Addressing Noise in Multidialectal Word Embeddings

Alexander Erdmann, Nasser Zalmout, Nizar Habash
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi
United Arab Emirates
{ae1541,nasser.zalmout,nizar.habash}@nyu.edu

Abstract

Word embeddings are crucial to many natural language processing tasks. The quality of embeddings relies on large non-noisy corpora. Arabic dialects lack large corpora and are noisy, being linguistically disparate with no standardized spelling. We make three contributions to address this noise. First, we describe simple but effective adaptations to word embedding tools to maximize the informative content leveraged in each training sentence. Second, we analyze methods for representing disparate dialects in one embedding space, either by mapping individual dialects into a shared space or learning a joint model of all dialects. Finally, we evaluate via dictionary induction, showing that two metrics not typically reported in the task enable us to analyze our contributions' effects on low and high frequency words. In addition to boosting performance between 2-53%, we specifically improve on noisy, low frequency forms without compromising accuracy on high frequency forms.

1 Introduction

Many natural language processing tasks require word embeddings as inputs, yet quality embeddings require large, non-noisy corpora. Dialectal Arabic (DA), the low register of highly diglossic Arabic (Ferguson, 1959), is problematically noisy. While the high register, Modern Standard Arabic (MSA), is uniform across educated circles in the Arab World, many varieties of DA are not even mutually intelligible (Chiang et al., 2006). The lexical correspondences across four Arab city dialects in Table 1[1] demonstrate that this variation is not limited to sound change among cognate forms, but involves significant lexical changes due to borrowing, semantic shift, etc.

Rabat	Cairo	Beirut	Doha	MSA	Gloss
مطيشة *mTyšħ*	قوطة *qwTħ*	بندورة *bndwrħ*	طماطم *TmATm*	طماطم *TmATm*	*tomato*
طبلة *Tblħ*	طربيزة *Trbyzħ*	طاولة *TAwlħ*	طاولة *TAwlħ*	مائدة *mAŷdħ*	*table*
لديد *ldyd*	حلو *Hlw*	طيب *Tyb*	لذيذ *lðyð*	لذيذ *lðyð*	*declicious*

Table 1: Lexical correspondences between four urban Arabic dialects and MSA.

Seldom written previously, DA is becoming the dominant form of Arabic on social media, yet annotated data are still scarce (Muhammad Abdul-Mageed and Elaraby, 2018; Israa Alsarsour and Elsayed, 2018; Kareem Darwish and Kallmeyer, 2018). While complex morphology contributes to sparsity in both MSA and DA (Habash, 2010), noise from inter-dialect variation and unstandardized spelling further reduces token-to-type ratios in DA. This limits opportunities to learn accurate vector representations for any given word. Table 2 shows that the MSA token-to-type ratio is over three times larger than DA, controlling for corpus size. This is still not nearly as large as English due to English's morphological simplicity.[2] Furthermore, the percentage of tokens belonging to low frequency types is three times greater in DA.

Many previous works ignore inter-dialect variation, training dialect agnostic embeddings, yet we show that modeling dialects individually yields

[1]Examples are drawn from the MADAR lexicon (Bouamor et al., 2018). Arabic script follows CODA guidelines (Habash et al., 2018) and transliteration is presented in the HSB scheme (Habash et al., 2007).

[2]Our DA corpora are described in Section 3 whereas the MSA and English sentences are randomly drawn from the parallel corpus described in Almahairi et al. (2016).

	Egyptian	Levantine	MSA	English
Tokens per type	20	19	68	128
Tokens with type frequency < 5	6%	6%	2%	1%

Table 2: Token and type based comparisons between two dialects of Arabic, MSA, and English in corpora of 13 million words each.

strong performances in a dictionary induction task when noise is systematically addressed. To that end, we make three contributions. First, we describe simple but effective adaptations to word embedding tools to maximize the informative content leveraged in each training sentence. Second, we compare methods for representing disparate dialects in one embedding space, by mapping individual dialects into shared space or learning a joint model of all dialects. Finally, we evaluate our techniques via dictionary induction, showing that two metrics not typically reported are quite informative. In addition to improving accuracy 2-53%, our adaptations specifically boost performance on noisy, low frequency forms without compromising accuracy on high frequency forms.

