
Proceedings Paper:
Ezeani, I. orcid.org/0000-0001-8286-9997, Hepple, M. orcid.org/0000-0003-1488-257X, Onyenwe, I. et al. (1 more author) (2018) Multi-task projected embedding for Igbo. In: Sojka, P., Horák, A., Kopeček, I. and Pala, K., (eds.) Text, Speech, and Dialogue : 21st International Conference, Proceedings. 21st International Conference on Text, Speech, and Dialogue, 11-14 Sep 2018, Brno, Czech Republic. Springer , pp. 285-294. ISBN 9783030007935

https://doi.org/10.1007/978-3-030-00794-2_31

Multi-task Projected Embedding for Igbo

Ignatius Ezeani, Mark Hepple, Ikechukwu Onyenwe, and Chioma Enemuo

Department of Computer Science,
The University of Sheffield, United Kingdom.
{ignatius.ezeani,m.r.hepple,i.onyenwe,clenemuo1}@sheffield.ac.uk
http://www.sheffield.ac.uk

Abstract. NLP research on low resource African languages is often impeded by the unavailability of basic resources: tools, techniques, annotated corpora, and datasets. Besides the lack of funding for the manual development of these resources, building from scratch will amount to the reinvention of the wheel. Therefore, adapting existing techniques and models from well-resourced languages is often an attractive option. One of the most generally applied NLP models is word embeddings. Embedding models often require large amounts of data to train which are not available for most African languages. In this work, we adopt an alignment based projection method to transfer trained English embeddings to the Igbo language. Various English embedding models were projected and evaluated on the *odd-word*, *analogy* and *word-similarity* tasks intrinsically, and also on the diacritic restoration task. Our results show that the projected embeddings performed very well across these tasks.

Keywords: low-resource, Igbo, diacritics, embedding models, transfer learning

1 Background

The core task in this paper is embedding-based diacritic restoration. Training embedding models requires large amounts of data which are unavailable in low resource languages. Web-scraped data are often relied upon but they are of poor quality. Languages with diacritics have most of the words wrongly written with missing diacritics. Diacritic restoration helps to improve the quality of corpora for NLP systems.

This work focuses on Igbo, mainly spoken in the south-eastern part of Nigeria and worldwide by about 30 million people. Igbo has diacritic characters (Table **??**) which often determine the pronunciation and meaning of words with the same latinized spelling.

1.1 Previous Approaches

Key studies in diacritic restoration involve word-, grapheme-, and tag-based techniques [**?**]. Earlier examples include Yarowsky's works [**?**,**?**] which combined decision list with morphological and collocational information. POS-tags and language models have also been applied by Simard [**?**] to well resourced languages (French and Spanish). Hybrid of techniques are common with this task e.g. Yarowsky [**?**] used decision list, Bayesian

Char	Ortho	Tonal
a	–	à,á, ā
e	–	è,é, ē
i	ị	ì, í, ī, ị̀, ị́, ị̄
o	ọ	ò, ó, ō, ọ̀, ọ́, ō̩
u	ụ	ù, ú, ū, ụ̀, ụ́, ū̩
m	–	m̀,ḿ, m̄
n	ṅ	ǹ,ń, n̄

Table 1. Igbo diacritic complexity

classification and Viterbi decoding while Crandall [**?**] applied Bayesian- and HMM-based methods. Tufiş and Chiţu [**?**] combined the two approaches by backing off to character-based method when dealing with "unknown words".

However, these methods are mostly on well-resourced languages (French and Spanish) with comparatively limited diacritic complexity. Mihalcea *et al* [**?**] proposed an approach that used character based instances with classification algorithms for Romania. This inspired the works of Wagacha *et al* [**?**], De Pauw *et al* [**?**] and Scannell [**?**] on a variety of relatively low resourced languages. However, it is a common position that the word-based approach is superior to character-based approach for well resourced languages. Diacritic restoration can also be modelled as a classification task. For Maori, Cocks and Keegan [**?**] used naïve Bayes algorithms with word *n*-grams to improve on the character based approach by Scannell [**?**].

