Morphological Analysis and Disambiguation for Dialectal Arabic

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh
Center for Computational Learning Systems
Columbia University
{habash,ryanr,rambow,reskander,nadi}@ccls.columbia.edu

Abstract

The many differences between Dialectal Arabic and Modern Standard Arabic (MSA) pose a challenge to the majority of Arabic natural language processing tools, which are designed for MSA. In this paper, we retarget an existing state-of-the-art MSA morphological tagger to Egyptian Arabic (ARZ). Our evaluation demonstrates that our ARZ morphology tagger outperforms its MSA variant on ARZ input in terms of accuracy in part-of-speech tagging, diacritization, lemmatization and tokenization; and in terms of utility for ARZ-to-English statistical machine translation.

1 Introduction

Dialectal Arabic (DA) refers to the day-to-day native vernaculars spoken in the Arab World. DA is used side by side with Modern Standard Arabic (MSA), the official language of the media and education (Holes, 2004). Although DAs are historically related to MSA, there are many phonological, morphological and lexical differences between them. Unlike MSA, DAs have no standard orthographies or language academies. Furthermore, different DAs, such as Egyptian Arabic (henceforth, ARZ), Levantine Arabic or Moroccan Arabic have important differences among them, similar to those seen among Romance languages (Holes, 2004; Abdel-Massih et al., 1979). Most tools and resources developed for natural language processing (NLP) of Arabic are designed for MSA. Such resources are quite limited when it comes to processing DA, e.g., a state-of-the-art MSA morphological analyzer only has 60% coverage of Levantine Arabic verb forms (Habash and Rambow, 2006).

In this paper, we describe the process of retargeting an existing state-of-the-art tool for modeling MSA morphology disambiguation to ARZ, the most commonly spoken DA. The MSA tool we extend is MADA – Morphological Analysis and Disambiguation of Arabic (Habash and Rambow, 2005). The approach used in MADA, which was inspired by earlier work by Hajič (2000), disambiguates in context for every aspect of Arabic morphology, thus solving all tasks in "one fell swoop". The disadvantage of the MADA approach is its dependence on two complex resources: a morphological analyzer for the language and a large collection of manually annotated words for all morphological features in the same representation used by the analyzer. For ARZ, such resources have recently become available, with the development of the CALIMA ARZ morphological analyzer (Habash et al., 2012b) and the release by the Linguistic Data Consortium (LDC) of a large ARZ corpus annotated morphologically in a manner compatible with CALIMA (Maamouri et al., 2012a). In the work presented here, we utilize these new resources within the paradigm of MADA, transforming MADA into MADA-ARZ. The elegance of the MADA solution makes this conceptually a simple extension.

Our evaluation demonstrates that our Egyptian DA version of MADA, henceforth MADA-ARZ, outperforms MADA for MSA on ARZ morphological tagging and improves the quality of ARZ to English statistical machine translation (MT).

The rest of this paper is structured as follows: Section 2 discusses related work. Section 3 presents the challenges of processing Arabic dialects. Section 4 outlines our approach. And Section 5 presents and discusses our evaluation results.

Proceedings of NAACL-HLT 2013, pages 426–432,
Atlanta, Georgia, 9–14 June 2013.

2 Related Work

There has been a considerable amount of work on MSA morphological analysis, disambiguation, part-of-speech (POS) tagging, tokenization, lemmatization and diacritization; for an overview, see (Habash, 2010). Most solutions target specific problems, such as diacritization (Zitouni et al., 2006), tokenization or POS tagging (Diab et al., 2007). In contrast, MADA provides a solution to all of these problems together (Habash and Rambow, 2005).

Previous work on DA morphological tagging focused on creating resources, using noisy or incomplete annotations, and using unsupervised/semi-supervised methods. Duh and Kirchhoff (2005) adopt a minimally supervised approach that only requires raw text data from several DAs, as well as a MSA morphological analyzer. They report a POS accuracy of 70.9% on a rather coarse-grained POS tagset (17 tags).

Al-Sabbagh and Girju (2012) describe a supervised tagger for Egyptian Arabic social networking corpora trained using transformation-based learning (Brill, 1995). They report 94.5% F-measure on tokenization and 87.6% on POS tagging. Their tokenization and POS tagsets are comparable to the set used by the Arabic Treebank (ATB). We do not compare to them since their data sets are not public.