2 Related Work

Common monolingual embedding models are trained to predict either the target word given the context (Continuous Bag of Words) or elements of the context given the target (SkipGram) (Mikolov et al., 2013a). These have been adapted to incorporate word order (Trask et al., 2015) or subword information (Bojanowski et al., 2016) to model syntax, morphology, etc.

Bilingual embeddings are vector representations of two languages mapped into shared space, such that translated word pairs have similar vectors (Gouws et al., 2015; Luong et al., 2015). They facilitate applications from parallel sentence extraction (Grover and Mitra, 2017) to machine translation (Zou et al., 2013; Cholakov and Kordoni, 2016; Artetxe et al., 2017b) and can be used to improve monolingual embeddings (Faruqui and Dyer, 2014). Bilingual embeddings are learned via one of three methods: mapping both spaces into a shared space (Mikolov et al., 2013b), monolingual adaptation of one language's embedding space into another's (Zou et al., 2013), or bilingually training both embeddings simultaneously (AP et al., 2014; Pham et al., 2015). We compare implementations of two state-of-the-art models for mapping embeddings that use the monolingual adaptation technique, as these best suited our data and resources: VECMAP (Artetxe et al., 2016, 2017a) and MUSE (Conneau et al., 2017). Both are equipped to learn either via supervision or by iteratively mapping with little or no supervision. Recently, another unsupervised approach leveraging local neighborhood structures was evaluated on French, English, and MSA (Aldarmaki et al., 2017). Such approaches address seed data scarcity, but have not previously been applied to sparse corpora lacking standardized spelling. While we address unstandardized spelling indirectly by learning better embeddings for low frequency types, Zalmout et al. (2018), Abidi and Smaïli (2018), and Dasigi and Diab (2011) attempt to map DA spelling variants to each other.

We are the first to use embeddings for multiple specific DA dialects, though DA embeddings are often used for sentiment analysis (Al Sallab et al., 2015; Altowayan and Tao, 2016). One such work, Dahou et al. (2016), uses pre-built dictionaries to deterministically identify phrases in mixed MSA-DA data before training embeddings. In MSA, embeddings have been used in additional tasks like morphological analysis (Zalmout and Habash, 2017) and POS tagging (Darwish et al., 2017).

3 Data

We adopt Zaidan and Callison-Burch (2011)'s 4-way coarse-grained dialect distinction of Gulf (GLF), Maghrebi (MAG), Egyptian (EGY), and Levantine (LEV). We collect corpora for each dialect by concatenating the relevant dialect identified portion of the following corpora: Almeman and Lee (2013)'s web crawl of forums, comments and blogs, Khalifa et al. (2016)'s Gumar corpus of internet novels,[3] the Broad Operational Language Translation corpus of primarily blogs described in Zbib et al. (2012), the dialectal Arabic travel corpus of Bouamor et al. (2018), Zaidan and Callison-Burch (2011)'s online news commentary corpus, and Jarrar et al. (2014)'s corpus of subtitles and tweets. This results in 1.7 million sentences of EGY, 1.5 million GLF, 1.3 million LEV, and 1.1 million MAG. These corpora are each about 200 times smaller than MSA's single-domain Gigaword (Parker et al., 2011), with lack of standard-

[3]Gumar's GLF portion is huge, making the GLF corpus less comparable to other dialects. Thus, we removed GLF Gumar as its inclusion did not help performance.

ized spelling and internal domain inconsistency compounding scarcity with noise.

To map dialects' embeddings into shared spaces and evaluate dictionary induction, we generate seed and test dictionaries similar to Artetxe et al. (2016). We use MGIZA (Koehn et al., 2007) to align 8,000 sentences from Bouamor et al. (2018)'s travel corpus. It contains 12,000 five-way parallel sentences between the DA varieties of Beirut (LEV), Cairo (EGY), Doha (GLF), Tunis (MAG), and Rabat (MAG), but we collapse Tunis and Rabat to match Zaidan and Callison-Burch (2011)'s granularity and hold out 4,000 sentences for development on downstream tasks. After alignment, we extract unigram translations from 2,000 sentences to form a bidialectal evaluation dictionary. This yields between 2,500 and 4,000 word pairs, with 1.3 to 1.7 average translations per word depending on the dialect pair. Lastly we realign the remaining 6,000 training sentences and extract a seed dictionary. Three annotators jointly evaluated 400 unigram pairs from the LEV–EGY evaluation dictionary. 89% were acceptable translations.