For Igbo, however, one major challenge to applying most of the techniques mentioned above that depend on annotated datasets is the lack of these datasets for Igbo e.g tags, morph-segmented or dictionaries. This work aims to apply a resource-light approach that is based on a more generalisable state-of-the-art representation model like word-embeddings which could also be tested on other tasks.

1.2 Igbo Diacritic Restoration

Igbo was among the languages in a previous work [**?**] with 89.5% accuracy using a version of their *lexicon lookup* methods, *LL2*. This technique used the most frequent word and a bigram model to determine the right replacement. However, we could not directly compare their work to ours as the task definitions are slight different. While their accuracy is based on the restoration of every word in a sentence, our work focuses on only the ambiguous words. Besides, their training corpus was too little (31k tokens and 4.3k types) to be representative and there was no language speaker in their team to validate their results. However, we re-implemented a version of the *LL2* and bigram model as our baseline for the restoration task reported in this work.

Ezeani *et al* [**?**] implemented a more complex set of n–gram models with similar techniques on a larger corpus but though they reported improved results, their evaluation method assumed a closed-world by training and testing on the same dataset. While a more standard evaluation method was used in [**?**], the data representation model was akin to *one-hot* encoding which is inefficient and could not easily handle large vocabulary sizes.

Another reason for using embedding models for Igbo is that diacritic restoration does not always eliminate the need for sense disambiguation. For example, the restored word *àkwà* could be referring to either *bed* or *bridge*. Ezeani *et al* [**?**] had earlier shown that with proper diacritics on ambiguous wordkeys (e.g. *akwa*), a translation system like *Google Translate* may perform better at translating Igbo sentences to other languages. This strategy, therefore, could be more easily extended to sense disambiguation in future.

Statement	***Google Translate***	**Comment**
O ji *egbe* ya gbuo *egbe*	He used his **gun** to kill *gun*	wrong
O ji **égbè** ya gbuo **égbé**	He used his **gun** to kill **kite**	correct
Akwa ya di n'elu *akwa* ya	It was on the **bed** in his room	fair
Ákwà ya di n'elu **àkwà** ya	his **clothes** on his **bed**	correct
Oke riri *oke* ya	Her addiction	confused
Òké riri **òkè** ya	**Mouse** ate his **share**	correct
O jiri *ugbo* ya bia	He came with his *farm*	wrong
O jiri **ụgbọ** ya bia	He came with his **car**	correct

Table 2. Disambiguation challenge for *Google Translate*

2 Experimental Setup

Our experimental pipeline follows four fundamental stages:

1. pre-processing of data (Section **??**);
2. building embedding models (Section **??**);
3. enhancing embedding models (Section **??**);
4. evaluation of models (Section **??**)

Models are intrinsically evaluated on the *word similarity*, *analogy* and *odd-word identification* tasks as well as the key process of diacritic evaluation.

2.1 Experimental Data

We used the Igbo-English parallel bible corpora, available from the *Jehova Witness* website[1], for our experiments. There are 32,416 aligned lines of text, bible verses, and chapter headings, from both languages. Total token sizes, without punctuations, are 902,429 and 881,771 with vocabulary lengths of 16,084 and 15,000 for Igbo and English respectively.

Over 50% of both the Igbo tokens (595,221) and vocabulary words (8,750) have at least one diacritic character. There are 550 ambiguous *wordkeys*[2]. Over 97% of the ambiguous wordkeys have 2 or 3 variants.

[1] `jw.org`

[2] A *wordkey* is a word stripped of its diacritics if it has any. Wordkeys could have multiple diacritic variants, one of which could be the same as the wordkey itself.

2.2 Embedding Models

Inspired by the concept of the universality of meaning and representation (Figure **??**) in distributional semantics, we developed an embedding-based diacritic restoration technique. Embedding models are very generalisable and therefore will constitute essential resources for Igbo NLP work. We used both trained and projected embeddings, as defined below, for our tasks.

