Transferred Embeddings for Igbo Similarity, Analogy and Diacritic Restoration Tasks

Ignatius Ezeani Mark Hepple Ikechukwu Onyenwe Chioma Enemuo
Department of Computer Science,
The University of Sheffield, United Kingdom.
https://www.sheffield.ac.uk/dcs
{ignatius.ezeani, m.r.hepple, i.onyenwe, clenemuo1}@shef.ac.uk

Abstract

Existing NLP models are mostly trained with data from well-resourced languages. Most minority languages face the challenge of lack of resources - data and technologies - for NLP research. Building these resources from scratch for each minority language will be very expensive, time-consuming and amount largely to unnecessarily re-inventing the wheel. In this paper, we applied transfer learning techniques to create Igbo word embeddings from a variety of existing English trained embeddings. Transfer learning methods were also used to build standard datasets for Igbo word similarity and analogy tasks for intrinsic evaluation of embeddings. These projected embeddings were also applied to the diacritic restoration task. Our results indicate that the projected models not only outperform the trained ones on the semantic based tasks of analogy, word-similarity and odd-word identifying, but they also achieve enhanced performance on the diacritic restoration with learned diacritic embeddings.

1 Background

Most NLP systems are modelled with English data. One major challenge to adapting these systems for low resource languages is lack of good quality data. Such languages often rely on poor quality web-crawled data. In our case the target language is Igbo, a language spoken by over 30 million indigenes who live mainly in the south-eastern part of Nigeria but also in different parts of the world.

Inspite of the relatively large number of speakers, Igbo is critically low-resourced in terms of NLP research (Onyenwe et al., 2018). Recent efforts to develop resources for Igbo include the design of Igbo POS tagset (Onyenwe et al., 2014), and the tagset refinement (Onyenwe et al., 2015) as well as the development of Igbo POS-tagger (Onyenwe, 2017). Works are also on-going with its automatic diacritic restoration and lexical disambiguation (Ezeani et al., 2016) (Ezeani et al., 2017) and morphological segmentation (Enemouh et al., 2017).

1.1 Embedding Models

Word embeddings are generic semantic representations from corpus. It enhances the concept of distributional hypothesis (Harris, 1954) and count-based distributional vectors (Baroni and Lenci, 2010) and provides an alternative to the *one task, one model* approach. Their application areas span most NLP tasks and other fields such as biomedical, psychiatry, psychology, philology, cognitive science and social science (Altszyler et al., 2016). There are many approaches to training embedding models, however predictive (Mikolov et al., 2013a) and count-based (Pennington et al., 2014) models are very commonly used.

Ideally, a model trained in one language should capture similar semantic distribution in other languages. Since the large amount of data required to train such a model are not often available for low resource languages, transfer learning techniques could be used to project learned knowledge from one language to another.

Proceedings of SemDeep-3, the 3rd Workshop on Semantic Deep Learning, pages 30–38
Santa Fe, New Mexico, USA, August 20, 2018.

1.2 Transfer and Cross-lingual Learning

Transfer learning generally refers to the transfer of knowledge acquired in one domain in solving a problem in another domain. It is commonly applied when the target domain training data is limited (Weiss et al., 2016). With transfer learning we could take advantage of a parallel data that exist across languages in the form of word-aligned data, sentence-aligned data (e.g. Europarl corpus), document-aligned data (e.g. Wikipedia), lexicon (bilingual or cross-lingual dictionary) or even zero-shot learning with no parallel data.

In a survey of cross-lingual embedding models (Ruder, 2017), four different approaches were identified, *monolingual mapping* (Mikolov et al., 2013b; Faruqui and Dyer, 2014; Guo et al., 2015) which trains embeddings on large monolingual corpora and then linearly maps a target language word to its corresponding source language embedding vectors; *pseudo-cross-lingual* (Duong et al., 2016; Gouws and Søgaard, 2015; Xiao and Guo, 2014) which trains embeddings with a pseudo-cross-lingual corpus i.e mixing contexts from different languages; *cross-lingual* (Hermann and Blunsom, 2013; Hermann and Blunsom, 2014; Kočiskỳ et al., 2014) trains embeddings on a parallel corpus constraining similar words to be close to each other in a shared vector space; *joint optimization* (Klementiev et al., 2012; Luong et al., 2015; Gouws et al., 2015) trains models on parallel or monolingual data but jointly optimise a combination of monolingual and cross-lingual losses. In this paper, we will adopt the projection approach described in (Guo et al., 2015).

