On the Role of Orthographic Variations in Building Multidialectal Arabic Word Embeddings

Abdellah El Mekki[†,*], Abdelkader El Mahdaouy[†], Ismail Berrada[†],Ahmed Khoumsi[‡]
[†] School of Computer Sciences, Mohammed VI Polytechnic University, Morocco
[‡] Dept. Electrical & Computer Engineering, University of Sherbrooke, Canada

Abstract

Dialectal Arabic (DA) is mostly used by over 400 million people across Arab countries as a communication channel on social media platforms, web forums, and daily life. Building Natural Language Processing systems for each DA variant is a challenging issue due to the lack of data and the noisy nature of the available corpora. In this paper, we propose a method to incorporate orthographic features into word embedding mapping methods, inducing a multidialectal embedding space. Our method can be used for both supervised and unsupervised cross-lingual embedding mapping approaches. The core idea of our method is to project the orthographic features into a shared vector space using Canonical Correlation Analysis (CCA). Then, it extends word embedding vectors using the resulting features and learns the multidialectal mapping. The overall obtained results of our proposed method show that our method enhances Bilingual Lexicon Induction of DA by 3.33% and 17.50% compared to state-of-the-art supervised and unsupervised cross-lingual alignment methods, respectively.

Keywords: Arabic Natural Language Processing, Dialectal Arabic, Multidialectal Word Embedding, Cross-lingual Word Embedding Mapping, Bilingual Lexicon Induction

1. **Introduction**

Many languages around the world are being used more for speech communication than for formal writing, resulting in a lack of textual data for downstream tasks in Natural Language Processing (NLP). Dialectal Arabic (DA) is a specific example of this observation. It is considered to be a low-resource language, which is mostly used today on social media, and it has many variants that make the problem of building NLP systems for each DA variant harder.

Annotating textual data for these languages might be the solution to this bottleneck, however this process is tedious, costly, and time-consuming, especially for languages varieties with no standard writing [1]. Recent studies [2, 3] have explored the use of transfer learning from a rich-resource language (e.g. English, French, Modern Standard Mandarin, Modern Standard Arabic (MSA)) to a low-resource language (e.g. Inuit, Sinhala, Sindhi, Nepali, DA), by finding a shared embedding space for multiple languages. Most of the time, finding this shared embedding space relies on learning a mapping between the word embedding spaces of a source language and a target language [4, 5]. In a more general setting, Bilingual Lexicon Induction (BLI) is used as an evaluation task for cross-lingual embedding mapping.

BLI's recent progress has shown promising results on low-resource languages using both supervised [2] and unsupervised [6] approaches. Recent work have incorporated orthographic features [7] as additional features to word vectors. The aim is to improve the BLI performance for close languages. Although several research works have showed that the DA variants share similarities on multiple linguistic levels (orthography, morphology, phonology, etc) [1].

In this paper, we propose a method to incorporate orthographic features for cross-lingual mapping between Arabic dialects. The aim is to project dialects' embedding spaces into a

[*] abdellah.elmekki@um6p.ma

single shared multidialectal space, leveraging their orthographic similarities. Unlike the work of [7], who incorporated orthographic features directly into the word vectors, we perform the Canonical Correlation Analysis (CCA) transformation between the orthographic features of the source and the target dialects to induce a shared space for these additional features. The resulting representation of words' orthographic features is then concatenated with the embedding vectors to learn the mapping. Besides, our method can be used successfully in both supervised and unsupervised mapping approaches. Experiments are performed on four Arabic dialects: Maghrebi (MAG), Egyptian (EGY), Gulf (GLF) and Levantine (LEV). For the evaluation of the obtained word embedding mappings, we use the BLI task. The obtained results show that our method yields very promising results, especially for unsupervised learning of mappings. In comparison to existing methods, it achieves an average gain of 3.33% and 17.50% for the supervised and the unsupervised mapping approaches, respectively.

The rest of the paper is organized as follows: Section 2 introduces the background and the notation used in this paper. Section 3 describes the proposed method. Section 4 presents experiments and the obtained results. Finally, Section 5 concludes the paper.

2. Background

As our method is based on the recent approaches and methods for learning a mapping between embedding spaces, in this section, we recall the basic concepts related to cross-lingual alignment systems and orthographic extension of word embedding. For the rest of this paper, we will consider two distinct languages L_s (source language) and L_t (target language), having the vocabularies V_s and V_t respectively. Let us denote by:

- $X^e \in \mathbb{R}^{|V_s| \times d}$ and $Z^e \in \mathbb{R}^{|V_t| \times d}$, the corresponding embedding matrices of L_s and L_t, respectively, where d is the word embedding vector dimension.
- X^e_{i*} and Z^e_{i*}, the ith row of matrices X^e and Z^e, respectively.
- Z^{e^T}, the transpose of the matrix Z^e.