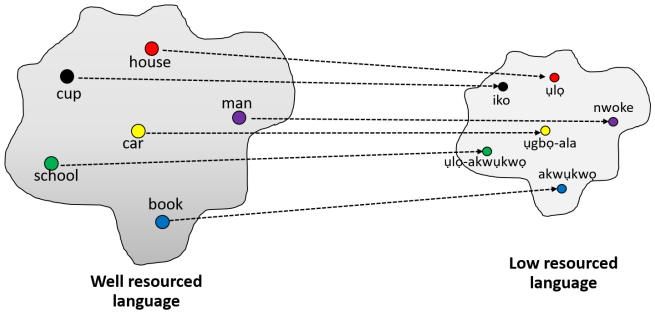

Fig. 1. Embedding Projection

Embedding Training We built the **igBbltrain** embedding from the data described in Section **??** using the Gensim *word2vec* Python libraries [**?**]. Default configurations were used apart from optimizing $dimension$ and $window\_size$ parameters to 140 and 2 respectively on the **Basic** restoration method described in Section **??** [3].

Embedding Projection We adopt an alignment-based projection method similar to the one described in [**?**]. It uses an Igbo-English alignment dictionary $A^{I|E}$ with a function $\boldsymbol{f}(w_i^I)$ that maps each Igbo word w_i^I to all its co-aligned English words $w_{i,j}^E$ and their counts $c_{i,j}$ as defined in Equation **??**. $|V^I|$ is the vocabulary size of Igbo and n is number of co-aligned English words.

$$\begin{aligned} A^{I|E} &= \{w_i^I, \boldsymbol{f}(w_i^I)\}; i = 1..|V^I| \\ \boldsymbol{f}(w_i^I) &= \{w_{i,j}^E, c_{i,j}\}; j = 1..n \end{aligned} \tag{1}$$

The projection is formalised as assigning the weighted average of the embeddings of the co-aligned English words $w_{i,j}^E$ to the Igbo word embeddings $\mathbf{vec}(w_i^I)$ [**?**]:

$$\mathbf{vec}(w_i^I) \leftarrow \frac{1}{C} \sum_{w_{i,j}^E, c_{i,j} \in f(w_i^I)} vec(w_{i,j}^E) \cdot c_{i,j} \tag{2}$$

[3] The pre-trained Igbo model from *fastText Wiki word vectors* project [**?**] was also tested but its performance was so bad that we had to drop it.

where $C \leftarrow \sum_{c_{i,j} \in f(w_i^I)} c_{i,j}$

Using this projection method, we built 5 additional embedding models for Igbo:

- **igBblproj** from a model we trained on the English bible.
- **igGNproj** from the pre-trained *Google News*[4] *word2vec* model.
- **igWkproj** from *fastText* Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset.
- **igSwproj** from same as **igWkproj** but with subword information.
- **igCrlproj** from *fastText* Common Crawl dataset

Table **??** shows the vocabulary lengths ($Vocabs^L$), and the dimensions ($Dimension$) of each of the models used in our experiments. While the pre-trained models and their projections have vector sizes of 300, our trained **IgboBible** performed best with vector size of 140 and so we trained the **IgboEnBbl** with the same dimension.

$Model$	$Dimension$	$Vocabs^I$	$Vocabs^E$	$Data$
igBbltrain	140	4968	–	902.5k
igBblproj	140	4057	6.3k	881.8k
igGNproj	300	3046	3m	100bn
igWkproj	300	3460	1m	16bn
igSwproj	300	3460	1m	16bn
igCrlproj	300	3510	2m	600bn

Table 3. Igbo and English models: vocabulary, vector and training data sizes

2.3 Enhancing embedding models

For this experiment, our dataset consists of 29 ambiguous *wordkeys*[5] from our corpus. For each wordkey, we keep a list of sentences (excluding punctuations and numbers), each with a place-holder (see Table **??**) to be replaced with the correct variant of the wordkey.

Variant	*Left context*	*Placeholder*	*Right context*	Meaning
àkwá	ka okwa nke kpokotara	____	o na-eyighi eyi otu	egg
ákwà	a kpara akpa mee	____	ngebichi nke onye na-ekwe	cloth
ákwá	ozugbo m nuru mkpu	____	ha na ihe ndi a	cry

Table 4. Instances of the wordkey ***akwa*** in context

In both trained and projected embedding models, vectors are assigned to each word in the dictionary, and that includes each diacritic variant of a wordkey. The **Basic**

[4] https://code.google.com/archive/p/word2vec/

[5] Highly dominant variants or very rarely occurring *wordkeys* were generally excluded from the datasets.

restoration process (Section **??**) uses this initial embedding model *as-is*. The models are then refined by "learning" new embeddings for each variant that correlate more with its context words embeddings.