2 Experimental Setup

Our experimental data consists of a collection of Igbo texts from the *Igbo Bible* and the translation of the *Universal Declaration of Human Rights*, two short novels: an Igbo version of *Eze Goes to School* and another Igbo novel *Mmadụ Ka A Na Arịa*. The pipeline has three stages. It starts with building the embedding models using training or projection methods (section 2.1). The next stage enhances the diacritic words with the embeddings of the its co-aligned English words (section 3.4.2). Lastly, the diacritic restoration is implemented as laid out in section 3.4.3.

In this experiment, we used only the Igbo-English parallel bible corpora, available from the *Jehova Witness* website[1], for the word alignment and projection of embedding models. The parallel data consist of 32,416 aligned lines of text. Additional data from the novels (3179 lines) and official documents (90 lines) make up the rest of the 35,685 lines of text with token sizes of 962,747 (without punctuations)[2] and vocabulary length 16,586 we used.

Although only 34% (328,591) of all tokens have diacritics, 54.8% (9,090) of vocabulary words are diacritic marked. There are 795 ambiguous *wordkeys*. 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. Over 97% of the ambiguous wordkeys have 2 or 3 variants.

2.1 Building Igbo Embedding Models

In this work, we used both trained and projected embeddings for our tasks. We built the **igBible** embedding from the data using the Gensim *word2vec* Python libraries (Řehůřek and Sojka, 2010) with its default parmeters. We also used the **igWiki**, a pre-trained Igbo model from *fastText Wiki* project (Bojanowski et al., 2016), but it was removed due to its unstable performance across tasks which we could not resolve at the time of submission of this paper.

For the embedding transfer, we applied an alignment-based projection method (Guo et al., 2015). An Igbo-English alignment dictionary $A^{I|E}$ uses 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 1. $|V^I|$ is the vocabulary size of Igbo and n is number of co-aligned English words.

[1] `jw.org`

[2] There will be 1,138,036 in total with punctuations, symbols and digits

Model	Igbo Vocabs	Dimensions	Eng Vocabs	Train data
igBible	4968	300	–	902.5k
igEnBbl	4057	300	6.3k	881.8k
igGglNews	3046	300	3m	100bn
igWkNews	3460	300	1m	16bn
igWkSbwd	3460	300	1m	16bn
igWkCrl	3510	300	2m	600bn

Table 1: Igbo and English models: vocabulary, vector and training data sizes

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

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 $\mathbf{vec}(w_i^I)$ (Guo et al., 2015):

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

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:

- **igEnBbl** from a model we trained on the English bible.
- **igGNews** from the pre-trained *Google News*[3] *word2vec* model.
- **igWkNews** from *fastText* Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset.
- **igWkSbwd** from same as **igWkNews** but with subword information.
- **igWkCrl** from *fastText* Common Crawl dataset

Table 1 shows the vocabulary lengths (*vocabs*), and the dimensions (*vectors*) of each of the models used in our experiments.

3 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.

3.1 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 2) 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*.

3.2 Analogy

This is based on the concept of analogy as defined by (Mikolov et al., 2013a) 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 noun and adjectives (Table 3) and randomly combined them to build the analogy data e.g. *di* (husband) – *nwoke* (man) + *nwaanyị*(woman) $\approx$ *nwunye*(wife) ?

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

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 2: Word categories for *odd word* dataset

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 3: Word pair categories for *analogy* dataset

3.3 Word Similarity

We created Igbo word similarity dataset by transferring the standard *wordsim353* dataset (Finkelstein et al., 2001). 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)[4].