4 Word Embedding Models

We consider the following models for training word embeddings:

FT refers to a FASTTEXT (Bojanowski et al., 2016) implementation of SkipGram with 200 dimensions and a context window of 5 tokens on either side of the target word. A word's vector is the sum of its SkipGram vector and that of all its component character n-grams between length 2 and 6. Since short vowels are not typically written in Arabic, many affixes only consist of a word start/end token and one character. Thus, these character n-gram parameters outperformed the range of 3 to 6 proposed by Bojanowski et al. (2016) for other languages. In preliminary experiments, FT outperformed WORD2VEC models (Mikolov et al., 2013a; Řehůřek and Sojka, 2010) which lack subword information and hence struggle with Arabic's morphological complexity. We also compared FT to variant implementations with larger and smaller context windows, though FT consistently performed the same or better.

EXT refers to an extended FT model where wide and narrow windowed embeddings, sizes 5 and 1 respectively, are trained separately. Resulting vectors are concatenated to build a 400 dimensional model. Given much work demonstrating that narrow context windows capture more syntactic information and wide windows, semantic information (Pennington et al., 2014; Trask et al., 2015; Goldberg, 2016; Tu et al., 2017), component vectors should complement each other, giving the concatenated vector access to a wider range of linguistic information. To ensure that the improvement came from vector concatenation and not simply from having higher dimensional vectors, we built 400 dimension FT models to compare to EXT, but they did not outperform 200 dimensional FT, likely due to sparsity.

PP+EXT refers to an EXT model trained on a preprocessed corpus where phrases have been probabilistically identified. To identify phrases, we recurse over each sentence R times, each time forming bigram phrases from component unigrams (which could have been longer n-grams in previous iterations) depending on the frequencies of the relevant unigrams and bigrams. We implement this step exactly as described in Mikolov et al. (2013c), but then we copy each output sentence C times and probabilistically decompose the deterministically identified phrases into smaller n-grams.[4] More precisely, for each deterministically identified n-gram phrase, we progress from the first to the $(n-1)^{\text{th}}$ gram, randomly splitting the phrase at that point with probability $\frac{e^n}{\sum_{r=1}^{R} e^r}$. The final result of the probabilistic phrase identification is C potentially unique copies of each sentence containing identified n-gram phrases of length $n \leq 2^R$. We experimented with linear distributions in addition to the exponential one used for phrase splitting, but the exponential performed better. The exponential distribution means that it is less likely to separate at any given potential break point in longer n-grams than in shorter ones.

Like Mikolov et al. (2013c)'s deterministic identification of phrases, PP+EXT avoids training vectors on individual words in non-compositional phrases, yet PP+EXT's probabilistic nature lets the model learn from multiple perspectives of every word/phrase's context, with more informative phrase distributions more likely to appear more frequently. Interestingly, identifying phrases can

[4] Using a development set, we found performance to plateau around R=5 and C=15 and thus adopt these parameters, though higher values of C could in theory marginally boost performance at the expense of runtime.

be harmful, as our evaluation is performed on unigrams. We implemented a deterministic version of PP+EXT but it did not outperform the baseline FT as too many unigrams were lost in longer phrases. Thus, identifying phrases probabilistically is crucial to PP+EXT's high performance.

In preliminary experiments, probabilistic phrase identification improved the FT model without extending vectors, yet the performance did not exceed EXT. Hence, we only report PP+EXT scores, as the technique is far more effective when coupled with EXT. The combination of techniques is actually designed to be complementary: FT leverages morphology, EXT combines syntax with semantics, and probabilistic phrase identification increases the number of meaningful contexts used for training. These enable the model to learn better representations for noisy, low frequency forms without requiring additional data.

5 Multidialectal Embedding Space

We consider two options for generating multidialectal embeddings for DA: (a) a dialect agnostic model trained on all DA corpora, and (b) training individual dialect models separately before mapping them into a shared embedding space. While (b) leverages less data per model, (a) is subject to more noise and ambiguity, as many words are unique to certain dialects or have disparate meanings in different dialects. (b) can be seeded with a bidialectal dictionary or parallel sentences. We found the dictionary approach to perform better.