For example, let mcw_v contain the top n (say n = 20) of the most co-occurring words of a certain variant, v and their counts, c. The diacritic embedding is derived by replacing each diacritic variant vector with the weighted average of the vectors of its most co-occurring words (see Equation (**??**)).

$$\textbf{diac}_{\textbf{vec}} \leftarrow \frac{1}{|mcw_v|} \sum_{w \in mcw_v} w_{vec} * w_c \tag{3}$$

where w_c is the 'weight' of w i.e. the count of w in mcw_v.

2.4 Model Evaluation

We evaluate the models on their performances on the following NLP tasks: *odd-words*, *analogy* and *word similarity* and diacritic restoration. As there are no standard datasets for these tasks in Igbo, we had auto-generate them from our data or transfer existing ones from English. Igbo native speakers were used to refine and validate instances of the dataset or methods used.

The *odd word* In this task, the model is used to identify the *odd word* from a list of words e.g. *breakfast, cereal, dinner, lunch* $\rightarrow$ *"cereal"*. We created four simple categories of words Igbo words (Table **??**) that should naturally be mutually exclusive. Test instances were built by randomly selecting and shuffling three words from one category and one from another e.g. *ọkpara, nna, ọgaranya, nwanne* $\rightarrow$ *ọgaranya*.

category	**Igbo words**
nouns(family) *e.g. father, mother*	ada, ọkpara, nna, nne, nwanna, nwanne, di, nwunye
adjectives *e.g. tall, rich*	ọcha, ọgaranya, ogbenye, ogologo, oji, ọjọọ, okenye, ọma
nouns(humans) *e.g. man, woman*	nwaanyị, nwoke, nwata, nwatakịrị, agbọghọ, okorobịa
numbers *e.g. one, seven*	otu, abụọ, atọ, anọ, ise, isii, asaa, asatọ, itoolu, iri

Table 5. Word categories for *odd word* dataset

Analogy This is based on the concept of analogy as defined by [**?**] which tries to find y_2 in the relationship: $x_1 : y_1$ as $x_2 : y_2$ using vector arithmetic e.g $king - man + woman \approx queen$. We created pairs of opposites for some common nouns and adjectives (Table **??**) and randomly combined them to build the analogy data e.g. *di* (husband) – *nwoke* (man) + *nwaanyị*(woman) $\approx$ *nwunye*(wife) ?

category	opposites
oppos-nouns	nwoke:nwaanyị, di:nwunye, okorobịa:agbọghọ, nna:nne, ọkpara:ada
oppos-adjs	agadi:nwata, ọcha:oji, ogologo:mkpụmkpụ, ọgaranya:ogbenye

Table 6. Word pair categories for *analogy* dataset

Word Similarity We created Igbo word similarity dataset by transferring the standard *wordsim353* dataset [**?**]. Our approach used *Google Translate* to translate the individual word pairs in the combined dataset and return their human similarity scores. We removed instances with words that could not be translated (e.g. cell → *cell* & phone→*ekwentị*,7.81) and those with translations that yield compound words (e.g. situation → *ọnọdụ* & conclusion → *nkwubi okwu*,4.81)[6].

Diacritic restoration process The restoration process computes the cosine similarity of the variant and context vectors and chooses the most similar candidate. For each wordkey, wk, candidate vectors, $D^{wk} = \{d_1, ..., d_n\}$, are extracted from the embedding model on-the-fly. C is defined as the list of the context words and vec_C is the context vector of C (Equation (**??**)).

$$\mathbf{vec_C} \leftarrow \frac{1}{|C|} \sum_{w \in C} vec_w \tag{4}$$

$$\mathbf{diac_{best}} \leftarrow \underset{d_i \in D^{wk}}{\mathbf{argmax}}\ sim(\mathbf{vec_C}, d_i) \tag{5}$$

3 Results and Discussion

Our results on the odd-word, analogy and word-similarity tasks (Table **??**, Figure **??**) indicate that the projected embedding models, in general, capture concepts and their relationships better. This is not surprising as the trained model, **igBible**, and the one from its parallel English data, **igEnBbl** are too little and cover only religious data. Although **igWkSbwd** includes subword information which should be good for an agglutinative language like Igbo, these subword patterns are different from the patterns in Igbo. Generally, the models from the news data, **igGNews, igWkNews**, did well on these tasks.

