Harnessing Cross-lingual Transfer Learning Techniques to Facilitate Interventions for Low-resourced Languages

by

Thapelo A. Sindane
(U20772077)

A thesis submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science
in the
Faculty of Engineering, Built Environment and Information Technology

University of Pretoria

Supervisors
Prof Vukosi Marivate
Dr Abiodun Modupe

Declaration

I, Thapelo Andrew Sindane, declare that this research project is my own work. All information sources used to conduct this research project have been referenced. This work has never been submitted to any other university for any other degree.

Signed: 

03 September 2024

Date: _____

Publication data:
Thapelo A. Sindane. Harnessing Cross-lingual Transfer Learning Techniques to Facilitate Interventions for Low-resourced Languages.. value(document.document), University of Pretoria, Department of Computer Science, cnr Lynnwood Road and, Roper St, Hatfield, Pretoria, 0083, September 2023.

Harnessing Cross-lingual Transfer Learning Techniques to Facilitate Interventions for Low-resourced Languages.

by

Thapelo A. Sindane
E-mail: Sindane.thapelo@tuks.co.za

Abstract

The world continues to witness increasingly complex technological, economic, and societal advancements at an accelerated pace in the space of Natural Language Processing (NLP) and Artificial Intelligence (AI). The availability of massive digital data in various forms such as language data, image data, and numeric data plays a profound role in supporting this upward trend. For example, the availability of tremendous volumes of English data and other high internet prevalent languages unlocks the ability to develop high-quality language technologies such as Generative AI systems, Question Answering systems, Translation systems, and other societally impactful technologies we see today. This new era unfolds a simple yet efficacious equation that takes the form (increased datasets = increased performance) operating with proportionality mechanics. Despite the remarkable strides, a concerning consequence has emerged − a widening horizontal divide among globally spoken languages. A divide that highlights disparities of benefits from available language technologies across the 7000-plus spoken languages. Key impedes that emerge in addressing such disparities for the underserved languages include data availability, data benchmarking, scaling, internet prevalence, sustainable pipelines, coverage, and lack of expertise. In this work, we extensively scrutinize some of these concerns by first grounding our work in the context of South African languages. South Africa has 12 official languages with varying states of resource-prevalence which provided a perfect case to demonstrate our proposed remedial approaches. To address benchmarking we proposed standard datasets for all spoken languages; Scaling is addressed by showcasing the use of bilingual lexicons as a resource with much higher linguistic

coverage to define various techniques that continuously improve our machine learning models; and Coverage is demonstrated by accounting for all South African languages in the development of technologies.

The main objective of this thesis is to investigate cross-lingual embeddings as cheaper interventions to administer transfer capabilities of various machine learning models across various downstream tasks, in order to foster the development, and accessibility of local technologies for low-resourced languages. Cross-lingual embeddings are intra-semantic and inter-translation equivalent representations between high-resourced and low-resourced languages. For this work, these cross-lingual embeddings have demonstrated efficacy in tasks such as News Headlines Classification (NHC), Named Entity Recognition (NER), Part of Speech (POS) Tagging, Machine Translation (MT), and have shown great potential for the development of localized technologies. The investigations showed that training NLP models with cross-lingual embeddings enhances both transfer and learning-from-scratch capabilities compared to monolingual embedding training. This study also highlighted that increasing supervision signals such as bilingual lexicons for training cross-lingual embeddings also improves their performance. Furthermore, our investigations indicated that no single cross-lingual model works well across all languages. We were able to address 4 key performance point and we hope the interventions proposed in this study will have a positive impact on the socio-economic status of South Africa and can be scaled to other contexts to empower societies and businesses. We released our code and datasets here [1,2].

Keywords: Natural Language Processing, Low-resourced languages, Monolingual embeddings, Cross-lingual embeddings, Language technologies.

Supervisors : Prof. V Marivate
Dr. A Modupe
Department : Department of Computer Science
Degree : Master of Science in Computer Sciences

[1] https://github.com/dsfsi/thapelo-sindane-msc-public.git

[2] https://github.com/dsfsi/za-bilingual-lexicons

" Let us not engage in this romanticist view that there can only be one way to learn. We first must indigenize knowledge, and learn from our cultures and traditions. Secondly, we need to diversify whom we learn from. We cannot only learn from the West, we need to learn from everybody who has something to teach."

Willy Mutunga, Former Chief Justice of Kenya.

Dedication

This work is dedicated to my late father Samson Tsei Sathura. We have spent very little time together, and spoke very few words to each other. However, in the limited encounters I had with you, you would wear a big smile on your face. You spoke very less, and acted more. If there was one thing I could change in this life, it would be the day I received a call informing me of your passing. In all that remains, thank you for all the support you showed in my education and life.

To my mother, Raisibe Matron Sathura - I dedicate this work to you. You have been a pillar of strength in my life. I could not have asked for a better mom. You showered not only me, but also my sibblings with love, support, protection, and blessings. Thank you for everything you have done for us Mma Sathura.

To my beautiful daughter, Hlalefo Mathula. Your smile is all the fuel I need to keep breaking boundaries, unlocking opportunities and making a difference. Keep smiling little girl. You are perfect just the way you are. Please remember that daughter.

And finally, to my two siblings, Kamogelo Sindane and Mmashela Sharon Sindane. I hope this work provides motivation and testimony that you can achieve all your heart's desires with consistency and hard work. Life taught me that I am not really smart as I thought I am. I just fight hard enough to achieve what I desire. I dedicate this work to you both.

Acknowledgments

This section serves to acknowledge all those who were instrumental to the completion of this work and my continued survival. Without you all, I would not have come this far. I thank you all and please receive my sincerest gratitude.

- To this very day, I have no idea what I have done for my stars to align and afford me the opportunity and privilege to meet him – Professor Vukosi Marivate, and for that I'm forever grateful. I wish him a healthy, fulfilling, and long life, he is a true deity in human form. What a true honor it has been to have crossed your path. Thank you Vukosi. Thank you for being around, for the support, and for your unwavering generosity.

- To my co-supervisor Dr Abiodun Modupe, thank you for bearing with me on this tough journey. You have been instrumental in the completion of this dissertation and without you, this would have not been possible. I thank you, and again congratulations on your PhD, well deserved.

- I would like to thank my sponsor MasterCard Scholarship Foundation, to whom without I would have not been able to sustain this journey. Thank you Dr Grace Ramafi, Dr Efe Esike, Sifiso Khuboni, Eloise Law-van Wyk, Lennox Wasara, Nandi Theka, and the supporting MasterCard team. I thank you.

- And finally. To the lab members of the Data Science for Social Impact Research group. You have made the University of Pretoria a home away from home for me. You gave me a sense of belonging. You gave me a voice, and for that, I will forever be grateful. I thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The cross-pollination of learnable language properties between languages with varying resources opens an opportunity to accelerate innovation and research in low-resource settings. Cross-lingual models are predominantly used methods for facilitating this sharing of learnable properties and have showcased efficacy in many downstream tasks such as Part-of-Speech Tagging (POS), Named Entity Recognition (NER), Grammatical Error Correction (GER), Sentiment analyses (SA), Machine Translation (MT), Natural Language Inference (NLI), etc. In applying these methods, interesting questions emerge such as the severity of low-resourcedness, the mathematical soundness of the methods, and what constitutes shareable language properties. Regardless, pertinent to our study is investigating the practicality of cross-lingual models in relation to all official South African languages.

Recent breakthroughs in the NLP arena of Artificial Intelligence (AI) have made a significant impact on society. With the emergent language technologies, businesses are able to seamlessly engage with their customers, meet their customer's needs effectively and efficiently, and realize unprecedented revenues [10]; institutions of teaching and learning can tailor educational content, and provide virtual environments that are specific to each individual student's learning styles with the potential to address various pedagogies [11]; in the healthcare sector, DNA and protein sequences can be unfolded with precision to gather insights that may have great potential in fields such as phar-

macogenomics and pharmacogenetics [12]. Acting as stewardships for the success of these NLP technologies is the availability of experimental resources (e.g. abundance of quality training examples, standardized dataset, computing resources, etc.), dedicated communities of researchers building critical mass, and availability of funding. Continuous refining of these critical resources complemented progress in research-centric tasks such as NER [13], POS [14], SA [15], MT [16], and recently the industry-centric tasks: Question Answering [15], Conversational AI [17], and Generative AI [18]. Regardless, the underlying language technologies that supported or in some cases motivated the vertical progression of the aforementioned socioeconomic constituents are only available in a few languages such as English, French, and Spanish often referred to as high-resourced languages. This uneven distribution of benefits raises alarming inclusivity concerns for horizontal counterparts (i.e. other globally spoken languages). Leading interventions aiming at addressing inclusivity vary across efforts that tackle increasing datasets and standardizing benchmarks [19], to those that increase research and development spend (R & D spend) [20], and those that develop techniques, tools, and processes that leverage existing resources (e.g. transfer learning, cross-lingual models, multilingual training, multimodality technologies, etc.) [2, 21, 22, 23]. In this work, we focus on the latter i.e. methods that supplement under-represented language resources with existing techniques of high-resourced languages, with a particular focus on cross-lingual models and embeddings [2].

