QALB: Qatar Arabic Language Bank

Nizar Habash¹, Behrang Mohit², Ossama Obeid², Kemal Oflazér², Nadi Tomeh³ and Wajdi Zaghouni²

¹Columbia University, ²Carnegie Mellon University Qatar, ³Université Paris 13

habash@ccls.columbia.edu, behrang@cmu.edu, oobeid@cmu.edu, ko@cs.cmu.edu
nadi.tomeh@lipn.univ-paris13.fr, wajdiz@cmu.edu

The Problem
Standard Arabic is a language with rich morphology and a complex grammar.

→ Many Arabic speakers make mistakes when spontaneously writing Arabic.
→ Arabic output of machine translation has numerous grammar errors.

Our Solution
QALB
Qatar Arabic Language Bank

Build a corpus of Arabic language errors with manually annotated corrections.

• Target size is 2 million words. The largest such corpus for Arabic.
• Native Arabic text is primarily Aljazeera comments
• Collected essays of non-native writers
• Machine translation of English Wikipedia pages into Arabic

ACLE
Automatic Correction of Language Errors

Develop models of automatic correction of Arabic language errors using the QALB corpus.

• Unsupervised models for error detection and correction are evaluated against QALB corpus.
• Supervised models of error correction are trained and evaluated using QALB corpus.

Status of the Project (Year 1)

QALB Annotation Guidelines
• Comprehensive guidelines covering errors of spelling, punctuation, lexical choice, morphology and syntax.
• A large number of examples to train annotators.

QAWI: QALB Annotation Web Interface
• An intuitive web-based annotation tool
• Easy administration of QALB’s large-scale annotations
• Complete record of all annotator actions
• Facilities for evaluating inter-annotator agreement
• Integrated automated annotators: MADA, a system for Morphological Analysis and Disambiguation of Arabic

QALB Annotation
• We have trained a group of eight annotators.
• We have so far annotated 600,000 words.

Next Steps
• We plan to reach our annotation goal of 2 million words in a year and a half.
• We started developing supervised ACLE solutions using QALB.
• We will host an international competition on Arabic text correction as part of the Arabic Natural Language Processing Workshop. The workshop is in the conference of Empirical Methods for Natural Language Processing – EMNLP 2014 in Doha.

Acknowledgment: This research was supported by Qatar National Research Fund (QNRF), NPRP grant 4-1058-1-168

English
Ten farm owners bought twenty-one birds.

MT output
Ten owners the farms bought (sing) roosters (plur) and twenty the birds.

Edit
Ten owners the farms bought (sing) roosters (plur) and twenty the birds.

Our challenge: How to fix these errors automatically?