On the diacritic restoration task, the embedding based approaches, with semantic information, generally performed comparatively well with respect to the *n*-gram models that capture syntactic details better. **IgBible**'s performance is impressive especially as it outperformed the bigram model[7].

Expectedly, compared to other projected models, **IgBible** and its parallel, **IgEnBbl**, clearly did better on this task. **IgBible** was originally trained with the same dataset and language of the task and its vocabulary directly aligns with that of **IgEnBbl**. Clearly,

[6] An alternative considered is to combine the word e.g. *nkwubi okwu* → **nkwubi-okwu** and update the model with a projected vector or a combination of the vectors of constituting words.

[7] We intend to implement higher level *n*-gram models.

the enhanced diacritic embeddings improved the performances of all the models which is expected as each variant is pulled to the center of its most co-occurring words.

	Odd-word	**Similarity**	**Analogy**	
Models	*Accuracy*	*Correlation*	*nouns*	*adjectives*
igBible	78.27	48.02	23.81	06.67
igGNews	**84.24**	60.00	64.29	**56.67**
igEnBbl	75.26	58.96	54.76	13.33
igWkSbwd	84.18	58.56	64.29	50.00
igWkCrl	80.72	**62.07**	78.57	21.37
igWkNews	81.51	59.69	**80.95**	50.00

Table 7. Trained and Project Embeddings on odd-word prediction

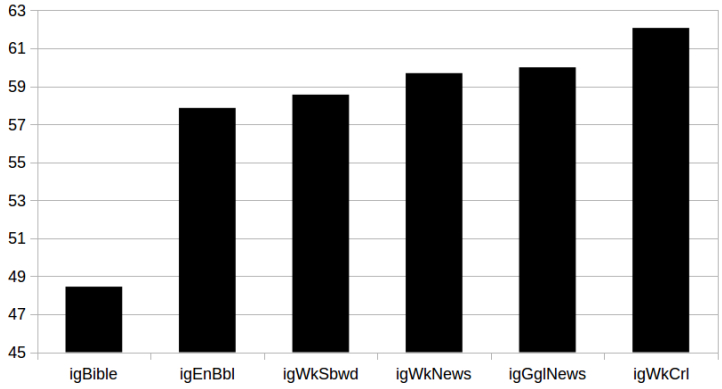

Fig. 2. Worst-to-Best Word Similarity Correlation Performance

4 Conclusion and Future Research Direction

This work contributes to the IgboNLP[8] [**?**] project. The goal of the project is to build a framework that can adapt, in an effective and efficient way, existing NLP tools to support the development of Igbo. In this paper, we demonstrated that projected embedding models can outperform the ones built with small language data on a variety of NLP tasks on low resource languages.

We also introduced a technique for learning diacritic embeddings which could be applied to the diacritic restoration task. Our next focus is to refine our techniques and datasets and train models with sub-word information as well as consider sense disambiguation task.

[8] See `igbonlp.org`

Baselines: *n*-gram models

	Unigram	*Bigram*
	72.25%	80.84%

Embedding models

	Accuracy		Precision		Recall		F1	
	Basic	Diac	Basic	Diac	Basic	Diac	Basic	Diac
igBible	69.28	**82.26**	61.37	77.96	61.90	**82.28**	57.19	**76.16**
igEnBbl	64.72	78.71	59.60	75.18	59.65	79.52	50.51	72.93
igGNews	57.57	74.14	32.20	72.50	49.00	74.56	19.06	62.47
igWkSbwd	62.10	73.83	13.82	73.81	47.64	74.03	10.65	66.62
igWkCrl	60.78	73.30	40.07	**78.02**	49.16	76.24	25.36	68.62
igWkNews	61.07	72.97	14.16	76.04	46.10	75.14	8.31	65.20