The objective of a cross-lingual mapping of word embeddings is to find a mapping matrix $W \in \mathbb{R}^{d \times d}$ such that WX^e best approximates Z^e.

2.1. Supervised and unsupervised cross-lingual alignment

In the case of cross-lingual mapping, both supervised and unsupervised approaches have been proposed. The state-of-the-art supervised approach [2] has used the Procrustes solution [8]. The optimal mapping matrix is equivalent to minimizing the following Frobenius norm:

$$\underset{W}{\operatorname{argmin}} \|X^e W - Z^e\|_F^2$$

This optimization problem has the analytic solution of $W = VU^T$ where $Z^{e^T} X^e = U\Sigma V^T$ is the Singular Value Decomposition (SVD) of $Z^{e^T} X^e$.

For the unsupervised setting, Conneau, Lample, Ranzato, Denoyer, and Jégou [6] have proposed the MUSE system. It consists of two phases: a distribution matching phase using adversarial training, followed by an iterative Procrustes refinement phase. For both supervised and unsupervised approaches, the pre-processing proposed by [2] (vector length normalization and zero mean centring) is applied to word vectors of L_s and L_t.

2.2. Orthographic extension for cross-lingual mapping

The use of orthographic features to improve the cross-lingual mapping has been proposed first in [7].

Their method consists of three steps:

- Step 1: compute the ordered set A containing all characters of both languages L_s and L_t: $A = A_{L_s} \cup A_{L_t}$
- Step 2: extend the embeddings of L_s and L_t with X^o and Z^o, where X^o (resp. Z^o) contains the counts of each character appearing in every word w_i of L_s (resp. L_t). The count value is then scaled by an empirical scalar α.
$$O_{ij} = \alpha.count(A_j, w_i), O \in \{X^o, Z^o\}$$
where i refers to ith row on the matrix O and j refers to the language L_s or L_t.
- Step 3: use the final embedding matrices X' and Z' to learn the mapping, such that: $X' = [X^e, X^o], Z' = [Z^e, Z^o]$

2.3. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is used in order to maximize the correlation between two vector spaces X and Z. This is achieved by learning two projection vectors a and b for X and Z. The projected vectors of a pair (X_{i*}, Z_{i*}) for words of index i are given by: $X'_{i*} = X_{i*}a \quad Z'_{i*} = Z_{i*}b$

The correlation between the projected vectors is expressed as:

$$\rho(X'_{i*}, Z'_{i*}) = \frac{E[X'_{i*}\ Z'_{i*}]}{\sqrt{E[(X'_{i*})^2]\ E[(Z'_{i*})^2]}}$$

CCA maximizes the correlation ρ by finding an optimal pair (a, b) of projection vectors: $a, b = CCA(X_{i*}, Z_{i*}) = \underset{a,b}{\operatorname{argmax}}\ \rho(X_{i*}a, Z_{i*}b)$.

3. Proposed method for cross-lingual mapping of Arabic dialects

In this section, we describe our proposed method for cross-lingual mapping of Arabic dialects into a shared embedding space.

3.1. DA orthographic variations

Arabic dialects share many features at different levels of linguistic representation (e.g. phonology, morphology, lexicon) [1]. These features can be leveraged to improve cross-lingual alignment of dialects embeddings.

Some of the orthographic variations can be attributed to phonological and morphological variations between DA and MSA, their proto-language [1]. Additionally, the absence of an orthographic standard is another main cause of non-standard writing which results in other forms of orthographic variations. For instance, DA writers sometimes tend to write words etymologically (based on the MSA origin of words), while other times phonologically, as it is the case of "ثعلب" vElb or "تعلب" tElb. Like MSA, other orthographic variations are due to the non-standard writing of some letters such as "أ، إ، ٱ، آ، ة، ي".

The next section shows how the aforementioned orthographic variations can be captured, using orthographic features, to enhance the alignment between DA embedding spaces.

3.2. Cross-lingual alignment method

The core idea of our method is to encode the orthographic features of Arabic dialects into a shared vector space. The resulting representations are then used to extend the word vectors for performing the embedding spaces mapping. For this purpose, we use the CCA transformation to maximize the correlation between the orthographic features of Arabic dialects.

