in recall. For more details, we refer the reader to the system description papers.

In addition to the official rankings, it is also interesting to analyze system performance for different types of mistakes. Note that here we are not interested in the annotation classification by action type. Instead, we automatically assign 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 data[5] . Table 7 shows 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. Observe that when punctuation errors are not taken into account, the CUFE team gets the first ranking (for each the results of the best-performing system were chosen).

7 Analysis of System Output

We conducted a couple of experiments to analyze the task challenges and system errors.

The Most and Least Challenging Sentences We examined some of the most, and the least challenging parts of the test data for the shared task systems. To identify these subsets, we ranked all sentences using their average sentence-level F_1 score and selected the top and bottom 50 sentences. Our manual examination of these two

[4]This year CoNLL was an extension of the CoNLL-2013 competition for all errors but in its evaluation favored precision twice as much as recall, so we are not comparing to this setting.

[5]For example, there are a lot of missing periods at the end of a sentence that may be due to the fact that the data was collected online.

Rank	Team	P	R	F_1
1	CLMB-1	73.34	63.23	**67.91**
2	CLMB-2	70.86	62.21	66.25
3	CUFE-1	87.49	52.63	65.73
4	CMUQ-1	77.97	56.35	65.42
5	CLMB-3	71.45	60.00	65.22
6	QCRI-1	71.70	56.86	63.43
7	GWU-1	75.47	52.98	62.25
8	GWU-2	75.34	52.99	62.22
9	QCRI-2	62.86	60.32	61.57
10	YAM-1	63.52	57.61	60.42
11	QCRI-3	60.66	59.28	59.96
12	TECH-1	73.46	50.56	59.89
13	TECH-2	73.50	50.53	59.88
14	TECH-3	72.34	50.51	59.49
15	CP13-2	76.85	47.33	58.58
16	CP13-1	77.85	38.77	51.76
17	GLTW-1	75.15	23.15	35.40
18	GLTW-2	69.80	12.33	20.96

Table 7: Official results on the test set. Column 1 shows the system rank according to the F_1 score.

Rank (test)	Rank (dev)	Team	P	R	F_1
1	2	CLMB-1	72.22	62.79	67.18
2	3	CLMB-2	69.49	61.73	65.38
3	1	CUFE-1	94.11	53.74	**68.42**
4	4	CMUQ-1	76.17	56.59	64.94
5	5	CLMB-3	69.71	59.42	64.15
6	6	QCRI-1	70.83	57.34	63.38
7	9	GWU-1	73.15	53.18	61.59
8	10	GWU-2	73.01	53.13	61.50
9	8	QCRI-2	62.21	61.30	61.75
10	14	YAM-1	57.81	59.19	58.49
11	12	QCRI-3	60.47	60.65	60.56
12	13	TECH-1	70.86	50.04	58.66
13	15	TECH-2	70.66	49.65	58.32
14	16	TECH-3	70.65	48.83	57.75
15	7	CP13-2	74.85	54.15	62.84
16	11	CP13-1	75.73	51.33	61.19
17	17	GLTW-1	73.83	22.80	34.84
18	18	GLTW-2	67.85	11.09	19.06

Table 8: Results on the **development** set. Columns 1 and 2 show the rank of the system according to F_1 score obtained on the test set shown in Table 7 and the development set, respectively.

sets shows that the differences between them relate to both the density and the type of errors. The more challenging sentences (with the lowest system performance) contain more errors in general, and their errors tend to be complex and challenging, e.g., the correction of the erroneous two-token string أدت عت (Âdt ς t) requires a character deletion and a merge to produce ادعت (Adςt). In contrast the less challenging sentences tend to have fewer and simpler errors such as the common Alif/Ya errors.

System Combination We took the 18 systems' output and conducted two system combination experiments: (a) an oracle upper-bound estimation and (b) a simple majority vote system combination. For these experiments we isolated and evaluated each sentence output individually to form a new combined system output.

In the oracle experiment, we combined different systems by selecting the output of the best performing system for each individual sentence. For that, we evaluated sentences individually for each system and chose the system output with the highest F_1 score. The combined output holds the best output of all systems for the test set. This is an oracle system combination which allows us to estimate an upper-bound combination of all 18 systems.

In the majority vote experiment, we combined system output based on majority vote of various systems at sentence level. For every sentence, we choose the output that is agreed by most systems. If all systems have different output for a sentence, we back-off to the best performing system (CLMB-1).