3.4 Diacritic restoration

The absence of proper diacritics in Igbo words causes ambiguities and may affect MT systems (Ezeani et al., 2016; Ezeani et al., 2017) (see Table 4). There are word-, grapheme-, and tag-based techniques (Francom and Hulden, 2013) for this task involving a huge amount of annotated data (Yarowsky, 1994; Yarowsky, 1999) which Igbo does not have. Techniques for low-resource languages (Mihalcea, 2002; Wagacha et al., 2006; De Pauw et al., 2011) but were not applied to Igbo. So far, works on Igbo used either too little data (Scannell, 2011), non-generic methods (Ezeani et al., 2016; Ezeani et al., 2017).

Statement	*Google Translate*	Comment
O ji *egbe* ya gbuo *egbe*	He used his **gun** to kill *gun*	wrong
O ji **ígbè** ya gbuo **ígbé**	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 4: Translation challenge for *Google Translate* (Ezeani et al., 2017)

3.4.1 Building the baseline n-grams

As our baseline, we used standard n-gram models with back-off and 10-fold cross validations. We focused on restoring only the ambiguous sets with a fair distribution of variants. To achieve this, we set a maximum threshold of 70% for any of the variants in a set i.e. if choosing the most common variant from a set gets 70% accuracy on that set, it is disqualified, leaving us with 215 (27%) of all 795 ambiguous wordkeys. Figure 2 shows that there is no significant improvement after the bigram model.

[4] 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.

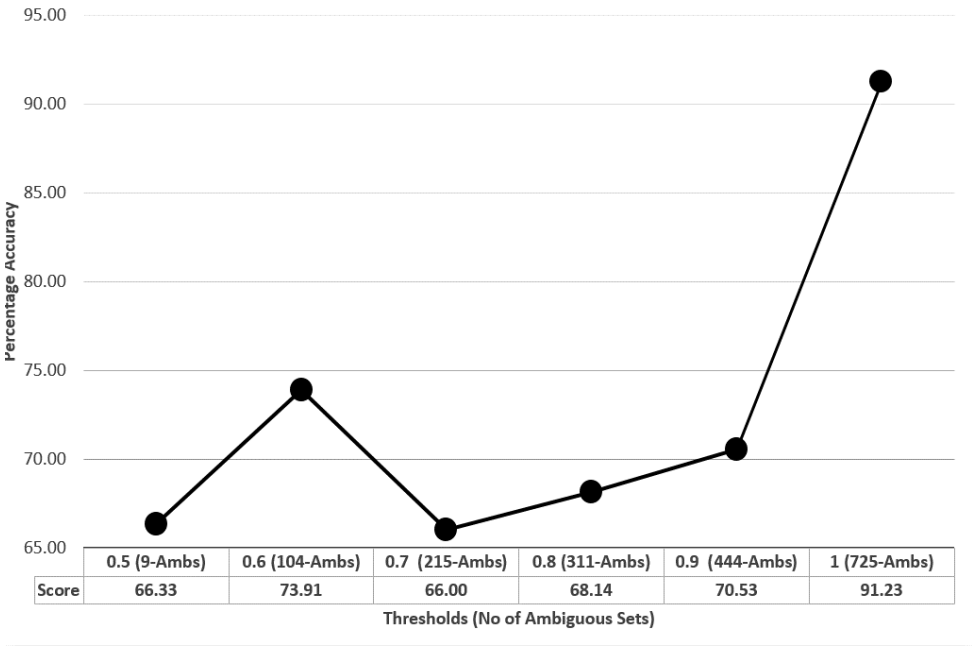

Figure 1: Average accuracy scores for all n-gram models: Thresholds [0.5 .. 1.0]

3.4.2 Deriving *diacritic* embedding models

The word **akwa** without context could mean **àkwá**(egg), **ákwà**(cloth), **ákwá**(cry/wail), **àkwà**(bed/bridge). The task is to ensure that the embedding for each of the variants of **akwa** exists in the model and is represented by the weighted combination of each of the most co-occurring words, mcw_v.

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

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

3.4.3 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 context words (i.e. all the words in the same sentence) and vec_C is the context vector of C (Equation (4)).

$$\textbf{vec}_{\textbf{C}} \leftarrow \frac{1}{|C|} \sum_{w \in C} vec_w \tag{4}$$

$$\textbf{diac}_{\textbf{best}} \leftarrow \underset{d_i \in D^{wk}}{\textbf{argmax}}\ sim(\textbf{vec}_{\textbf{C}}, d_i) \tag{5}$$

4 Results and Discussion

Our results on the odd-word, analogy and word-similarity tasks indicate that the projected embeddings (Table 5, Figure 3) capture better general concepts and their relationships. 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 6, the results compare the *basic* model (i.e. as trained or projected) with the *diac* (i.e. with variant vectors enhanced with the embeddings of their most co-occuring words.