Stallard et al. (2012) show that unsupervised methods for learning DA tokenization can outperform MSA tokenizers on MT from Levantine Arabic to English. We do not compare to them directly since our work is on ARZ. However, we carry a similar MT experiment in Section 5.

Mohamed et al. (2012) annotated a small corpus of Egyptian Arabic for morphological segmentation and learned segmentation models using memory-based learning (Daelemans and van den Bosch, 2005). Their best system achieves a 91.90% accuracy on the task of morpheme-segmentation. We compare to their work and report on their test set in Section 5.

There are some other morphological analyzers for DA. Kilany et al. (2002) worked on ARZ, but the analyzer has very limited coverage. Their lexicon was used as part of the development of CALIMA (Habash et al., 2012b). Other efforts are not about ARZ (Habash and Rambow, 2006; Salloum and Habash, 2011).

Given the similarity between MSA and DA, there has been some work on mapping DA to MSA to exploit rich MSA resources (Chiang et al., 2006; Abo Bakr et al., 2008; Salloum and Habash, 2011; Salloum and Habash, 2013). Other researchers have studied the value of simply combining DA and MSA data, such as Zbib et al. (2012) for DA to English MT. In our approach, we target DA directly, and we evaluate the use of additional MSA annotated resources to our training in Section 5.

3 Arabic Dialect Challenges

General Arabic Challenges Arabic, as MSA or DA, poses many challenges for NLP. Arabic is a morphologically complex language which includes rich inflectional morphology and a number of clitics. For example, the MSA word وسيكتبونها *wsyktbwnhA (wa+sa+ya-ktub-uwna+hA)*[1] 'and they will write it [lit. and+will+they-write-they+it]' has two proclitics, one circumfix and one pronominal enclitic. Additionally, Arabic has a high degree of ambiguity resulting from its diacritic-optional writing system and common deviation from spelling standards (e.g., Alif and Ya variants) (Buckwalter, 2007). The Standard Arabic Morphological Analyzer for (SAMA) (Graff et al., 2009) produces 12 analyses per MSA word on average.

Differences between ARZ **and** MSA As mentioned above, most tools developed for MSA cannot be expected to perform well on ARZ. This is due to the numerous differences between the two variants. Lexically, the number of differences is quite significant. For example, ARZ طربيزة *Trbyzħ* 'table' corresponds to MSA طاولة *TAwlħ*. Phonologically, there are many important differences which relate to orthography in DA, e.g., the MSA consonant ث /*θ*/ is pronounced as /t/ in ARZ (or /s/ in more recent borrowings from MSA); for a fuller discussion, see (Habash, 2010; Habash et al., 2012a). Examples of morphological differences include changes in the

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

morpheme form, e.g., the MSA future proclitic +س *sa+* appears in ARZ as +ه *ha+*. There are some morphemes in ARZ that do not exist in MSA such as the negation circum-clitic ما+ … +ش *mA+ … +š*. And there are MSA features that are absent from ARZ, most notably case and mood.

Since there are no orthographic standards, ARZ words may be written in a variety of ways reflecting different writing rules, e.g., phonologically or etymologically. A conventional orthography for Dialectal Arabic (CODA) has been proposed and used for writing ARZ in the context of NLP applications (Habash et al., 2012a; Al-Sabbagh and Girju, 2012; Eskander et al., 2013). Finally, MSA and ARZ coexist and are often used interchangeably, especially in more formal settings. The CALIMA morphological analyzer we use addresses several of these issues by modeling both ARZ and MSA together, including a limited set of inter-dialect morphology phenomena, and by mapping ARZ words into CODA orthography internally while accepting a wide range of spelling variants.

4 Approach

4.1 The MADA Approach

MADA is a method for Arabic morphological analysis and disambiguation (Habash and Rambow, 2005; Roth et al., 2008). MADA uses a morphological analyzer to produce, for each input word, a list of analyses specifying every possible morphological interpretation of that word, covering all morphological features of the word (diacritization, POS, lemma, and 13 inflectional and clitic features). MADA then applies a set of models (support vector machines and N-gram language models) to produce a prediction, per word in-context, for different morphological features, such as POS, lemma, gender, number or person. A ranking component scores the analyses produced by the morphological analyzer using a tuned weighted sum of matches with the predicted features. The top-scoring analysis is chosen as the predicted interpretation for that word in context.