Critical to cross-lingual embeddings is the development of mathematical models that facilitate the sharing, merging with, or projection of sufficiently learned language representation (typically a monolingual embedding space achieved through high-resourced settings [24]) from high-resourced language (s) to an insufficiently learned embeddings space (s) of low-resourced languages i.e cross-lingual models [2]. The objective of these models is to develop a strengthened shared vector space that preserves both intra-semantic relations within a low-resourced language (s) and inter-translation equivalences between low and high-resourced languages. The shared language space enables the training of NLP models (e.g. statistical, machine learning, or deep learning models) with the source data of high-resourced languages, or very few low-resourced language examples, or in

some cases no data at all (commonly referred to as zero-shot transfer learning), whose usability extends to the tasks of the low-resourced languages used to create the shared representations. This training setup has shown evidence of practical eminence in equivalent downstream tasks of low-resourced languages such as Dependency parsing [25], NER [26], POS [27], GEC [28], Entity linking and discovery [29], topic classification [30], and NLI also known as Textual Entailment [31]. From the ground up, these models have evolved into more sophisticated cross-lingual models such as multilingual transformer models (mBERT, XML-R, GPT) with significant performance gains on low-resourced downstream tasks [18, 21, 32]. Recently, Afri-centric cross-lingual models such as AfriBert, AfroLM, Afro-XMLR, and LaBSE with the aim of specializing in available African languages have been proposed [33, 34, 35]. Henceforth, available high-resourced technologies and resources such as quality datasets can be leveraged extensively to support low-resource settings. However, coverage remains a leading critical concern for many African languages with very low digital data or internet prevalence, and pertaining to the overarching issue, only 2-3 of the 11 official South African languages are included in active and impactful NLP research [13, 14, 36]. That is, the NLP research footprint of the majority of South African languages lags behind.

Downstream works using the empirically validated cross-lingual models for South African languages lag behind. Makgatho et al. [37], used an unsupervised VecMap projection model to create cross-lingual embeddings between Sepedi and Setswana and validated their performance on Word Similarity tasks. However, their work did not investigate extrinsic downstream tasks and only covered 2 South African languages. Ngomane et al. [38] created cross-lingual embeddings between English and Isizulu for news headlines classification. Notably, their work only considered 1 South African low-resourced language. Myoya et al. [39] compared the recently adopted multilingual models (AfriBerta, AfroXmlr, AfroLM) on news classification. However, only one (IsiZulu) South African language is considered in their work. Ifeoluwa Adelani et al. [13], Muhammad et al. [36] and the recent Dione et al. [14] works considered only three South African languages, namely, IsiXhosa, IsiZulu, and Setswana on NER, Sentiment Analyses, and POS downstream tasks respectively, using pre-trained multilingual models. Addition-

ally, Lastrucci et al. [40] created a multilingual corpus covering all South African languages and benchmarked their dataset through MT task using a multilingual M2M100 [41] translation model. However, their dataset is based on government content and therefore their application scope may be limited to their source domain. Besides, this remains an open question. Clearly, more work needs to be done for South African low-resourced languages, and surprisingly, no works explored alternative and rich resources such as bilingual lexicons with much higher language coverage and empirical validation and support from the cross-lingual community for creating cross-lingual embeddings [42]. Bilingual lexicons enable supervised training of cross-lingual embeddings. Considerable ablation works show that, by increasing bilingual pairs, the created shared representations become continuously refined [43]. That is, the more examples of similar words are available, the more precise and accurate the created cross-lingual embeddings will be generated by the cross-lingual models. Additionally, with bilingual lexicons, scaling to other unconsidered South African languages outside the typical 2-3 prevalent in research becomes a possibility.

1.1 Motivation

Cross-lingual models provide an alternative and a cost-effective approach to remediation strategies aiming at addressing technological inclusion such as data collection. Additionally, with limited domain expertise, language dataset creation becomes exponentially complex and expensive. As such, methods with cheaper views such as cross-lingual models become more attractive. Indeed, these models facilitated groundbreaking results on many complex tasks and continue to break inclusion barriers. Tasks such as cross-lingual document classification [44, 45], cross-lingual summarization [46], cross-lingual textual entailment [31], and cross-lingual question answering [47] provide empirical testimony for the efficacy of cross-lingual models. However, collectively, since the inception of cross-lingual models [2] (about 2 decades ago), not more than half ($\frac{1}{2}$) of the South African languages are found in active NLP research. This means fundamentals, benchmarks, baselines, or insights do not exist for all South African languages. Which are useful NLP commodities necessary to build state-of-the-art models and cutting-edge

technologies for South African languages. Moreover, bilingual lexicons played an integral part in cross-lingual models and have been shown to have better language coverage compared to other limited resources [42]. For these aforementioned reasons, we aim to investigate, benchmark, and develop cross-lingual-aided machine learning models for all South African languages using available datasets and bilingual lexicons. For our investigations, we considered traditional machine learning models such as Conditional Random Field, Long Short Term Memory, and Gated Recurrent Units. To support the horizontal development of these models we created monolingual datasets and bilingual lexicons for all South African languages.

1.2 Aim of the study

The study aimed at comparatively investigating cross-lingual modeling techniques to improve the quality of data processing for South African low-resourced languages and identifying key factors necessary for generating quality cross-lingual embeddings. Additionally, we aimed to investigate the impact of cross-lingual transfer learning on downstream tasks such as Named Entity Recognition, Part of Speech Tagging, News Headlines Classification, Word Similarity, and Machine Translation using machine learning models such as XGBoost, Long Short Term Memory (LSTM), Gated Recurrent Unit, Bidirectional LSTM, and e.t.c.

1.3 Research Objectives

The objective of this dissertation is to motivate the creation of benchmark datasets, models, and the use of innovative and cost-effective solutions in the development of cutting-edge technologies for low-resourced languages such as cross-lingual models. These objectives are succinctly organized as follows:

1. Conduct a comprehensive survey of available techniques in the field of cross-lingual models and their applications in natural language processing tasks for low-resourced languages.

2. Present empirical evidence of the efficacy of cross-lingual models for low-resourced languages targeting four languages: Sepedi, Setswana, Sesotho, and IsiXhosa.

3. Present practical case studies of using bilingual lexicons in scaling technology inclusion for low-resourced languages.

4. Explore a calibration study creating and analyzing cross-lingual embeddings across all South African languages: resulting in 110 cross-lingual embeddings.

1.4 Research Questions

This study addresses the following research questions to comply with the research study aim.

1. What is the current literature on cross-lingual models for low-resourced languages, particularly looking at South African low-resourced languages?

    - Given the current literature and advancements of cross-lingual models, techniques, and architectures, what are the recent trends, state-of-the-art, challenges, if any, reported in literature for the development of cross-lingual techniques for South African languages?

2. What cross-lingual modeling techniques are most effective for South African low-resourced languages?

    - Supplementary resources such as bilingual lexicons improve the quality of projection models. In this case, we aim to explore conditions that result in better cross-lingual embeddings and which projection technique works best for South African languages. Targeting modeling techniques such as Canonical Correlation Analyses [43], VecMap [48], and Muse [49].

3. What is the impact of cross-lingual representations on downstream tasks compared to monolingual representation?

- For tasks such as NER, POS, News Headlines Classification, and Machine Translation, what performance variations emerge when training traditional models (e.g Conditional random field, Logistic regression, etc.) and Deep neural networks (e.g FeedForward Neural network, Long Short Term Memory (LSTM), Gated Recurrent Unit, Bidirectional LSTM, e.t.c) and why?

1.5 Contributions

The unique contributions of this study are organized into the following seven paradigms:

1. We present a comprehensive survey of cross-lingual models, multilingual models, and their prevalence in the application arena of low-resourced languages.
2. We present a thorough evaluation of cross-lingual embeddings between English and four languages, namely, Sepedi, Setswana, Sesotho, and IsiXhosa on NER, and NHC.
3. Collection of the first bilingual lexicon dataset covering all South African languages.
4. Creation of the first benchmark of cross-lingual embeddings for all 110 combinations between South African languages.
5. Creation of hierarchy of transfer performance based on a cross-lingual target language calibration study.

1.6 Limitation

Recent literature evaluates cross-lingual embeddings efficacy as a consequence of their implementation and not their innate form. That is, the challenge of understanding which properties are transferred between embedddings to supplement one another has not been addressed. Which makes it even harder to truly understand cross-lingual embeddings. Cross-lingual models depend on the quality of monolingual embeddings as there is no additional training on the embeddings themselves but rather on the model

that projects them into the same shared space. Additionally, the lack of diverse annotated datasets for various downstream tasks such as Word sense disambiguation, Natural Language Inference, and Natural Language Understanding limits the scope of investigation of these cross-lingual models across varying downstream task types. Moreover, cross-lingual models lag behind in the transparency of shareable linguistic properties that are transferred to this so-called cross-lingual embeddings space so that it is easier to associate the performance of specific tasks with specific linguistic traits and patterns. However, other exploratory means to evaluate cross-lingual models in creating and using shared representations for model training make this domain worth investigating.

1.7 Dissertation Outline

This dissertation comprises 5 chapters, all addressing low-resourced languages in the realm of NLP. It is important to note that each chapter is self-contained in that it is a full paper on its own. The experiments are grounded in South African low-resourced languages. Each chapter aims to gather experimental and linguistic insights that can be useful in improving, scaling, and benchmarking technologies for low-resourced languages. The 5 chapters are as follows:

1. **Chapter 2** focusses on a comprehensive survey of cross-lingual models and their application in low-resourced settings.
2. **Chapter 3** covers the comparative analyses of two cross-lingual models: VecMap and CCA on two downstream tasks, namely, NER and News Headlines Classification on four South African languages: Sepedi, Setswana, Sesotho, and IsiXhosa.
3. **Chapter 4** presents a calibration study that explores the creation of cross-lingual embeddings from all pairs of South African languages and the implication of language pairs on downstream tasks and transfer performance.
4. **Chapter 5** gives concluding remarks and future direction of low-resourced languages and NLP tools.