Our method incorporates orthographic features into DA embeddings alignment throughout the following three phases:

(1) Phase 1: Building the seed dictionary for encoding the orthographic features of Arabic dialects. For supervised alignment, we consider the same seed dictionary provided to train the cross-lingual mapping. For the unsupervised alignment, we apply the edit distance on the initial seed dictionary, generated using the adversarial mapping [6]. We limit our training data to seed dictionary pairs for which the edit distance is smaller than 2^1. Hence, our CCA alignment of orthographic features can cover most orthographic variations between Arabic dialects.

(2) Phase 2: Extracting the orthographic features. We convert each word to its orthographic vector representation. Instead of considering all characters that appear in both languages L_s and L_t, we limit this rule in our method to Arabic letters, covered by the Safe Buckwalter transliteration scheme [9]. The aim is to prevent some special characters, such as \, >, <..., from appearing in many vocabulary words.

(3) Phase 3: Maximizing the correlation between orthographic features of DA translation pairs using the CCA. The latter measures the linear relationship between multidimensional variables.

Let X^o and Z^o be the orthographic feature matrices of the source and target dialects, respectively. We use CCA to maximize the correlation between these two matrices.

$$\rho(X^{o'}_{i*}, Z^{o'}_{i*}) = \frac{E[X^{o'}_{i*}\ Z^{o'}_{i*}]}{\sqrt{E[(X^{o'}_{i*})^2]\ E[(Z^{o'}_{i*})^2]}}$$

CCA maximizes the correlation ρ by finding an optimal pair $(a,\ b)$ of projection vectors:

$$a, b = CCA(X^o_{i*}, Z^o_{i*}) = \underset{a,b}{\operatorname{argmax}}\ \rho(X^o_{i*}a, Z^o_{i*}b)$$

$X^{o'}$ and $Z^{o'}$ are the resulting encoded orthographic features matrices.

After encoding the orthographic features, we perform independent normalization of:

(1) The word embedding matrices X^e and Z^e:

$$X^{e'}_{i*} = \frac{X^e_{i*}}{\|X^e_{i*}\|}, Z^{e'}_{i*} = \frac{Z^e_{i*}}{\|Z^e_{i*}\|}$$

(2) The orthographic features matrices X^o and Z^o:

$$X^{o'}_{i*} = \frac{X^{o'}_{i*}}{\|X^{o'}_{i*}\|}, Z^{o'}_{i*} = \frac{Z^{o'}_{i*}}{\|Z^{o'}_{i*}\|}$$

Finally, the encoded orthographic features matrices X''_o and Z''_o are concatenated with the word embedding matrices X'_e and Z'_e of the source and target dialects:

$$X = [X^{e'}, X^{o'}] \text{ and } Z = [Z^{e'}, Z^{o'}]$$

The resulting matrices are then used to learn the cross-lingual alignment. For the supervised cross-lingual alignment, we use the official Vecmap tool [2], while for the unsupervised approach, we train our cross-lingual embedding using MUSE tool [6]. For nearest neighbor retrieval, we employ Cross-Domain Similarity Local Scaling (CSLS) [6]. The CSLS retrieval distance is defined as follows: $CSLS(x, z) = 2\cos(Wx, z) - \Gamma_Z(Wx) - \Gamma_{WX}(z)$. Where $\Gamma_A(b)$ is the average cosine similarity between b and its k nearest neighbors in A.

[1] After testing several values of edit distance limit, a value of 2 shown to give the optimal results for our study.

4. **Experiments**

In this section, we evaluate our method and compare it with previous approaches for DA word embedding alignment. Experiments include both the supervised and unsupervised cross-lingual alignment approaches. The evaluation is performed using the BLI task which is considered as the main evaluation task for cross-lingual word embedding alignment. The objective of this task is to assess translation pairs of the source language and the target language using a bilingual dictionary. We follow the same evaluation procedure as [2, 6] and use precision@k=1 (P@1) as a metric for BLI.

4.1. **Experimental setup**

The experiments are performed under the following setting:

(1) Four Arabic dialects are considered: MAG, EGY, GLF, and LEV. We evaluate our method using off-the-shelf word embedding, pretrained by [3]. The word vectors are the concatenation of separately trained wide and narrow windowed FastText embedding models of dimension 200 [10]. The wide context window is set to 5, while the narrow context window is fixed to 1. The aim is to capture both syntactic and semantic information of words [3]. The resulting embeddings are 400-dimensional vectors. We use the same dictionaries of [3], produced by aligning 8000 parallel sentences of the four evaluated regions. This yields between 3000 and 7000 pairs for the training dictionaries and between 2000 and 3000 for the evaluation dictionaries.

(2) We investigate the performance of our method against several state-of-the-art supervised and unsupervised alignment methods. For the supervised approach, we compare our work with three existing embedding alignment methods [2, 3, 7]. For the unsupervised approach, we compare our work with [3].