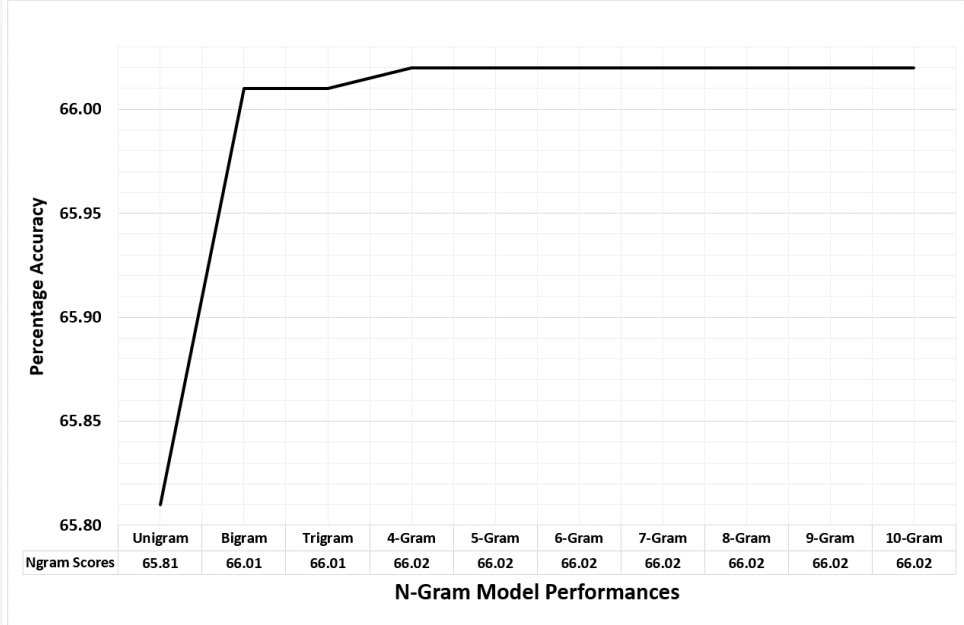

Figure 2: N-Gram accuracy scores for Threshold=0.7 (215) ambiguous sets

These models with semantic information, generally out-performed the *n*-gram models that capture more of syntactic details.

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

Clearly, the learned 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 5: Trained and Project Embeddings on odd-word prediction

5 Conclusion and Future Research Direction

This work is part of the IgboNLP[5] (Onyenwe et al., 2018) project which aims at build a framework that can adapt, in an effective and efficient manner, existing NLP tools to support the development of NLP resources for Igbo. In this paper, we showed that, projected embedding models can outperform the one built with small language data on a variety of tasks. 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 subword information as well as consider sense disambiguation task.