Table 8. Performances of Basic and Diacritic versions of the *Trained* and *Projected* embedding models on diacritic restoration tasks

References

1. Crandall, D., Automatic Accent Restoration in Spanish text, 2005, `http://www.cs.indiana.edu/~djcran/projects/674\_final.pdf`, [Online; accessed 7-January-2016]
2. De Pauw, G., De Schryver, G. M., Pretorius, L., Levin L., 2011 Introduction to the Special Issue on African Language Technology, Language Resources and Evaluation, 45, 263-269, Springer Online
3. Ezeani, I., Hepple, M., Onyenwe, I., 2016, Automatic Restoration of Diacritics for Igbo Language, Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno , Czech Republic, Sep 12–16, Springer International Publishing, 198–205, 978-3-319-45510-5
4. Ezeani, I., Hepple, M., Onyenwe, I., 2017, Lexical Disambiguation of Igbo using Diacritic Restoration. In Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications (pp. 53-60).
5. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E., 2001, Placing Search in Context: The Concept Revisited., In Proceedings of the 10th international conference on World Wide Web (pp. 406-414). ACM.
6. Francom, J., Hulden, M, Diacritic Error Detection and Restoration via POS tags, Proceedings of the 6th Language and Technology Conference, 2013
7. Guo, J., Che, W., Yarowsky, D., Wang, H., Liu, T., 2015, Cross-Lingual Dependency Parsing Based on Distributed Representations, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Vol1: Long Papers), 1234–1244
8. Mihalcea, R., 2002, Diacritics Restoration: Learning from Letters Versus Learning from Words, Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, CICLing '02, 3-540-43219-1, 339–348, 10, http://dl.acm.org/citation.cfm?id=647344.724003, 724003, Springer-Verlag, London, UK
9. Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013, Efficient Estimation of Word Representations in Vector Space, arXiv preprint arXiv:1301.3781
10. Onyenwe, I. E., Hepple, M., Chinedu, U., Ezeani, I., 2018, A Basic Language Resource Kit Implementation for the IgboNLP Project, ACM Trans. Asian Low-Resource. Lang. Inf. Process., February 2018, vol 17.2, Jan,2018, issn 2375-4699, pages 10:1–10:23 ACM

11. Radim Řehůřek and Petr Sojka, 2010, Software Framework for Topic Modelling with Large Corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45–50, May, 22, ELRA, Valletta, Malta, http://is.muni.cz/publication/884893/en
12. , John Cocks and Te-Taka Keegan A Word-based Approach for Diacritic Restoration in Māori, Proceedings of the Australasian Language Technology Association Workshop 2011, December,2011, Canberra, Australia, 126–130, http://www.aclweb.org/anthology/U/U11/U11-2016
13. Tufiş, D. and Chiţu, A., Automatic Diacritics Insertion in Romanian Texts, Proceedings of the International Conference on Computational Lexicography, Pecs, Hungary,185–194, 1999
14. Scannell, K. P., 2011, Statistical Unicodification of African Languages, Language Resource Evaluation, 45, 3, Sep, 375–386,Springer-Verlag New York, Inc., Secaucus, NJ, USA
15. Simard, Michel, Automatic Insertion of Accents in French Text, Proceedings of the Third Conference on Empirical Methods for Natural Language Processing, 27–35, 1998
16. Wagacha P. W., De Pauw, G., Githinji P. W., 2006 A Grapheme-based Approach to Accent Restoration in G̃ikũyũ, Fifth International Conference on Language Resources and Evaluation
17. Yarowsky, D., A Comparison of Corpus-based Techniques for Restoring Accents in Spanish and French Text, Proceedings, 2nd Annual Workshop on Very Large Corpora, 1994, Kyoto, 19–32
18. Yarowsky, D., 1999, Corpus-based Techniques for Restoring Accents in Spanish and French Text, Natural Language Processing Using Very Large Corpora, Kluwer Academic Publishers, 99-120,1999
19. Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas, Enriching Word Vectors with Subword Information, arXiv preprint arXiv:1607.04606, 2016