1.8 Publications

This section outlines the works carried out in the completion of this thesis as well as others we collaborated with other researchers on, in relation to this thesis.

1. Publications

   - Dione, C.M.B., Adelani, D., Nabende, P., Alabi, J., Sindane, T., Buzaaba, H., Muhammad, S.H., Emezue, C.C., Ogayo, P., Aremu, A. and Gitau, C., 2023. MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African Languages. arXiv preprint arXiv:2305.13989.

2. Preprints

   - Sindane T., V. Marivate, and A. Modupe (2023), Zero-Shot Transfer Learning using Affix and Correlated Cross-Lingual Embeddings, Q.J.R. Meteorol. Soc., 2017;00:1–6.

3. Submissions

   - ACM Computing Surveys – Cross-Lingual Embedding Methods and Applications: A Systematic Review for Low-Resourced Scenarios.
   - Special Interest Group on Under-resourced Languages (SIGUL) – Point of Pivot: Calibration of Cross-lingual Embeddings for Southern Nguni and Niger-Congo Low-Resourced Languages.

Chapter 2

Cross-Lingual Embedding Methods and Applications: A Systematic Review for Low-Resourced Scenarios

This chapter peruses available Natural Language Processing (NLP) literature on cross-lingual models and provides a comprehensive introductory message looking at the history of language technologies, their current state, current trends, current application, challenges of inclusivity for low-resourced languages (e.g annotated datasets, language resources, lack of expertise on indigenous knowledge, and tools), and recently proposed solutions for language technology inclusion in this space. Furthermore, we succinctly organized the available works on cross-lingual models into a taxonomy, highlighting gaps, considered downstream tasks, and pointed out future prospects in the field of cross-lingual models for low-resourced languages.

2.1 Introduction

The ideology of equipping machines with the intelligence to process, learn, bind patterns, and manipulate written language became a shared responsibility over the past decades in the field of Natural language Processing (NLP)[50, 51, 52, 53]. With this vision emerged a complex task of establishing a mirror image of language in the compute space, which is commonly referred to as language models. Older approaches limited by available resources and linguistic knowledge took a natural standpoint on modeling language as independent linguistic units (words) without considering nuanced complexities of a language such as morphology, word order, temporal nature, context, and interrelat-

edness of words, etc [53]. Methods that were in active research participation included the rule-based or case-based methods [54], Bag of Words [55], N-gram models [56, 57], Term Frequency (TF) models [58], and Term Frequency Inverse Document Frequency (TF-IDF) [56, 58, 59]. Rule-based methods operate in a similar fashion as case-based reasoning, in that, the objective is to exploit an already existing knowledge base and transfer capabilities to new unknown domains [54, 60]. For example, a list of hand-coded rules can be pre-defined and used as a system's reasoning engine to determine new features or new representations in the study domain. Bag of Words method first converts the document into a discriminative list of words and represents the document as a vector of occurrences of the discriminative words represented in binary [61, 62]. N-gram models consider n consecutive subsequences as a representative token to determine the probability of the next token and they approximate the Markov's property [56, 57, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74]. TF models count the occurrence of a token in a document and the weight of that token is represented by the number of times it occurs in the document [59]. That is, a document will be represented by a matrix of a number of occurrences, where each dimension represents a token. TF-IDF models on the other hand complement TF models due to TF models being susceptible to amplifying non-informative features such as stopwords (e.g. $\{a, on, the, at, e.t.c.\}$) [59]. That is, TF-IDF models multiply the occurrence matrix (the TF matrix) by regularizing factor δ (see eq 2.1) that is low (i.e damp it down) for plausible stopwords and high (i.e. intensify) for uncommon words. More concretely, the IDF part of the TF-IDF models regularizes the TF part. These traditional approaches were bounded by their insufficiency in capturing essential language constructs such as semantics, syntax, temporal dependency, and relatedness of words. Other compute-related weaknesses of such encoding schemes included the high dimensionality of representations which conflicted with the limited compute capabilities at the time. These impedes, and others fueled the unequivocal need to consider alternative language models. Motivated by the linguistic insight "You shall know a word by the company it keeps" (Firth, 1957), a new era of models with a different mechanical viewpoint of language processing emerged. These models learned continuous representation of words in vector space, commonly known as word embeddings [1]. Word embeddings are word-vector representations that are

obtained by taking the neighboring words into account during training. These neighboring considerations can be two-fold, 1) a vector representation of a word is obtained by learning to predict the neighboring word vectors within a given sentence, and 2) the neighboring vectors are used to predict the obfusticated word vector. A popular model that adopts the former process is the Skip-gram (SG) model, and a model that adopts the latter is the CBOW model [1].

$$\delta(t, D) = \log\left(\frac{N}{count(d \in D : t \in d)}\right) \tag{2.1}$$

where t is the token, D is the corpus of documents d, and N is the number of documents.

For a while, word embeddings were considered state-of-the-art objectively speaking in relation to language models. Indeed, these models championed many research breakthroughs in many NLP downstream tasks such as Sentiment Analyses, Question Answering, Authorship Attributions, Machine translation, etc [75]. Embeddings of varying linguistic units such as characters, phrases, sentences, paragraphs, and documents also gained attention for different tasks and applications [26, 76, 77]. This varying level of linguistic granularity arose due to varying complexities of modeling downstream tasks. Regardless, the underlying support that strengthened progress in these breakthroughs included the availability of digital datasets, computing power, an active research community, and sophisticated language algorithms such as the continuous bag of words(CBOW) and Skip-Gram model that effectively processed millions of tokens in limited time constraints [1]. However, continuing efforts (e.g. developing standardized datasets and the advent of sophisticated algorithms) and the resulting proliferation of language models in the NLP technology space highlighted disparities in the benefits of available language technologies in terms of access to these technologies, representation of languages and datasets, and control or ownership of these datasets (as their use can often result in positive or negative economic and social implications). As a result of emerging inclusivity concerns - cross-lingual models were coined in 2013 as an intervention to address these disparities [2].

The emergence of cross-lingual models was based on the geometric constellation of two

monolingual embeddings plotted abreast [2]. That is, words in the embedding space of one language occupied the same geometric positions as their translation-equivalent words in the embedding space of another language. This observation sparked the question "if there is a possibility to create projection techniques that can map the embedding space of one language to another and vice-versa?". Albeit, the birth of cross-lingual models. With this capability lay the possibility to utilize already learned language properties in these spaces to supplement capabilities in low-resourced settings. From this question, emerged a series of projection techniques which we will cover in detail in the next subsections 2.3.

The projection techniques were grouped into the following categories: Linear (Offline) projection models, Pseudo cross-lingual (Pseudo-CL) models, cross-lingual (objective-CL) models, joint cross-lingual (joint objective-CL) models, and finally, large pre-trained multilingual (LPML) models. Offline projection models map two or more monolingual embedding spaces into a single joint embedding space with the optional use of a bilingual lexicon as a language equivalence signal. Pseudo-CL models create a pseudo corpus tied together by language equivalences or pure heuristic-based configurations of words or sentences. These models continue to train a monolingual embedding model such as skip-gram [1] using the created pseudo-corpus to create the shared embedding space. The objective-CL models use parallel data as a language-equivalence signal and are designed to optimize an objective function.

Joint-objective CL models complement objective-CL models by ensuring that not only inter-language semantics knowledge is preserved in the extracted space but also intra-lexico-semantic knowledge exists in the induced cross-lingual space. LPML models on the other hand implicitly extract shared embedding spaces through joint language training and shared sub-word vocabulary. However, this school of thought has been questioned extensively [78, 79]. The advent of such techniques yields an imperative need to systematically review them by looking at their application, domain, model soundness from a mathematical standpoint, metrics, languages in active participation, available datasets, etc. This chapter addresses these concerns and creates a taxonomy that highlights important insights related to the current state of cross-lingual models. Hitherto discussing each category of cross-lingual models, the next section describes monolingual embeddings and how they revolutionized language processing.

2.2 Monolingual embeddings

Training machine learning algorithms can be an exigent task, especially in highly complex domains such as NLP. The lack of enough and properly annotated training data exacerbates the challenge for many low-resourced languages across the world. However, for scenarios where enough and properly supervised data is available, a more pressing issue becomes that of encoding a representation of language into a space such that intra-language idiosyncrasies (e.g., semantics, syntactic, and word order) are preserved for machines to learn in downstream tasks. Traditional methods used to encode language representations included statistical-based methods such as n-grams, TF-IDF or bag of words, where each word in a vocabulary is represented as a single and isolated unit with no inherent relationship with other words. However, these methods failed to capture syntactic and semantic regularities or subtleties of language such as the relationship between words and word order, which are essential for applications such as Information retrieval, recommendation systems, machine translation, etc. Many works have been proposed to address these limitations [76, 80, 81, 82, 83] including the popular works of Bengio et al. [84]. In these works, the idea was to exploit the established expressive power of neural-based language models to learn continuous representations of words in latent vector space with the objective of ensuring that similar words have similar vector representations. These continuous vector representations of words are popularly known as word embeddings. These works follow from Rumelhart et al. [81], where the notion of distributed representation of words was initially proposed.

In a time where computational freedom was more restrictive, Mikolov et al. [1] introduced two less computationally greedy architectures for learning continuous representation of words, namely, CBOW and SG model which (amongst others) popularised embeddings and were able to train on billions of word tokens. The objective of CBOW is to learn the center word representation based on the context words and SG predicts the context words given a center word.