[5]See `igbonlp.org`

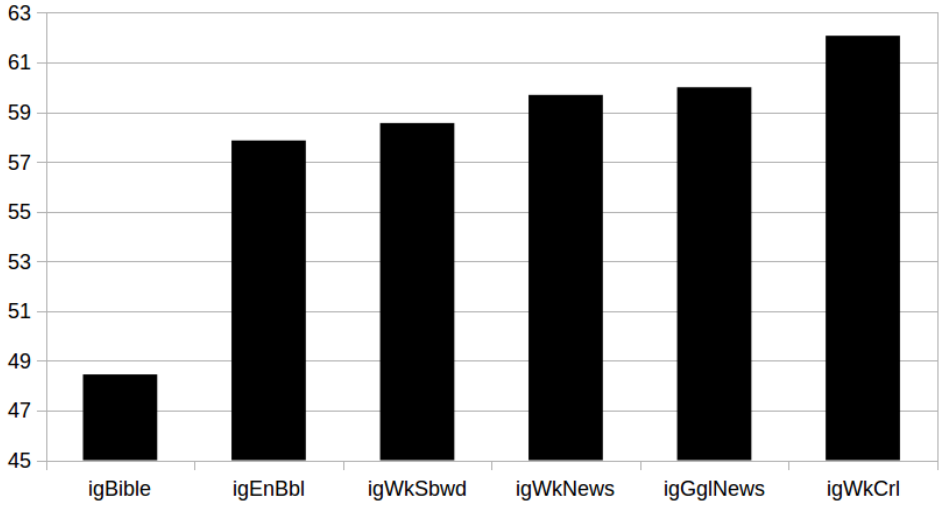

Figure 3: Worst-to-Best Word Similarity Correlation Performance

Baselines: *n*-gram models

	Unigram	*Best N-gram*
	65.81%	66.02%

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 6: Performances of Basic and Diacritic versions of the *Trained* and *Projected* embedding models on diacritic restoration tasks

References

Edgar Altszyler, Mariano Sigman, Sidarta Ribeiro, and Diego Fernández Slezak. 2016. Comparative study of lsa vs word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520*.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

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

G. De Pauw, G. M. De Schryver, L. Pretorius, and L. Levin. 2011. Introduction to the Special Issue on African Language Technology. *Language Resources and Evaluation*, 45:263–269.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*.

C. Enemouh, M. Hepple, I. Ezeani, and I. Onyenwe. 2017. Morph-inflected word detection in igbo via bitext. *Widening NLP Workshop co-located with ACL 2017, Vancouver, July 30th 2017*.

Ignatius Ezeani, Mark Hepple, and Ikechukwu Onyenwe, 2016. *Automatic Restoration of Diacritics for Igbo Language*, pages 198–205. Springer International Publishing, Cham.

Ignatius Ezeani, Mark Hepple, and Ikechukwu Onyenwe. 2017. Lexical disambiguation of igbo through diacritic restoration. *SENSE 2017*, page 53.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Jerid Francom and Mans Hulden. 2013. Diacritic error detection and restoration via part-of-speech tags. *Proceedings of the 6th Language and Technology Conference*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*, pages 748–756.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1234–1244.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Karl Moritz Hermann and Phil Blunsom. 2013. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *Proceedings of COLING 2012*, pages 1459–1474.

Tomáš Kočiskỳ, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. *arXiv preprint arXiv:1405.0947*.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Rada F Mihalcea. 2002. Diacritics restoration: Learning from letters versus learning from words. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 339–348. Springer.

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*.

Ikechukwu E Onyenwe, Chinedu Uchechukwu, and Mark R Hepple. 2014. Part-of-speech tagset and corpus development for igbo, an african language. *LAW VIII - The 8th Linguistic Annotation Workshop.*, pages 93–98.

Ikechukwu Onyenwe, Chinedu Uchechukwu, Mark Hepple, and Ignatius Ezeani. 2015. Use of Transformation-Based Learning in Annotation Pipeline of Igbo, an African Language. In *Recent Advances in Natural Language Processing, Hissar, Bulgaria*. Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects.

Ikechukwu E Onyenwe, Mark Hepple, Uchechukwu Chinedu, and Ignatius Ezeani. 2018. A basic language resource kit implementation for the igbonlp project. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(2):10:1–10:23, January.

Ikechukwu Onyenwe. 2017. Developing methods and resources for automated processing of the african language igbo. *Doctoral dissertation*.

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.

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*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902.

Kevin P. Scannell. 2011. Statistical unicodification of african languages. *Language Resource Evaluation*, 45(3):375–386, September.

Peter W. Wagacha, Guy De Pauw, and Pauline W. Githinji. 2006. A Grapheme-based Approach to Accent Restoration in Gĩkũyũ. In *In Proceedings of the fifth international conference on language resources and evaluation*.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data*, 3(1):9.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129.

D. Yarowsky. 1994. A Comparison of Corpus-based Techniques for Restoring Accents in Spanish and French Text. In *Proceedings, 2nd Annual Workshop on Very Large Corpora*, pages 19–32, Kyoto.

D. Yarowsky, 1999. *Corpus-based Techniques for Restoring Accents in Spanish and French Text*, pages 99–120. Kluwer Academic Publishers.