Table 10 compares the results of these two experiments against the best performing system (CLMB-1). We observe a large boost of performance in the oracle experiment. This promising result reflects the complementary nature of the different methods that have been applied to the shared task, and it motivates further research on system combination. The result for the majority-vote system combination is very close to the CLMB-1's performance. This is not surprising; since, for 92% of sentences, there was no sentence-level agreement among systems. As a result, the combined system was very close to the back-off CLMB-1 system.

8 Conclusion

We have described the framework and results of the first shared task on automatic correction of Arabic, which used data from the QALB corpus. The shared task received 18 systems submissions

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
CLMB-1	82.63	72.50	77.24	64.05	50.86	**56.70**	76.99	49.91	60.56
CMUQ-1	82.89	68.69	75.12	68.32	40.51	50.86	74.25	41.46	53.21
CP13-2	80.51	59.97	68.74	65.09	28.00	39.16	68.67	25.34	37.02
CUFE-1	85.22	78.79	**81.88**	83.34	36.21	50.48	80.63	63.25	**70.89**
GLTW-1	65.18	34.84	45.41	48.52	15.29	23.26	49.25	26.78	34.70
GWU-1	76.28	64.17	69.70	64.67	39.61	49.13	59.07	41.48	48.74
QCRI-1	76.74	74.93	75.82	59.66	41.90	49.23	63.22	55.10	58.88
TECH-1	81.23	62.99	70.95	59.39	34.59	43.72	64.93	35.69	46.06
YAM-1	77.38	63.99	70.05	50.77	43.43	46.81	64.63	34.71	45.17

Table 9: 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.

System	Precision	Recall	F_1
Oracle	83.25	68.72	75.29
Majority-Vote	73.96	62.88	67.97
CLMB-1	73.34	63.23	67.91

Table 10: Comparing the best performing system with two experimental hybrid systems.

from nine teams in six countries. We are pleased with the extent of participation, the quality of results and the diversity of approaches. We plan to release the output of all systems. Such dataset and all the methods used in this shared task are expected to introduce new directions in automatic correction of Arabic. We feel motivated to extend the shared task's framework and text domain to conduct new research competitions in the near future.

9 Acknowledgments

We would like to thank the organizing committee of EMNLP-2014 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 grants NPRP-4-1058-1-168 and YSREP-1-018-1-004 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors. Nizar Habash performed most of his contribution to this paper while he was at the Center for Computational Learning Systems at Columbia University.

References

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

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. 2014. GWU-HASP: Hybrid Arabic Spelling and Punctuation Corrector. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing: QALB Shared Task*.

T. Buckwalter. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0.

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 and O. Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of ACL*.

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, O. Rambow, and R. Roth. 2009. MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*.

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

Y. Hassan, M. Aly, and A. Atiya. 2014. Arabic Spelling Correction using Supervised Learning. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing: QALB Shared Task*.

S. Jeblee, H. Bouamor, W. Zaghouani, and K. Oflazer. 2014. CMUQ@QALB-2014: An SMT-based System for Automatic Arabic Error Correction. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing: QALB Shared Task*.

D. Mostefa, O. Asbayou, and R. Abbes. 2014. TECH-LIMED System Description for the Shared Task on Automatic Arabic Error Correction. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing: QALB Shared Task*.

H. Mubarak and K. Darwish. 2014. Automatic Correction of Arabic Text: a Cascaded Approach. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing: QALB Shared Task*.

M. Nawar and M. Ragheb. 2014. Fast and Robust Arabic Error Correction System. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing: QALB Shared Task*.

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*.

N. Tomeh, N. Habash, R. Eskander, and J. Le Roux. 2014. A Pipeline Approach to Supervised Error Correction for the QALB-2014 Shared Task. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing: QALB Shared Task*.

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)*. European Language Resources Association (ELRA).

T. Zerrouki, K. Alhawiti, and A. Balla. 2014. Autocorrection Of Arabic Common Errors For Large Text Corpus. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing: QALB Shared Task*.

Appendix A: Sample annotation file

Below is the complete list of correction actions for the example in Table 1 as they appear in the training and evaluation data. The first two columns are the error index linking to Table 1 and the original word, respectively. Only the column titled Correction Action is in the training and evaluation data. The two numbers following the `A` specify the start and end positions of the sentence token string to change. Following that (and delimited by `|||`) are the action type and the correction string. The last three fields are irrelevant to this discussion.

Error Index	Original Word	Correction Action
#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