ALLDA is a PP+EXT model trained on a combined corpus of all dialects. To avoid code switching issues, ALLDA assigns words only to those dialects for which its relative frequency in that dialect's corpus is greater than 5% of its maximum relative frequency in any dialect. Thus, a word assigned to multiple dialects will take the same vector in each dialect and be its own nearest neighbor for any dialect pairs where it belongs to both.

VECMAP is Artetxe et al. (2016, 2017a)'s tool that uses a seed dictionary (or shared numerals) to learn a mapping function which minimizes distances between seed dictionary unigram pairs. In data scarce settings, the function can be learned iteratively, inducing a larger seed dictionary each round, yet the noise in our DA corpora prevents this process from getting off the ground, producing scores of zero after a few iterations.

MUSE is Conneau et al. (2017)'s tool, using adversarial learning (and optionally a seed) to identify similarly behaving high frequency anchor words, bootstrapping into fine tuning the mapping of less frequent words. MUSE is specifically designed for data scarce and unsupervised settings. It assumes shared embedding structures to be identifiable, and the authors demonstrate that domain differences can strain this assumption.

6 Experiments and Results

To evaluate the quality of our DA word embeddings, we use the task of dictionary induction. Given source dialect words from the evaluation dictionary, we attempt to recall appropriate translations in the target dialect based on cosine distance in multidialectal embedding space. The standard metric for this task is precision@k=1 (P@1) (Artetxe et al., 2016, 2017a; Conneau et al., 2017), measuring the fraction of source words in the evaluation dictionary for which the nearest target dialect neighbor matches any of the possible translations in the evaluation dictionary.

We, however, are also concerned with how well multiple translations are recalled, as many words become polysemous in DA with short vowels omitted and spelling not standardized. For this reason, many words appearing both in the seed and evaluation dictionaries do not map to the exact same set of possible translations in each. Thus, many precision errors may be forgiveable, so we focus on recall, reporting the metric recall@k=5 (R@5). Lastly, because types appear in a Zipfian distribution and type-based metrics disproportionately reflect accuracy in the tail, we report a frequency weighted recall@k=5 (WR@5) as well.[5] Considering both R@5 and WR@5 avoids the risk of improving performance on high or low frequency types at the expense of the other.

In Table 3, models FT, EXT, and PP+EXT are trained on individual dialects, then mapped using supervised SVECMAP into bidialectal embedding spaces. We experimented with all combinations of mapping algorithms and embedding models, yet SVECMAP consistently outperformed the other mapping algorithms. We also report results for unsupervised UMUSE leveraging PP+EXT embeddings. ID is an identity dictionary mapping

[5] R@5 and WR@5 are normalized by the score of an oracle that correctly recalls up to 5 translations of every source word, but no more should more exist. Thus, the maximum score for these metrics is 1, making them comparable to P@1.

			SVECMAP			ALLDA	UMUSE
	Metric	ID	FT	EXT	PP+EXT	PP+EXT	PP+EXT
MAG	WR@5	28.9	35.3	42.2	**47.0**	32.6	26.8
⇓	R@5	24.9	36.2	40.4	**51.1**	26.2	14.9
LEV	P@1	33.6	35.3	39.7	**54.0**	33.7	12.2
MAG	WR@5	37.5	46.9	49.7	**50.8**	40.5	42.3
⇓	R@5	30.4	36.9	41.2	**45.2**	29.0	25.4
GLF	P@1	35.0	31.1	37.9	**40.0**	29.6	19.1
MAG	WR@5	42.4	48.2	**48.3**	47.9	45.8	43.1
⇓	R@5	30.7	34.5	39.4	**42.9**	34.0	25.5
EGY	P@1	36.0	29.4	**38.0**	36.6	36.3	20.9
EGY	WR@5	42.9	51.3	51.3	**52.8**	47.8	40.5
⇓	R@5	40.9	48.2	49.9	**52.8**	38.4	33.1
GLF	P@1	47.7	43.3	**48.5**	48.3	41.7	24.0
LEV	WR@5	43.2	50.6	50.4	**51.7**	48.5	40.9
⇓	R@5	33.6	37.8	38.9	**46.4**	31.8	24.7
GLF	P@1	39.0	34.1	37.5	**41.7**	33.1	20.0
LEV	WR@5	44.0	50.3	49.8	**52.4**	50.6	48.1
⇓	R@5	33.0	27.6	39.6	**42.3**	36.5	31.1
EGY	P@1	**39.6**	33.8	38.8	37.7	39.2	25.9