For example, Fig. 2.1 represents the CBOW architecture for predicting the current word based on the context (on the left), while the SG predicts surrounding terms based on the current word (on the right). Mikolov et al. [85] demonstrated that such repre-

sentations can capture syntactic and semantic regularities by devising an offset-based approach that uses basic arithmetic reasoning to model vector relationships and evaluated the feasibility of the technique on analogy datasets[1]. For example, their offset-based reasoning demonstrated that such spaces can capture the concept of royalty (semantics), where the vector ("King") - vector ("Man") + vector ("Woman") resulted in a vector representation close to that of "queen" obtained through cosine distance without being explicitly taught such relations. Additionally, they explored syntactic information of the embedding space through Simlex tasks, where the task was to query this embedding space for similar words given a predefined list of words. Consequently, such observations resulted in a great increase in research attention for learning continuous word representations [86, 87, 88].

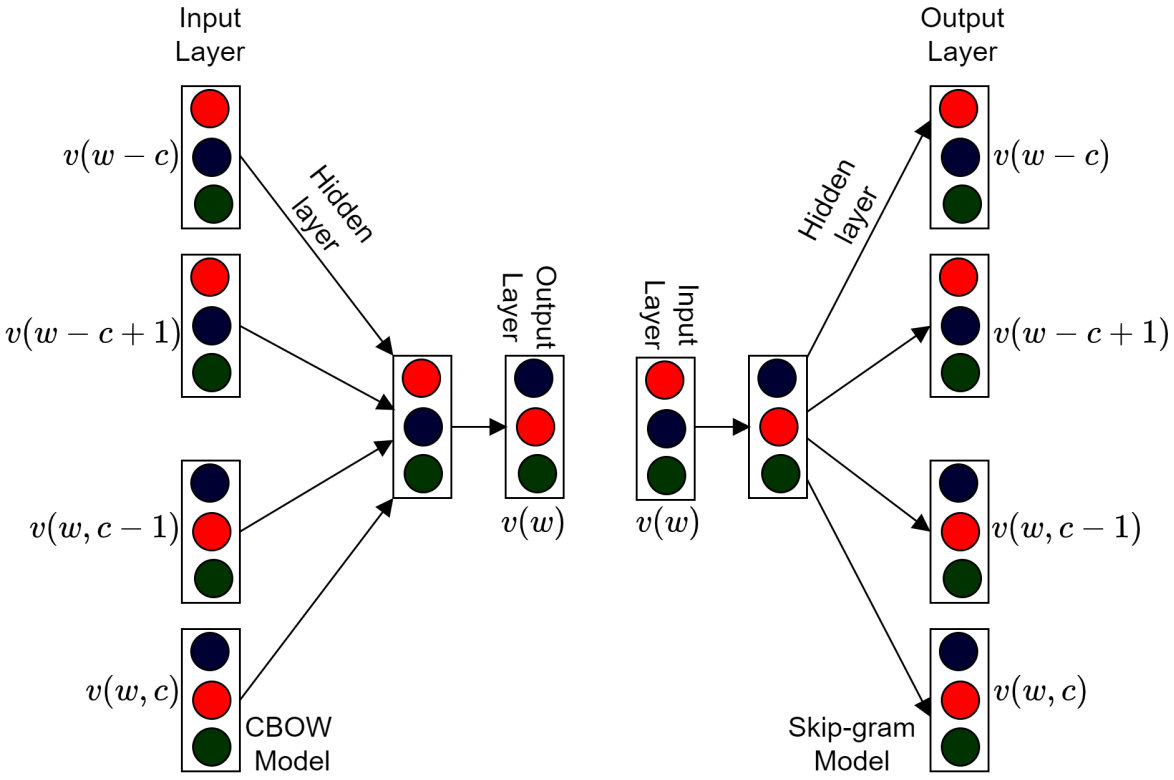


Figure 2.1: The CBOW and SG architectures Mikolov et al. [1]

[1] http://research.microsoft.com/en-us/projects/rnn/default.aspx

The major convenience with embeddings (often not disclosed) is that they allow prediction models to learn a subspace of similar words using a single example in that subspace. This is possible because such embedding spaces are semantically and syntactically tied through correlated and collocated vector representations. Hence, during parameter updates for a single word example, the adjustments are carried over to related words in the space. This is what makes embedding spaces an indispensable and extremely powerful technology of its time.

Statistical-based methods such as latent semantic analyses (LSA) and latent Dirichlet allocations (LDA) have also been attempted to model continuous representations of words and have been shown to be less effective compared to neural-based techniques [89]. The spoken language is a very complex artifact, so creating a space that encodes language dynamics in vector space is non-trivial. The major requirement for building a functional embedding space is having large amounts of representative data for that language. The immediate drawback with this is that most underrepresented languages have very little to no digital data available. Furthermore, the cost of creating datasets for these languages is very high. For these reasons and more, NLP research on low-resourced languages resorted to devising methods that can share learned knowledge ingrained in such a space (generally from a high-resourced language) to another space (a low-resourced language) with the hope of establishing or exploiting inter-language relations to improve performance in limited settings. This is what cross-lingual models aim to achieve. The next section discusses models that construct cross-lingual spaces grouped into different categories and their application for low-resourced languages.

2.3 Cross-lingual Projection Models

There are generally five types of projection techniques for mapping embeddings from one space to another to construct cross-lingual embeddings. These are offline projection models, pseudo-cross-lingual models, cross-lingual training models, joint-objective cross-lingual training models, and large pre-trained multilingual language models. Offline projection models map two or more monolingual embedding spaces into a single shared space, with the optional use of a bilingual lexicon as a language equivalence

signal. Pseudo-cross-lingual models create a pseudo corpus tied together by language equivalences or pure heuristic-based configurations of words or sentences. These models continue to train a monolingual embedding model such as skip-gram [1] using the created pseudo-corpus to create the cross-lingual space. The cross-lingual training models use parallel data as a language-equivalence signal and are designed to optimize an objective function. Joint-objective cross-lingual models work with cross-lingual training models to make sure that both inter-language semantic knowledge and intra-lexico-semantic knowledge are kept in the cross-lingual space that is created. Multilingual language models, on the other hand, implicitly extract cross-lingual spaces through joint language training and shared sub-word vocabulary. The following subsections provide a detailed discussion of the aforementioned methods.

2.3.1 Linear Projections or offline alignment of embeddings

The possibility of constructing joint representations was inaugurated through the observation that embedding spaces of different languages can have similar geometric properties [2]. In Fig. 2.2, geometric patterns of word representations in one language, "English", show similar qualities to those in another language, "Spanish" in the vector space. Mikolov et al. [2] created a basic linear projection model based on this knowledge, as illustrated in Fig. 2.2, which represents the number and animals in English (on the left) and Spanish (on the right). This model utilises monolingual embeddings from two languages to transfer the embeddings of one language to another. They trained their projection model, which is shown by the translation matrix M, using stochastic gradient descent (SGD) to get the least-squared error as low as possible, as shown in Eq. 2.2. To understand the translation matrix M, the author initially compiled a bilingual dictionary by translating the 5000 most common words from the source language (English). This dictionary was then employed to train model M for transferring vector representations from the source language to the target language embedding space.

$$\min_{M} \sum_{i=1}^{n} ||Ma_i - b_i||^2 \tag{2.2}$$

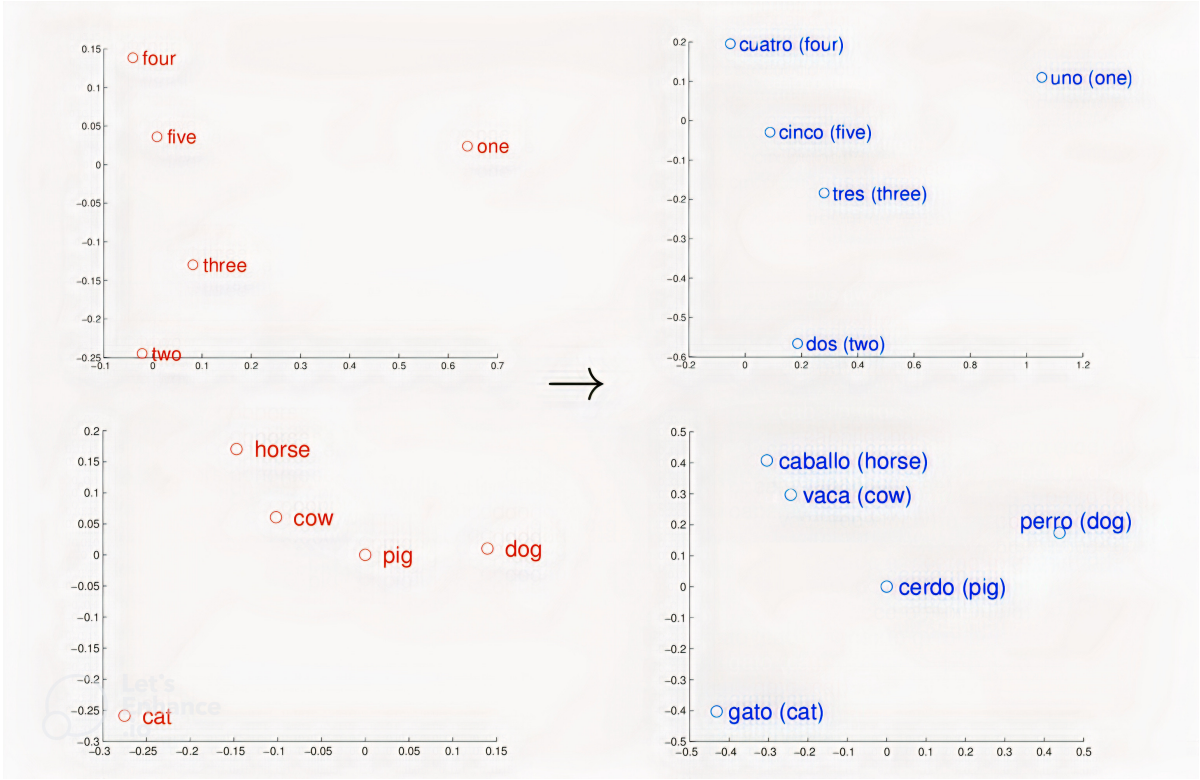


Figure 2.2: Distributed word vector representations Mikolov et al. [2].

where a_i and b_i are vector representations of aligned words from the initial dictionary of source and target translations.