4.2 Extending MADA into MADA-ARZ

Adjusting MADA to handle DA requires a number of modifications. The most significant change is replacing the MSA analyzer SAMA with the ARZ analyzer CALIMA to address the differences outlined in Section 3. In addition, new feature prediction models are needed; these are trained using ARZ data sets annotated by the LDC (Maamouri et al., 2006; Maamouri et al., 2012b). The data sets were not usable as released due to numerous annotation inconsistencies and differences from CALIMA, as well due to gaps in CALIMA. We synchronized the annotations with the latest version of CALIMA following a technique described by Habash and Rambow (2005). The result of this synchronization step is the data we use in this study (for training, development and testing). Our synchronized annotations fully match the LDC annotations in 90% of the words (in full morphological tag). We performed a manual analysis on randomly chosen 100 words that did not fully match. The choice we made is correct or acceptable in 55% of the cases of mismatch with the LDC annotation, which means that the our choice is accurate in over 95% of all cases.

Some of the original MADA features (which were needed for MSA) are not used in ARZ and so are dropped in MADA-ARZ; these features are case, mood, the question-marking proclitic, state and voice. Additional ARZ feature values have been added, e.g., to handle the progressive particle and future marker, among others. These are provided by CALIMA and are classified and selected by MADA-ARZ. In our current implementation, ARZ features that are not present in MSA, such as the negation and indirect-object enclitics, are not classified by MADA-ARZ classifiers, but since they are provided by CALIMA they can be selected by the whole MADA-ARZ system.

5 Evaluation

We evaluate MADA-ARZ intrinsically — in terms of performance on morphological disambiguation — and extrinsically in the context of MT.

5.1 POS Tagging, Diacritization, Lemmatization and Segmentation

Experimental Settings We use two sets of annotated data from the LDC: ATB-123, which includes parts 1, 2 and 3 of the MSA Penn Arabic Treebank

Train Data	Development			Test		
	MADA	MADA-ARZ		MADA	MADA-ARZ	
	MSA	ARZ	ALL	MSA	ARZ	ALL
Morph Tag	35.8	**84.0**	77.3	35.7	**84.5**	75.5
Penn POS	77.5	89.6	**90.2**	79.0	90.0	**90.1**
MADA POS	80.7	90.8	**91.3**	82.1	91.1	**91.4**
Diacritic	31.3	**82.6**	72.9	32.2	**83.2**	72.2
Lemma	64.0	**85.2**	81.6	67.1	**86.3**	82.8
Full	26.2	**74.3**	65.4	27.0	**75.4**	64.7
ATB Segmentation	90.6	97.4	**97.6**	90.5	97.4	**97.5**

Table 1: Evaluation metrics on the ATB-ARZ development and test sets. The best results are **bolded**. We compare MADA and MADA-ARZ with different training data conditions. Definitions of metrics are in Section 5.1. MSA training data is ATB-123. ARZ training data is ATB-ARZ. ALL training data is ATB-123 plus ATB-ARZ.

(Maamouri et al., 2004); and ATB-ARZ, the Egyptian Arabic Treebank (parts 1-5) (Maamouri et al., 2012a). For ATB-123 training, we use all of parts 1 and 2 plus the training portion of ATB-3 (as defined by Zitouni et al. (2006)); for development and test, we split Zitouni et al. (2006)'s devtest set into two. We sub-divide ATB-ARZ into development, training, and test sets (roughly a 10/80/10 split). The ATB-ARZ training data has 134K words, and the ATB-123 training data has 711K words.

We evaluate two systems. We used the latest release of MADA for MSA (v3.2), trained on ATB-123 (MSA), as our baseline. For MADA-ARZ, we compare two training settings: using ATB-ARZ (ARZ) and combining ATB-ARZ with ATB-123 (ALL). We present our results on the ATB-ARZ development and blind test sets (21.1K words and 20.4K words). Tuning for MADA-ARZ was done using a random 10% of the ATB-ARZ training data, which was later integrated back into the training set.