Table 3: Dictionary induction results comparing various multidialectal embedding models mapped via supervised (SVECMAP) and unsupervised (ALLDA, UMUSE) techniques.

all source words to themselves, thus representing dialect similarity. PP+EXT or EXT always outperform the baseline FT, with PP+EXT being the best model in all but one instance according to WR@5 and R@5. PP+EXT successfully addresses noise as its gains are larger on non-frequency weighted R@5 than WR@5; i.e., it improves on low frequency words without compromising high frequency word accuracy. Additionally, the consistency in results for WR@5 and R@5 as compared to P@1 suggests the small k is contributing to noise in the P@1 metric.

While ALLDA generally performs worse than the supervised mapping approaches, it typically performs slightly better on words which were not found in their seed dictionaries according to R@5, likely because it can leverage more data to learn better representations for non-ambiguous, low frequency shared forms. Depending on the intended application, system combination could be ideal, querying ALLDA for low frequency forms appearing in multiple dialects, but not the seed.

	SVECMAP	SMUSE	UMUSE	ALLDA
WR@5	0.70	0.97	0.90	0.99
R@5	0.24	0.48	0.89	0.86
P@1	0.03	0.18	0.68	0.78

Table 4: Correlation between mapping performance and dialect similarity, i.e., the ID baseline, using PP+EXT embeddings.

As for supervised mapping algorithms, Table 4 shows that, depending on the dialect pair in question, SMUSE's adversarial learning approach correlates with ID's metric of dialect similarity 20-30% more strongly than SVECMAP, which takes greater advantage of seed–evaluation domain similarity. Accordingly, SVECMAP beats SMUSE on in-seed forms by 3-23%. That said, SMUSE is more robust to seed coverage, slightly outperforming SVECMAP on out-of-seed forms and UMUSE successfully bootstraps without supervision, unlike UVECMAP. Still, the best performing option in the unsupervised set up is ALLDA. UMUSE's performance does not approach that of supervised alternatives as reported in Conneau et al. (2017). This is likely because they (as do Artetxe et al. (2017a)) impose bilingual data scarcity constraints on high resource languages but do not consider the sparsity effects of noise common in low resource languages. They use large quantities of domain consistent, spelling standardized monolingual data which are not available for DA.

7 Conclusion and Future Work

We presented techniques for generating multidialectal word embeddings from noisy DA corpora. Due to linguistic differences, modeling dialects independently and mapping embeddings into multidialectal space generally outperformed training dialect agnostic embeddings on combined corpora. Our novel techniques include concatenating narrow and wide windowed vectors and probabilistically identifying phrases before training embeddings. These techniques improved performance on bidialectal dictionary induction 2-53% over a state-of-the-art baseline, with most of the improvement realized on noisy, low frequency word forms. Our approach can easily be applied to other, similarly noisy corpora. In future work, we will improve the handling of orthographically ambiguous words, which are very prevalent in DA, and we will evaluate on the downstream applications of machine translation and morphological disambiguation.

Acknowledgments

The second author was supported by the New York University Abu Dhabi Global PhD Student Fellowship program. We are also grateful for the support of the High Performance Computing Center at New York University Abu Dhabi.

References

Karima Abidi and Kamel Smaïli. 2018. An automatic learning of an Algerian dialect lexicon by using multilingual word embeddings. In *11th edition of the Language Resources and Evaluation Conference, LREC 2018*.

Ahmad A Al Sallab, Ramy Baly, Gilbert Badaro, Hazem Hajj, Wassim El Hajj, and Khaled B Shaban. 2015. Deep learning models for sentiment analysis in Arabic. In *ANLP Workshop*. volume 9.

Hanan Aldarmaki, Mahesh Mohan, and Mona Diab. 2017. Unsupervised word mapping using structural similarities in monolingual embeddings. *arXiv preprint arXiv:1712.06961* .

Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron C. Courville. 2016. First result on Arabic neural machine translation. *CoRR* abs/1606.02680. http://arxiv.org/abs/1606.02680.