Following the work in [2], many studies have been proposed to learn shared representations of words. Faruqui and Dyer [43] proposed a canonical correlation analysis (CCA) approach to learn a linear mapping of embedding spaces. In their approach, they find correlated projection vectors V and W from sampled embedding spaces of the two languages ($A^{'}$ and $B^{'}$) by training to maximise ρ (Eq. 2.4). To achieve this, they first let a and b be corresponding words in embedding spaces $A^{'}$ and $B^{'}$ respectively and v and w be projection vectors of a and b respectively. They then compute the projected vectors a', b using the projection directions v and w as follows:

$$\begin{aligned} a' &= av, \\ b' &= bw \end{aligned} \tag{2.3}$$

From these projected vectors, they maximise ρ to get v and w using the equation 2.5 below:

$$\rho(a^{'}, b^{'}) = \frac{E[a'b']}{\sqrt{E[a'^2]E[b'^2]}} \tag{2.4}$$

$$v, w = \text{CCA}(a, b) = \underset{v,w}{\text{argmax}}\, \rho(av, bw) \tag{2.5}$$

To generalize this approach to the entire sampled embedding space they used the equation below:

$$V, W = CCA(A', B') \tag{2.6}$$

To find the final cross-lingual embedding space using A and B (the original monolingual embedding spaces), they utilized the learned correlated vectors V an W to compute the projected embedding spaces $A^* = AV$ and $B^* = BW$. A^* and B^* can be observed as the same embedding space.

The linear model proposed by Mikolov et al. [2] was later discovered to have inconsistencies among the objective functions used to create the embeddings, the error function for learning M, and the similarity measure. Xing et al. [90] proposed to normalize the embedding vectors to be the unit length and learn an orthogonal translation matrix M' instead.

Lazaridou et al. [91] spotted another issue associated with the objective function used to learn the projection matrix M of Mikolov Mikolov et al. [2]. They observed that the linear projection model learned through the least-squares error function maps some vectors to become universal neighbors of other vectors .i.e., a concept referred to as hubness. They addressed this by implementing a max-margin ranking loss proposed by Collobert and Weston [24]. The max-margin loss did not perform well on a cross-modal task experimented in Lazaridou et al. [91] for which they resolved through data augmentation.

Ammar et al. [92] presented two methods (multiCluster, multiCCA) for estimating shared embedding space between 59 languages and evaluated their performance on both intrinsic and extrinsic tasks. For multiCluster, they first create a multilingual cluster using bilingual dictionaries and replaced each token in each monolingual corpus with a cluster ID if it belongs to a particular cluster in the multilingual cluster. They merged the monolingual corpora (with replaced tokens) into a shared multilingual corpus and trained a simple skip-gram model to create a cross-lingual embedding space. For multiCCA, they extended the CCA [43] to a multilingual setting using English as the source language for each of the available languages. In addition to their contribution, they proposed an evaluation metric (multiQVEC-CCA), which is an extension of QVEC proposed in Tsvetkov et al. [93]. Furthermore, their study covered a wide range of intrinsic evaluation tasks (translation-variance, word similarity, and word translation) to assess semantics and syntactic information of their constructed joined representations and extrinsic evaluation tasks (document classification, cross-lingual parsing) to assess their usability in practice.
It is important to also mention the work of Valerio Miceli Barone [94], which introduced an interesting avenue of research for finding cross-lingual representations using only monolingual embeddings from source and target language by exploring adversarial neural networks. Adopting adversarial networks seemed like a good way to do research because of the bad situation of having little data in some languages[95, 96].

A limiting pattern in the research line of adversarial-based methods pointed out that these methods are being over-fitted to favorable evaluation conditions such as comparable corpora or using closely related languages which resulted in them performing poorly in realistic conditions. This gave birth to a new era of fully unsupervised approaches doing away completely with adversarial neural networks [48, 97, 98, 99, 100]. Artetxe et al. [48] adopted a fully unsupervised approach that executes in four steps (explained below).

1. The first step involves normalizing the embedding spaces of each language to be unit length and then mean centers for each dimension. The mean centering is followed by the second length normalization to ensure that the resulting embeddings are unit length and their dot product is equivalent to cosine similarity and is directly related to their Euclidean distance.

2. The second step is based on the isometry assumption that exploits the distribution similarity between monolingual embeddings A and B to construct semi-aligned embedding matrices $A^{'}$ and $B^{'}$. To construct these embeddings, they first create two similarity matrices W_A and W_B using singular value decomposition of A and B respectively. They then sort these similarity matrices in the j-th dimension (column-wise) and normalize them using step one explained above. The result of this process provides matrices $A^{'}$ and $B^{'}$ such that given a word w and its embedding in $A^{'}$ an equivalent word k can be extracted in $B^{'}$ using the nearest neighbor. This initial solution of weakly aligned embeddings is used to create an initial dictionary D. Then the initial solution is iteratively refined using a self-learning technique (using the original embeddings A and B instead).

3. The third step optimizes the objective function:

$$\underset{W_A, W_B}{\operatorname{argmax}} \sum_i \sum_j D_{ij}((A_{i*}W_A) * (B_{j*}W_B))$$

with the help of four dictionary induction strategies that assist in escaping a local optimum: Stochastic dictionary induction, frequency-based vocabulary cut-off, CSLS retrieval, and bidirectional dictionary induction. A_{i*} and B_{j*} are word embeddings in A and B respectively. Once the initial dictionary is computed, both $A^{'}$ and $B^{'}$ are discarded and iterations are executed on the original embedding A and B.

4. The last step applies symmetric re-weighting of the converged transformation matrices W_A, W_B.

Hoshen and Wolf [97] proposes a method similar to the method by Artetxe et al. [48]. However, in their case structural similarity is exploited to extract an initial solution using principal component analyses (PCA), motivated by the theory of 3D point cloud matching. That is, they use PCA to extract the top n principal components of the embeddings A, B and use those as initial embeddings of $A^{'}$ and $B^{'}$. To extract a translation model $M^{'}$ that can be used to align embeddings $A^{'}$ and $B^{'}$, they optimize an iterative closest point (ICP) objective with an added constraint of ensuring that

translation in both backward and forward direction results in the same representation using the Equation 2.7:

$$\sum_{j} ||b_j - M'_{ab} a_{f_b(j)}|| + \sum_{i} ||a_i - M'_{ba} b_{f_a(i)}|| + \lambda \sum_{i} ||a_i - M'_{ab} M_{ba} a_i|| + \lambda \sum_{j} ||b_j - M'_{ab} M'_{ba} j_j|| \tag{2.7}$$

where a_j,$b_{f(j)}$ is an embedding in $A^{'}$ and $B^{'}$, respectively, and $f(j)$ is the index of the embedding. $M^{'}$ is then used to initialize M to learn alignments on the original embeddings A and B using the same ICP objective.

Alvarez-Melis and Jaakkola [98] on the other hand formulates the task of mapping embedding into a joint space as an optimal transport (OT) optimization task using Gromov Wasserstein distance [101]. Simply learning a mapping function that minimizes a measure of discrepancies between the mapped source word and the target word is usually insufficient for inference. The underlying objective is mostly supplemented with an inverted softmax function (ISF) or cross-domain local similarity (CSLS) criteria to avoid the issue of hubness during inference. Joulin et al. [99] draws from this setup and asserts that the underlying objective functions are inadequate for both learning and inference of embeddings simultaneously and proposes a unified learning objective based on a convex relaxation of the CSLS criterion. Their formulation resulted in optimizing the following objective function using projected sub-gradient descent:

$$\min_{M \epsilon \vartheta_d} \frac{1}{n} \sum_{i=1}^{n} -2 a_i^T M^T b_i + \frac{1}{k} \sum_{b_j \epsilon N_B(M a_i)} a_j^T M^T b_j + \frac{1}{k} \sum_{M a_j \epsilon N_A(b_i)} a_j^T M^T b_i \tag{2.8}$$

where ϑ_d is a set of orthogonal matrices, M is the translation matrix, and $a_i \epsilon \mathbb{R}^d$, $b_i \epsilon \mathbb{R}^d$ are aligned word embeddings in source and target languages respectively. Finally, Ruder et al. [100] proposes a multi-step discriminative latent variable model formulated as solving a combinatorial optimization problem of bipartite graphs [102] to project monolingual embeddings into a joint vector space.

A noticeable drawback to these methods is that they rely heavily on the existence of monolingual embeddings. This is because the inception of these models was driven by

the objective of attempting to capture more linguistic information shared between languages to aid performance gains of models that are already in active research. However, this assumption of the existence of functional monolingual embeddings does not hold for languages with extreme scenarios. For this reason, the main focus of this paper is to investigate how these proposed methods were scaled and modified to truly low-resource settings with very little to no digital data available. Additionally, we intend to inspect how the induced spaces were evaluated by looking deeply at the source languages. Moreover, we analyze how these models were applied successfully in practice by looking at both intrinsic and extrinsic tasks and this is covered in section 2.5.