4.2. **Results**

	Supervised				Unsupervised	
	Erdmann et. al [3]	**Artetxe et. al [2]**	**Riley and Gildea [7]**	**Our method**	**Erdmann et. al [3]**	**Our method**
MAG TO LEV	54.00	62.7	57.01	**64.53**	12.2	**32.24**
MAG TO GLF	40.00	44.92	45.27	**47.87**	19.1	**34.75**
MAG TO EGY	36.5	41.13	41.96	**44.48**	20.9	**35.80**
EGY TO GLF	48.3	52.34	53.56	**55.27**	24.0	**46.33**
LEV TO GLF	41.7	46.85	46.03	**48.49**	20.0	**38.58**
LEV TO EGY	37.7	42.48	42.52	**45.67**	25.9	**39.44**
Average	43.03	48.40	47.72	**51.05**	20.35	**37.85**

Table 1. BLI task P@1 (%): comparing P@1 scores of various supervised and unsupervised methods for multi-dialectal embedding alignment.

Supervised cross-lingual alignment. Table 1 reports the obtained results using supervised methods for the BLI task. The overall results prove that incorporating orthographic features (using our method and that of [7]) into DA embedding alignment improves the BLI performance. The results also show that our method outperforms the evaluated state-of-the-art methods on all Arabic dialects pairs. On average, our method surpasses the previous orthographic feature-based method [7] and previous work on DA [3] by 3.33% and 8.01%, respectively.

Unsupervised cross-lingual alignment. Table 1 summarizes the obtained results of our method for unsupervised cross-lingual alignment of DA. The obtained results show that our method outperforms the state-of-the-art DA alignment method by a large margin on all evaluated DA pairs. On average, our method surpasses the previous DA embedding mapping method of [3] by 17.50%.

These results prove the effectiveness of encoding orthographic features into a shared space and then incorporating them into embedding spaces alignment.

5. Conclusion

In this paper, we have proposed a method to incorporate orthographic features for Arabic dialects embeddings alignment. Our method relies on CCA to encode orthographic features in a shared space. The aim is to capture orthographic variations across Arabic dialects. The encoded features are then used to extend the word vectors for learning multidialectal embedding. Our method is successfully used in both supervised and unsupervised approaches for cross-lingual embedding alignment. Experiments have been conducted using off-the-shelf DA embeddings of four Arabic dialects (MAG, EGY, LEV, and GLF). For multidialectal embedding evaluation, we have used the BLI task. Our method outperforms state-of-the-art cross-lingual alignment method under both supervised and unsupervised settings.

Future work will explore our DA alignment method for cross-dialect transfer learning in NLP tasks, such as text classification and sentiment analysis. Another direction of research work is to employ our method for DA unsupervised machine translation.

References

[1] N. Habash, M. Diab, and O. Rambow. "Conventional Orthography for Dialectal Arabic." In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 711–718. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/579_Paper.pdf.

[2] M. Artetxe, G. Labaka, and E. Agirre. "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, pp. 2289–2294.

[3] A. Erdmann, N. Zalmout, and N. Habash. "Addressing Noise in Multidialectal Word Embeddings." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 558–565. DOI: 10.18653/v1/P18-2089. URL: https://www.aclweb.org/anthology/P18-2089.

[4] T. Mikolov, Q. V. Le, and I. Sutskever. "Exploiting Similarities among Languages for Machine Translation." In: *CoRR* abs/1309.4168 (2013). arXiv: 1309.4168. URL: http://arxiv.org/abs/1309.4168.

[5] W. Ammar, G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer, and N. A. Smith. "Massively Multilingual Word Embeddings." In: *CoRR* abs/1602.01925 (2016). arXiv: 1602.01925. URL: http://arxiv.org/abs/1602.01925.

[6] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. "Word Translation Without Parallel Data." In: *arXiv preprint arXiv:1710.04087* (2017).

[7] P. Riley and D. Gildea. "Orthographic Features for Bilingual Lexicon Induction." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 390–394. DOI: 10.18653/v1/P18-2062. URL: https://www.aclweb.org/anthology/P18-2062.

[8] P. H. Schönemann. "A generalized solution of the orthogonal procrustes problem." In: *Psychometrika* 31.1 (Mar. 1966), pp. 1–10. DOI: 10.1007/bf02289451. URL: https://doi.org/10.1007/bf02289451.

[9] B. Tim. *Buckwalter transliteration*. 2002. URL: http://www.qamus.org/transliteration.htm.

[10] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. "Enriching Word Vectors with Subword Information." In: *arXiv preprint arXiv:1607.04606* (2016).