QCN System Description for NIST OpenMT15

Hassan Sajjad, Nadir Durrani, Francisco Guzman, Preslav Nakov, Ahmed Abdelali, Stephan Vogel
{hsajjad, ndurrani, fguzman, pnakov, aabdelali, svogel}@qf.org.qa
Qatar Computing Research Institute

Wael Salloum, Ahmed El Kholy
{wss2113, ame2127}@columbia.edu
Columbia University

Nizar Habash
nizar.habash@nyu.edu
New York University Abu Dhabi

Abstract

This document describes Qatar-Columbia-New York Submission of Arabic-to-English systems for NIST OpenMT15. We trained a phrase-based SMT system using state-of-the-art features such as sparse features, operation sequence models, class-based models, joint neural network model, neural reranking, and unsupervised transliteration mining. We additionally tried phrase-table merging and an MEMT-based system combination was performed. The data was processed using Aarib and MADA-Mira tools.

1. Site affiliation

Qatar Computing Research Institute (QCRI)
Columbia University (CU)
New York University Abu Dhabi (NYUAD)

2. Submissions

NIST_ara2eng_cn_primary
NIST_ara2eng_cn_contrastive1

3. Primary system specs

We tune a separate system for each type of input: SMS, CTB, and CTS

3.1 Core MT engine algorithmic approach

Phrase-based Statistical Machine Translation

3.2 Critical additional features and tools used

Phrase-based Decoder (Moses)
Class-based Models
Operation Sequence Model
Joint Neural Network Language Model
Sparse Features
Phrase Table Merging
Lexicalized Reordering
Unsupervised Transliteration Model
Interpolated Models
Pair-wise Neural Re-ranking

Pair-wise Ranked Optimization (PRO-Fix)

3.3 Significant data pre/post-Processing

Egyptian Tokenization (ATB, S2, D3) (MADAMIRA)
Arabizi to Arabic Script (3arrib Tool)
MSA Tokenization (MADA)
Normalization (Elongation Removal, Emoticons)

3.4 Other data used (outside the LDC training data)

None

4. Key differences in contrastive systems

The primary systems combine ATB, D3 and S2 segmentations, while the contrastive systems only used ATB or D3 segmentations

References

Al-Badrashiny, Mohamed, Ramy Eskander, Nizar Habash and Owen Rambow. Automatic Transliteration of Romanized Dialectal Arabic. In Proceedings of the Conference on Computational Natural Language Learning (CONLL), Baltimore, Maryland, 2014.

Devlin, J., R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, Fast and robust neural network joint models for statistical machine translation. In Proceedings of the Association for Computational Linguistics, Baltimore, 2014.

Durrani, Nadir, Philipp Koehn, Helmut Schmid, Alexander Fraser (2014). Investigating the Usefulness of Generalized Word Representations in SMT. In Proceedings of the 25th Annual Conference on Computational Linguistics (COLING). Dublin, Ireland. August

Durrani, Nadir, Hassan Sajjad, Hieu Hoang, Philipp Koehn (2014). Integrating an Unsupervised Transliteration Model into Statistical Machine Translation. In Proceedings of Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden.

Durrani, Nadir, Helmut Schmid, Alexander Fraser, (2011). A Joint Sequence Translation Model with Integrated Reordering. In Proceedings of the Association for Computational Linguistics, Portland, Oregon, USA.

Durrani Nadir, Haddow Barry, Koehn Philipp, and Heafield Kenneth (2014). Edinburgh’s Phrase-based Machine Translation Systems for WMT-14. In Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA

Habash, Nizar and Fatiha Sadat. Arabic Preprocessing Schemes for Statistical Machine Translation, In Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL), New York, 2006.

Habash, Nizar and Owen Rambow. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In Proceedings of the Conference of American Association for Computational Linguistics, 2005.

Hasler, E., B. Haddow, and P. Koehn, “Sparse Lexicalised Features and Topic Adaptation for SMT,” in Proc. of the Int. Workshop on Spoken Language Translation(IWSLT), Hong Kong, Dec. 2012, pp. 268–275.

Heafield, Kenneth, and Alon Lavie. "Combining Machine Translation Output with Open Source: The Carnegie Mellon Multi-Engine Machine Translation Scheme." The Prague Bulletin of Mathematical Linguistics 93 (2010): 27-36.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. J. Dyer, O. Bojar, A. Constantin,and E. Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation,” in Proceedings of the Association for Computational Linguistics, Demo and Poster Sessions, Prague, Czech Republic, 2007.

Nakov, Preslav, Francisco Guzman, and Stephan Vogel. "Optimizing for Sentence-Level BLEU+1 Yields Short Translations." COLING. 2012.

Pasha, Arfath, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow and Ryan M. Roth. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2014.