Glavas et al. [103] highlight difficulties in appraising projection-based cross-lingual embeddings, despite tremendous progress in the field. In their paper, they address the issue of a unified evaluation criteria for evaluating shared spaces and the need to avoid overfitting these embeddings to shallow intrinsic tasks such as bilingual lexicon induction but rather construct robust shared spaces with enough capacity to generalize to downstream tasks. However, recent approaches have overcome these limitations by providing new insights into projection models. In the following part, we describe pseudo-cross-lingual models that rely heavily on pseudo-data rather than matched words to learn a translation matrix M. These models have shown promising results in various studies and have the potential to revolutionise the field of cross-lingual modelling.

2.3.2 Pseudo-cross lingual projections driven embeddings

These models exploit a cross-lingual corpus created using small seed dictionaries as a strategy to tie two languages together. Xiao and Guo [3] relied on the expressive power of non-linear neural networks to create cross-lingual representations between languages. As shown in Fig. 2.3, the word w_i from the training set x is linked to an inter-language representation vector $\mathcal{R}(w_i)$ by the embedding matrix $\mathcal{R}$. In their approach, they find all source and target pairs using Wiktionary [104] to translate all source words into target words and filter the induced dictionary based on polysemous words and out-of-vocabulary (OOV) words. They then use this semi-clean dictionary to construct a unified bilingual vocabulary V, where V is made up of the merged source and target words. They further replace each v in V that appears as a pair of another word in the bilingual dictionary with

the same representation vector w. Their study provides empirical evidence supporting the efficacy of their induced bilingual embedding space on a dependency parsing task. Similar to the limitations of the majority of studies reviewed in this paper, their study did not evaluate their induced embedding space with an intrinsic task to inspect the linguistic content preserved in the embeddings. This is done so that the nature of the extrinsic tasks (e.g. POS tagging, Opinion mining, etc) can be chosen wisely to align with induced embedding space.

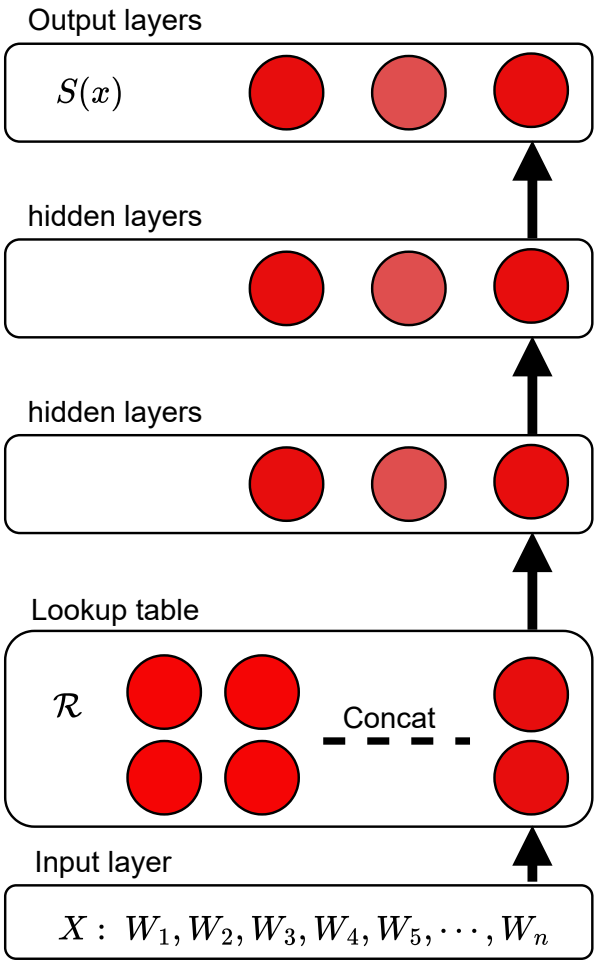


Figure 2.3: Cross-lingual word representations using the deep learning method [3].

Bilingual lexicons continue to become pivotal in connecting two languages in order

to extract cross-lingual embeddings. Many works exploited this resource obtained either from human annotations or utilizing translation tools such as Google translate [105], and Wiktionary [104]. Gouws and Søgaard [106] proposes an approach that relaxes the reliance of bilingual dictionaries on inducing shared spaces by using near-equivalence pairs of words in both languages as signals to connect languages. For example, they considered words such as a car in English to be equivalent to the word maison ('house') in French since they are both nouns and used this information as paired signals. Besides word translation (GT) pairs and POS equivalence, they investigated other forms of equivalences (e.g. clerk and chauffeur for super sense tagging) and their approach proved to be a novelty. Once this equivalence is constructed, they merged corpus C_1 from language 1 with corpus C_2 from language 2 and shuffled the words in the resulting merged corpus C_g. Then, for each word w in C_g, if that word appears in the equivalence dictionary it is replaced with its pair with a probability of $\frac{1}{2k}$ with k = cardinality. Lastly, they used a simple CBOW on C_g to extract bilingual embeddings. Similar to other works, they evaluated their embeddings on an extrinsic task (POS tagging).

A good cross-lingual space should preserve both monolingual and cross-lingual qualities. For example, similar words within a language should occupy the same position in the embedding space (monolingual quality) and likewise, similar words across languages should have similar representation (cross-lingual quality). Duong et al. [107] demonstrated how this property can be achieved by proposing an Expectation Maximisation(EM) driven approach for learning a shared embedding space using CBOW and adding the resulting context and embedding matrix. The inclusion of EM was solely to handle polysemy from a noisy dictionary. A distinguishing feature of their study is that they evaluated the induced shared embedding space across all task types i.e. both intrinsic and extrinsic tasks.

In Vulić and Moens [108], two methodologies are proposed for creating a pseudo-CL corpus with affordable document-aligned data. The first strategy creates a multilingual corpus by combining paired documents from each monolingual corpus and randomly rearranging the words in the final document. The second strategy appends each word from two paired documents to a third document based on their length and ratio (known as the length-ratio shuffle). Similar to Ammar et al. [92], the researchers trained the SG

model using negative sampling (SGNS) to retrieve the embedding from the joint corpus. This demonstrated tremendous gains over the state-of-the art (SOTA) method.

In pseudo-CL approaches, heuristics are primarily used to select words for inclusion in a joint document, which may result in less effective outcomes. Moreover, assigning identical representations to seemingly synonymous words in the pseudo-document can restrict the development of a cohesive shared embedding. For instance, in Sepedi (a South African Bantu language), the word 'Lena' in English, which translates to 'Them', is used as a term of respect to address an individual (with the actual meaning being 'you'). Therefore, neglecting this information when creating pseudo-documents can result in undesired outcomes. Thus, there is a need for more effective techniques to build joint embedding spaces. These methods have been shown to outperform traditional bilingual dictionary-based approaches in various tasks. These include traditional bilingual dictionary-based approaches to various tasks that leverage larger amounts of data and capture more nuanced relationships between languages, which have proven to be highly effective for cross-lingual tasks. The research demonstrated that these approaches surpass conventional bilingual dictionary methods for diverse tasks. Hence, the following section explores alternative methods for constructing cross-lingual embedding spaces, focusing on objective functions and parallel data instead of relying on bilingual dictionaries.

2.3.3 Cross-lingual training models

This family of cross-lingual models are tailored towards optimizing a CL objective function based on sentence-aligned corpora. The motivation is to gravitate from the requirement of word-aligned bilingual corpora as seen from previous methods to the extraction of higher-level distributed representation using objective functions.

2.3.3.1 *L2* regulariser and Noise contrastive estimation

The first approach to optimizing an objective function to learn shared representations of parallel sentences (later extended to multilingual) was proposed by Hermann and Blunsom [5]. In their approach, the idea was to minimize an $L2$ regularizer (see Equation

2.9) in order to drive representations of similar words closer together, while simultaneously enforcing a margin between a dissimilar word with a noise contrastive estimation criterion. This led to the resulting loss function $J(\theta)$ as seen in Equation 2.10.

$$E(a,b) = ||a_{root} - b_{root}|| \tag{2.9}$$

$$J(\theta_{bj}) = \sum_{(a,b)\epsilon C_{A,B}} (\sum_{i=1}^{k} E_{noise}(a,b,n_i)) + \frac{\lambda}{2}||\theta_{bi}||^2 \tag{2.10}$$

where a and b are pairs of parallel sentences, n are the negative samples such that each sentence in n is not semantically equivalent to a, λ is a regularisation term, and θ is the set of all model parameters.

Specifically, they trained two sentence-level compositional vector models (CVM) for each pair of languages. They proposed a CVM : ADD a model that adds a vector representation of each word to construct a sentence representation. For the resulting sentence representation of each CVM, they minimized the objective function $J(\theta)$ and propagated the signal back to word embeddings using SGD [4]. See Figure 2.4 of their model.