Metrics We use several evaluation metrics to measure the effectiveness of MADA-ARZ. **Morph Tag** refers to the accuracy of correctly predicting the full CALIMA morphological tag (i.e., not the diacritics or the lemma). **Penn POS** and MADA **POS** are also tag accuracy metrics. **Penn POS**, also known as the *Reduced Tag Set*, is a tag set reduction of the full Arabic morphological tag set, which was proposed for MSA (Kulick et al., 2006; Diab, 2007; Habash, 2010); since it retains no MSA-specific morphological features, it also makes sense for ARZ. MADA **POS** is the small POS tag set (36 tags) MADA uses internally. **Diacritic** and **Lemma** are the accuracies of the choice of diacritized form and Lemma, respectively. **Full** is the harshest metric, requiring that every morphological feature of the chosen analysis be correct. Finally, **ATB Segmentation** is the percentage of words with correct ATB segmentation (splitting off all clitics except for the determiner +ال *Al*+).

Results The results are shown in Table 1. MADA-ARZ performs much better than the MADA baselines in all evaluation metrics. Comparing the two MADA-ARZ systems, it is evident that adding MSA data (ATB123) results in slightly better performance only for the Penn POS, MADA POS, and ATB Segmentation metrics. Including the MSA data results in accuracy reductions for the other metrics, but the resulting system still outperforms the MADA MSA baseline in all cases. The results are consistent for development and blind test.

The CMUQ-ECA Test Set Mohamed et al. (2012) reported on the task of ARZ raw orthography morph segmentation (determining the morphs in the raw word). The CMUQ-ECA test data comprised 36 ARZ political comments and jokes from the Egyptian web site `www.masrawy.com`. The set contains 2,445 words including punctuation. Their best system gets a 91.9% word-level accuracy. Since MADA-ARZ modifies the spelling

Tokenization	OOV	BLEU	METEOR	TER
Punct	9.2	22.1	27.2	63.2
MADA ATB	5.8	24.4	29.6	60.5
MADA-ARZ ATB	4.9	**25.2**	**29.9**	**59.4**

Table 2: Machine translation results on the test set. "Punct" refers to the baseline which only tokenizes at punctuation.

of the word when it maps into CODA, we needed a manual analysis where no exact match with the gold occurs (11.8% of the time). We determined MADA-ARZ's accuracy on their test set for morph-segmentation to be 93.2%.

5.2 Egyptian Arabic to English MT

MT Experimental Settings We use the open-source Moses toolkit (Koehn et al., 2007) to build a phrase-based SMT system. We use MGIZA++ for word alignment (Gao and Vogel, 2008). Phrase translations of up to 8 words are extracted in the phrase table. We use SRILM (Stolcke, 2002) with modified Kneser-Ney smoothing to build two 4-gram language models. The first model is trained on the English side of the bitext, while the other is trained on the English Gigaword data. Feature weights are tuned to maximize BLEU (Papineni et al., 2002) on a development set using Minimum Error Rate Training (Och, 2003). We perform case-insensitive evaluation in terms of BLEU , METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006) metrics.

Data We trained on DA-English parallel data (Egyptian and Levantine) obtained from several LDC corpora. The training data amounts to 3.8M untokenized words on the Arabic side. The dev set, used for tuning the parameters of the MT system, has 15,585 untokenized Arabic words. The test set has 12,116 untokenized Arabic words. Both dev and test data contain two sets of reference translations. The English data is lower-cased and tokenized using simple punctuation-based rules.

Systems We build three translation systems which vary in tokenization of the Arabic text. The first system applies only simple punctuation-based rules. The second and third systems use MADA and MADA-ARZ, respectively, to tokenize the Arabic text in the ATB tokenization scheme (Habash and Sadat, 2006). The Arabic text is also Alif/Ya normalized.

Results The MT results are in Table 2, which also shows the percentage of out-of-vocabulary (OOV) words – test words not in the training data. MADA-ARZ delivers the best translation performance according to all metrics. All MADA-ARZ improvements over MADA are statistically significant at the .01 level (except in the case of METEOR). All improvements over Punct by MADA and MADA-ARZ are also statistically significant. For BLEU scores, we observe 3.1% absolute improvement to Punct (14% relative), and 0.8% absolute improvement to MADA (3.3% relative). In addition to better morphological disambiguation, MADA-ARZ reduces the OOV ratio (16% relative to MADA), which we suspect contributes to the observed improvements in MT quality.