Khalid Almeman and Mark Lee. 2013. Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. In *Communications, signal processing, and their applications (iccspa), 2013 1st international conference on*. IEEE, pages 1–6.

A Aziz Altowayan and Lixin Tao. 2016. Word embeddings for Arabic sentiment analysis. In *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, pages 3820–3825.

Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*. pages 1853–1861.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *EMNLP*. pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017a. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 451–462.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017b. Unsupervised neural machine translation. *CoRR* abs/1710.11041. http://arxiv.org/abs/1710.11041.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *The International Conference on Language Resources and Evaluation*. Miyazaki, Japan.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. In *Proceedings of EACL*. Trento, Italy.

Kostadin Cholakov and Valia Kordoni. 2016. Using word embeddings for improving statistical machine translation of phrasal verbs. *ACL 2016* page 56.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* .

Abdelghani Dahou, Shengwu Xiong, Junwei Zhou, Mohamed Houcine Haddoud, and Pengfei Duan. 2016. Word embeddings and convolutional neural network for Arabic sentiment classification. In *COLING*. pages 2418–2427.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, and Mohamed Eldesouki. 2017. Arabic POS tagging: Don't abandon feature engineering just yet. *WANLP 2017 (co-located with EACL 2017)* page 130.

Pradeep Dasigi and Mona Diab. 2011. Codact: Towards identifying orthographic variants in dialectal arabic. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. pages 318–326.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. Association for Computational Linguistics.

Charles F Ferguson. 1959. Diglossia. *Word* 15(2):325–340.

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *J. Artif. Intell. Res.(JAIR)* 57:345–420.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. pages 748–756.

Jeenu Grover and Pabitra Mitra. 2017. Bilingual word embeddings with bucketed cnn for parallel sentence extraction. In *Proceedings of ACL 2017, Student Research Workshop*. pages 11–16.

Nizar Habash, Salam Khalifa, Fadhl Eryani, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al Shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In *The International Conference on Language Resources and Evaluation*. Miyazaki, Japan.

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 Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Esraa Mohamed Reem Suwaileh Israa Alsarsour and Tamer Elsayed. 2018. Dart: A large dataset of dialectal arabic tweets. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, HÃl'lÃÍne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Paris, France.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a corpus for Palestinian Arabic: a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*. Association for Computational Linguistics, Doha, Qatar, pages 18–27. http://www.aclweb.org/anthology/W14-3603.

Hamdy Mubarak Ahmed Abdelali Mohamed Eldesouki Younes Samih Randah Alharbi Mohammed Attia Walid Magdy Kareem Darwish and Laura Kallmeyer. 2018. Multi-dialect arabic pos tagging: A crf approach. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, HÃl'lÃÍne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Paris, France.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. *arXiv preprint arXiv:1609.02960* .

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pages 177–180.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL15)*. pages 151–159.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Hassan Alhuzali Muhammad Abdul-Mageed and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, HÃl'lÃÍne Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Paris, France.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Arabic Gigaword Fifth Edition. LDC catalog number No. LDC2011T11, ISBN 1-58563-595-2.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

Hieu Pham, Thang Luong, and Christopher D Manning. 2015. Learning distributed representations for multilingual text sequences. In *Proceedings the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL15)*. pages 88–94.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pages 45–50. http://is.muni.cz/publication/884893/en.

Andrew Trask, David Gilmore, and Matthew Russell. 2015. Modeling order in neural word embeddings at scale. *arXiv preprint arXiv:1506.02338* .

Lifu Tu, Kevin Gimpel, and Karen Livescu. 2017. Learning to embed words in context for syntactic tasks. *arXiv preprint arXiv:1706.02807* .

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the Association for Computational Linguistics*. Portland, Oregon, USA.

Nasser Zalmout, Alexander Erdmann, and Nizar Habash. 2018. Noise-robust morphological disambiguation for dialectal Arabic. In *Proceedings of the 16th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL18)*. New Orleans, Louisiana, USA.

Nasser Zalmout and Nizar Habash. 2017. Don't throw those morphological analyzers away just yet: Neural

morphological disambiguation for Arabic. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 715–724.

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*. Association for Computational Linguistics, pages 49–59.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pages 1393–1398.