Hermann and Blunsom [109] extends their paper [5] to document-level representations by simply recursively applying the sentence-level CVM and objective function. Additionally, they proposed a new CVM: BI - a compositional model that applies hyperbolic tangent to the vector representations of each word in the sentence and sums the resulting vector (see equation 2.11).

$$f(x) = \sum_{i=1}^{n} \tanh(x_{i-1} + x_i) \tag{2.11}$$

2.3.3.2 Bag-of words Autoencoder

Lauly et al. [110] added to the knowledge base of optimizing robust objective functions to extract inter-language semantic representations. Their objective resembled that of a decoder reconstruction loss function. In their study, they trained an encoder with two decoders for each source and target language such that reconstruction of the original

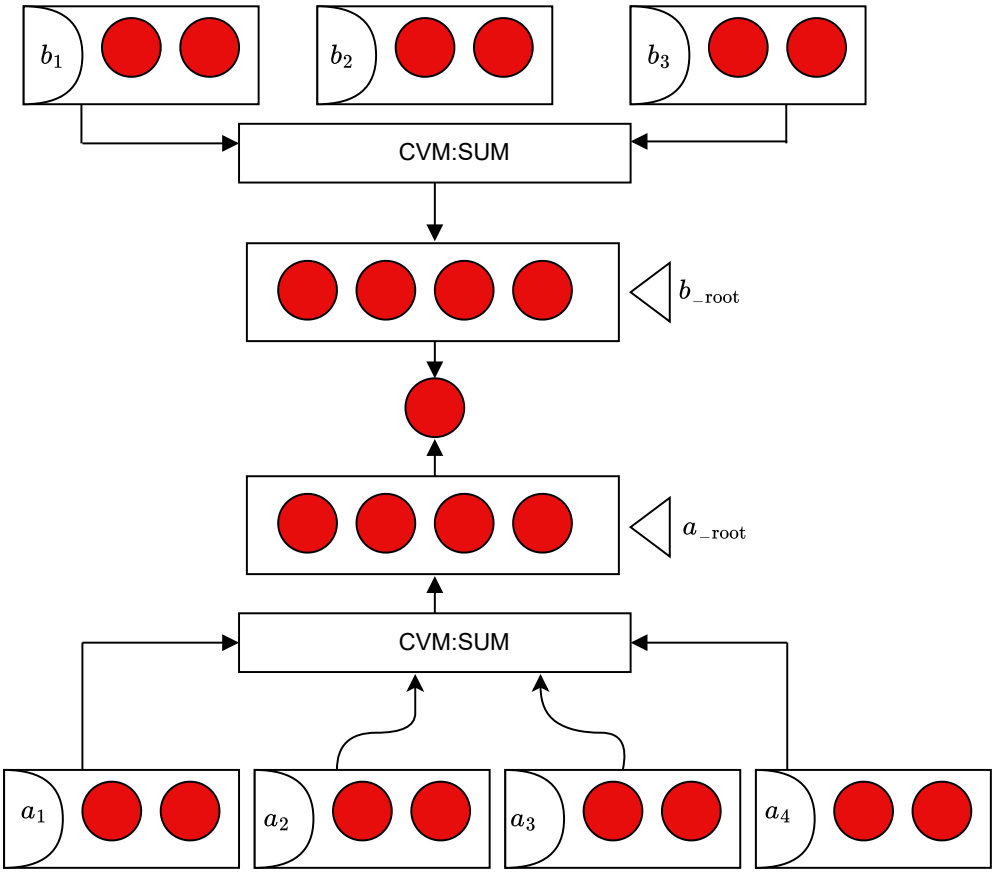


Figure 2.4: A bilingual model that minimizes the discrepancy between parallel input sentences a and b encoded using CVM [4] as extracted from [5].

sentence is possible in any of the given languages. The full architecture is a multilingual autoencoder capable of learning semantic representations across languages. The added advantage to this work is that they do not rely on word-aligned datasets, which are not available for low-resourced languages. This means that their approach is scalable to low-resourced settings.

2.3.3.3 Probabilistic distributed word alignments(PDWA)

While probability alignments have proven to be such an integral part of learning translation matrices, Kočiskỳ et al. [111] proposes an adaptive model that learns alignments, translation probabilities, transformation matrix, and cross-lingual representations all at the same time. To achieve this, they adopted a hybrid model consisting of two parts: 1. A modified version of FastAlign trained only for 5 epochs, 2. a probabilistic language model that minimizes the energy function as denoted in 2.12.

$$E(f, e_i) = -(\sum_{s=-k}^{k} r_{e_{i+s}}^T M_s) r_f - b_r^T r_f - b_f \tag{2.12}$$

where $r_{e_i} and r_f$ are vector representations for source and target words e, f in V respectively, M is the transformation matrix, b in R is the representation bias, b_f are biases for each word in V. The translation probability is given by equation 2.13:

$$p(f|e_i) = \frac{1}{Z_{e_i}} exp(-E(f, e_i)) \tag{2.13}$$

where the normalizer Z_{e_i} is defined in 2.14 below:

$$Z_{e_i} = \sum_{f} exp(-E(f|e_i)) \tag{2.14}$$

Unlike other methods proposed, their model takes two parallel sentences as input and for each word in the source sentence, they attempt to predict the associated word in the target sentence. This of course can be easily adapted to take context words of length k around the center word as demonstrated in their study. Their joint architecture was able to extract useful alignment and a shared embedding space. Empirical evidence supporting this was obtained on alignment tasks, cross-lingual document classification, and qualitative analyses. The added advantage of their model is that their derived representations are explainable i.e. the model provides better reasoning for the derived vector representations, which is often not the case for their contenders.

2.3.3.4 Multilingual Autoencoder

Chandar AP et al. [112] also explored autoencoders for extracting bilingual representation. In their approach, they investigated and compared the optimization of two objective

functions used to pull similar word representation closer together (cross-entropy loss-equation 2.15) and negative log-likelihood-equation 2.16)). Their study highlighted the following added advantages of autoencoder models: 1. word-aligned and exact sentence-level aligned resources are not essential, and 2. single-step extraction of multilingual embeddings is possible. Similarly, they argued the usefulness of their embeddings through cross-lingual document classification and compared their results with(at the time) state-of-the-art [44] cross-lingual model.

$$l(v(x)) = -\sum_{i=1}^{V} v(x)_i log(\hat{v}(x)_i) + (1 - v(x)_i) log(1 - \hat{v}(x)_i)) \tag{2.15}$$

$$l(x) = \sum_{i=1}^{V} -log(p(\hat{x} = x_i | \phi(x))) \tag{2.16}$$

where $\phi(x)$ denotes the probability distribution output of the decoder.

2.3.3.5 Translation invariant embeddings

An interesting turn of events is the work of Huang et al. [113] that extends the use of LSA [114] to the multilingual case. In their work, they extract similar representations by optimizing the objective function (eq 2.17):

$$\min_{A,V} ||\tilde{\mathbf{X}} - \mathbf{A}\mathbf{V}^T||_F^2 \tag{2.17}$$

where $\tilde{\mathbf{X}} = \frac{1}{4}(\mathbf{X} + \mathbf{D}_1\mathbf{X} + \mathbf{X}\mathbf{D}_2^T + \mathbf{D}_1\mathbf{X}\mathbf{D}_2^T)$, and X is the single co-occurrence matrix of joint words from two corpora C_1 and C_2 as rows and joint context (N_1, N_2) as columns of the co-occurrence matrix, D_1 is the word dictionary matrix, and D_2 is the context dictionary matrix, and lastly, A and V are word embedding and context embedding matrices respectively.

2.3.3.6 Taking advantage of Wikipedia's structure: Inverted indexing

Wikipedia has proven to be a resourceful tool for many NLP applications including machine translation, dictionary induction, etc. Apart from that, Wikipedia houses a plethora of multilingual documents (e.g. articles) written in multiple languages related

by concepts. Søgaard and Bohnet [115] exploits this resource containing over 6 million articles covering 309 languages to extract shared representations. In their approach, they constructed a matrix of concepts and associated terms used to describe that concept from multiple languages. With this setting, a single word is a row in this matrix. They used LSA technique similar to Huang et al. [113] to learn embeddings.

2.3.3.7 Language-Agnostic sentence representations (LASER)

Recently there has been an increasing interest in learning representations for longer linguistic units such as sentences and phrases to supplement the limitations of word embeddings [86, 116, 117, 118, 119]. In [120], authors proposed a language-agnostic general-purpose massively multilingual representations for over 93 languages using a simple stacked Bidirectional LSTM model trained with a cross-entropy objective and bite-pair encoding(BPE) vocabulary learned over the concatenation of all training data as input. The advantage of their work is that not only do they extract embeddings for multiple languages at the same time, but the induced embeddings can be used for any downstream tasks without additional task-specific fine-tuning as opposed to other works. This achievement is invaluable and has major positive implications for low-resourced languages with extreme scenarios.

CL training models focus on preserving correspondences between projected words and neglect the monolingual quality of an individual language separately. This issue is addressed by joint objective CL models.

2.3.4 Joint objectives cross-lingual models

Formally, a general joint-objective CL model takes the form: $\alpha(Mono_1 + Mono_2) + \beta B_i$, where $Mono_1$ and $Mono_2$ are monolingual objective functions, B_i is the bilingual objective, and finally, α and β are constant regularizers scaling contribution of the objectives combined. In the following subsections, we discuss models that use joint-objective CL to learn cross-lingual embedding spaces.

2.3.4.1 Multitask learning to jointly induce continuous cross-lingual representations

The first practical work for inducing joint representations using joint objectives was initially proposed in Klementiev et al. [44]. In their work, they used a language objective (first term in equation 2.18) to learn syntactic and semantic similarity in a language and combined this with the multitask objective (second term) to ensure that the induced representations are equivalent across languages as follows:

$$L(\theta) = \sum_{l=1}^{2} \sum_{t=1}^{T^{(l)}} log\hat{P}_{\theta^l}(w_t^l | w_{t-n+1:t-1}^l) + \frac{1}{2} c^T (A \otimes I_d) c \tag{2.18}$$

where l denotes a language, T^l is the number of words in l, w_t is a word at time t, and A is the interaction matrix capturing relatedness of tasks modeled as a bipartite graph, and lastly, $\otimes$ denotes the Kronecker product and can be used for multitask learning as illustrated by Klementiev et al. [44].