6 Conclusion and Future Work

We have presented MADA-ARZ, a system for morphological tagging of ARZ. We have shown that it outperforms an state-of-the-art MSA tagger (MADA) on ARZ text, and that it helps ARZ-to-English machine translation more than MADA.

In the future, we intend to perform further feature engineering to improve the results of MADA-ARZ, and extend the system to handle other DAs.

Acknowledgments

This paper is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-12-C-0014. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

References

Ernest T. Abdel-Massih, Zaki N. Abdel-Malek, and El-Said M. Badawi. 1979. *A Reference Grammar of Egyptian Arabic*. Georgetown University Press.

Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In *The 6th International Conference on Informatics and Systems, INFOS2008*. Cairo University.

Rania Al-Sabbagh and Roxana Girju. 2012. A supervised POS tagger for written Arabic social networking corpora. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 39–52. ÖGAI, September. Main track: oral presentations.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Eric Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4):543–565.

Tim Buckwalter. 2007. Issues in Arabic Morphological Analysis. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. In *Proceedings of the European Chapter of ACL (EACL)*.

Walter Daelemans and Antal van den Bosch. 2005. *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.

Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. 2007. Automated methods for processing arabic text: From tokenization to base phrase chunking. In Antal van den Bosch and Abdelhadi Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Kluwer/Springer.

Mona Diab. 2007. Improved Arabic Base Phrase Chunking with a New Enriched POS Tag Set. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 89–96, Prague, Czech Republic, June.

Kevin Duh and Katrin Kirchhoff. 2005. POS tagging of dialectal Arabic: a minimally supervised approach. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Semitic '05, pages 55–62, Ann Arbor, Michigan.

Ramy Eskander, Nizar Habash, Owen Rambow, and Nadi Tomeh. 2013. Processing Spontaneous Orthography. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tim Buckwalter. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

Nizar Habash and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, Sydney, Australia.

Nizar Habash and Fatiha Sadat. 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the 7th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL06)*, pages 49–52, New York, NY.

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

Nizar Habash, Mona Diab, and Owen Rabmow. 2012a. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012b. A Morphological Analyzer for Egyptian Arabic. In *NAACL-HLT 2012 Workshop on Computational Morphology and Phonology (SIGMORPHON2012)*, pages 1–9.

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

Jan Hajič. 2000. Morphological tagging: Data vs. dictionaries. In *Proceedings of the 1st Meeting of the*

North American Chapter of the Association for Computational Linguistics (NAACL'00), Seattle, WA.

Clive Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

H. Kilany, H. Gadalla, H. Arram, A. Yacoub, A. El-Habashi, and C. McLemore. 2002. Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Seth Kulick, Ryan Gabbard, and Mitch Marcus. 2006. Parsing the Arabic Treebank: Analysis and Improvements. In *Proceedings of the Treebanks and Linguistic Theories Conference*, pages 31–42, Prague, Czech Republic.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank : Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow, and Dalila Tabessi. 2006. Developing and Using a Pilot Dialectal Arabic Treebank. In *The fifth international conference on Language Resources and Evaluation (LREC)*, pages 443–448, Genoa, Italy.

Mohamed Maamouri, Ann Bies, Seth Kulick, Dalila Tabessi, and Sondos Krouna. 2012a. Egyptian Arabic Treebank Pilot.

Mohamed Maamouri, Sondos Krouna, Dalila Tabessi, Nadia Hamrouni, and Nizar Habash. 2012b. Egyptian Arabic Morphological Annotation Guidelines.

Emad Mohamed, Behrang Mohit, and Kemal Oflazer. 2012. Annotating and Learning Morphological Segmentation of Egyptian Colloquial Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proceedings of the 41st Annual Conference of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *ACL 2008: The Conference of the Association for Computational Linguistics; Companion Volume, Short Papers*, Columbus, Ohio.

Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland.

Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Matt Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the Association for Machine Transaltion in the Americas (AMTA 2006)*, Cambridge, Massachusetts.

David Stallard, Jacob Devlin, Michael Kayser, Yoong Keok Lee, and Regina Barzilay. 2012. Unsupervised Morphology Rivals Supervised Morphology for Arabic MT. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 322–327, Jeju Island, Korea.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada.

Imed Zitouni, Jeffrey S. Sorensen, and Ruhi Sarikaya. 2006. Maximum entropy based restoration of arabic diacritics. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 577–584, Sydney, Australia.