2.3.4.2 Bilingual word embeddings for phrase-based machine translation(MT)

A similar attempt to model both monolingual and bilingual semantics within and across languages is presented by Zou et al. [26]. In their approach, they use an objective function proposed by Collobert and Weston [24] to learn the monolingual embedding of English. They then learn Chinese monolingual embeddings bootstrapped by alignments obtained from Berkely Aligner [121] using the equation 2.19 below:

$$J_{CO}^{c,d} = \sum_{w^r \epsilon V_R} max(0, 1 - f(c^w, d) + f(c^{w^r}, d)) \tag{2.19}$$

From 2.19 above, f is the function defined by the neural network, w^r is a word chosen in a random subset of the vocabulary V_R, and c^{w^r} is the context window containing the word c^{w^r}. With these monolingual embedding extracted, they can then jointly learn a shared space between English and Chinese to preserve translation equivalence using the following equations:

$$J_{TEO-en-zh} = ||B_{zh} - A_{en-zh} B_{en}||^2 \tag{2.20}$$

$$J_{TEO-zh-en} = ||B_{en} - A_{zh-en}B_{zh}||^2 \tag{2.21}$$

where B_{en}, B_{zh} is the embedding matrix of English and Chinese respectively, A_{en},A_{zh} are alignment matrices. TEO refers to translation equivalent objective. During bilingual training to preserve both monolingual and bilingual similarity, they optimized the sum of the monolingual and bilingual objectives :

$$J_{CO-zh} + \lambda J_{TEO-en-zh} \tag{2.22}$$

for extracting Chinese embeddings and

$$J_{CO-en} + \lambda J_{TEO-zh-en} \tag{2.23}$$

for extracting English embeddings.

2.3.4.3 BiSkip

Luong et al. [122] learns semantic similarities in a language and equivalences across languages by using co-occurrence signals from the monolingual context and semantic equivalence from parallel data. In their approach, they extend the SG model into the bilingual context. To achieve this, instead of using the center word w in language $l1$ to predict the context c in $l1$ they replaced w in $l1$ with aligned word w' in $l2$ and vice versa. This can be observed collectively as training four SG models jointly to predict words between pairs of the following language: $l1$ -> $l2$, $l1$ -> $l1$, $l2$ -> $l1$, $l2$ -> $l2$.

2.3.4.4 BilBOWA

Gouws et al. [123] is amongst the great efforts that place emphasis on the notion that besides learning how words relate to each other within a language, it is equally important to learn representations that capture equivalences across two or more languages when developing joint representations. To demonstrate the feasibility of their argument they proposed bilingual representations trained using a monolingual embedding objective of SGNS and a bilingual objective in (eq 2.24):

$$\Omega_A^{(t)}(\mathbf{B}^e, \mathbf{B}^f) \approx ||\frac{1}{m}\sum_{w_i \epsilon s^e}^{m} b_i^e - \frac{1}{n}\sum_{w_j \epsilon s^f}^{n} b_j^f||^2 \tag{2.24}$$

where B^e, B^f denote the embedding representations for the two languages, and b_i describes the embedding learned for a particular word w_i. This bilingual objective assumes uniform alignments of words between parallel data and hence comes with the added advantage of not calculating alignment matrices which tend to be computationally costly.

2.3.4.5 Trans-gram

Coulmance et al. [124] proposed another extension to the SG model for the multilingual case. Monolingual embeddings are extracted using a conventional SG objective while cross-lingual embeddigs with translation equivalence in mind are extracted with a modified SG objective[1]. Contrary to the conventional SG model, the CL objective defines the context of a word w in language L_1 as all the words in parallel sentences in language L_2.

2.3.4.6 Sparse representations

Although these CL objective models perform well in extracting useful representations for downstream tasks, finer-grained representations that can capture semantics beyond translation equivalence are also important for tasks such as cross-lingual lexical entailment. Vyas and Carpuat [125] creates the first cross-lingual entailment tasks and benchmarks this problem using sparse representations as opposed to dense representations of earlier discussed methods. In their approach, they first extract monolingual embeddings for each language using a method proposed in [87], i.e. using GloVE embeddings. To obtain sparse bilingual embeddings, they optimized a modified version of the model proposed in [43] as follows:

$$\begin{aligned}\underset{A_e, D_e, A_f, D_f}{\operatorname{argmax}} \sum_{i=1}^{v_e} ||\mathbf{A}_{e_i}\mathbf{D}_e^T - \mathbf{B}_{ei}||_2^2 + \lambda_e||\mathbf{A}_{ei}||_1 + \sum_{i=1}^{v_f}\frac{1}{2}||\mathbf{A}_{ff}\mathbf{D}_f^T \\ -\mathbf{B}_{fj}||_2^2 + \lambda_f||A_{fj}||_1 + \sum_{i=1}^{v_e}\sum_{j=1}^{v_f}\frac{1}{2}\lambda_b\mathbf{S}_{ij}||\mathbf{A}_{ei} - \mathbf{A}_{fj}||_2^2\end{aligned} \tag{2.25}$$

2.3.4.7 Cross-lingual context with and without word alignments and cross-lingual similarity(CLSim) objective

Co-occurrence matrix co-factorization techniques for deriving monolingual embeddings have shown good results in practice [87]. Shi et al. [126] exploits the expressive power of global vector representation(GloVe) to extend it to a cross-lingual setting. In their approach, they extracted both monolingual and cross-lingual embedding for each language using the objective defined in [87]. Matrix factorization methods require co-occurrence statistics. To obtain context alignment information used to tie languages together, they proposed two approaches:

1. Using a technique that assumes uniform distribution between parallel sentences similar to definition in [123],
2. A method that replaces the context of each word w in language $l1$ by the context of word k in $l2$, where k is the translation equivalent of w in $l1$.

Alternative to the initial CL objective used in their study, they used the Euclidean distance between monolingual embeddings weighted by their similarity scalar as the cross-lingual constraint. This is referred to as the CLSim objective in their study and showed to be the best-performing model.

2.3.4.8 Bilingual paragraph vector(BRAVE)

[127] proposed a model inspired by Mikolov's paragraph vector distributed model(PV-DM)[88]. Unlike other works, their approach relaxes the reliance of both word-aligned and sentence-aligned corpora to weakly aligned corpora - topic-aligned or label-aligned corpora, which is effectively cheap to find. Furthermore, they model cross-language equivalences using the concept of elastic net regularization [128]. In recent work on cross-lingual embeddings has gravitated from models that explicitly extract cross-lingual embeddings to large models that implicitly extract joint representation through shared subword vocabulary training prior to the advent of the transformer models [21]. We discuss this group of models in the next subsection.

2.3.5 Large multilingual models for extracting joint representation

Recently, research has gravitated towards extracting shared embeddings between languages using large pre-trained language models such as BERT [21]. Surprisingly, BERT has no explicit cross-lingual objective, however, this model works exceptionally well for learning multilingual representations. BERT only uses a masked language modeling objective, where tokens within a sentence are randomly hidden and the task of the model is to predict the identifier of the hidden token [129]. Another interesting characteristic of BERT models is that it uses large transformer models which are currently state-of-the-art models for extracting abstract representations [130]. There has been active research on adopting BERT-based models for extracting contextual representations in order to draw conclusions on their surprisingly good performance for downstream tasks [131, 132, 133].

On that note, Lauscher et al. [134] conducted an experiment on mBERT (variation of BERT trained on multiple languages covering up to 104 languages) and XLM-R (another variation of BERT) [135] and found empirical evidence supporting the correlation between cross-lingual transfer performance and language similarity. In addition, they found that the size of the target data plays a vital role in extreme scenarios such as zero-short transfer learning. Furthermore, their study demonstrates the efficacy of few-shot transfer learning. On a similar trend, Nooralahzadeh et al. [136] supplemented a large multilingual model with meta-learning [137] and converged to the same conclusions made by Lauscher et al. [134] - morphosyntactic commonalities between languages may be a positive contributor of transfer performance.

de Vries and Nissim [138] proposes a model adoption approach for generative language models. Their approach starts by first training a GPT-2 [139] model on the English data to initialize the model. They then adapt this model on low-resourced languages by removing its lexical embedding layer(first layer) and freezing its parameters during additional fine-tuning on the target data. That is, only the first layer is adapted to the low-resourced language. The intuition is that by preserving the learned GPT-2 model's weights during adaption, the newly learned lexical embeddings are still aligned with the space of the English model(where there is high knowledge density[140])

Gogoulou et al. [79], applies a similar model adoption approach as de Vries and Nissim [138]. In their paper, they first initialize the embedding layer(constructed using BERT) by training on a source monolingual task and using the same BERT to further fine-tune on the target monolingual task. They evaluated their approach on the English language dataset and their method superseded a model trained and tested on English monolingual data. Interestingly, their findings contradict Lauscher et al. [134] by showing that language similarity has no impact on transfer performance. As these discoveries are still new, they merit additional examination.

Multilingual language models such as BERT are currently in their preliminary phases of research in the context of cross-lingual NLP, thus no general consensus has been reached as to what constitutes their generalization ability. Initially, the attribute was given to shared subword vocabulary and joint training across multiple languages. However, recent works are in conflict with this conclusion [78, 79]. Therefore, more work is essential in this research direction in order to understand the root cause of high-level abstractions extracted using these multilingual language models. Furthermore, BERT models require large amounts of computing resources to be effective, therefore exploring alternative approaches that can scale to settings with limited computing without compromising performance on downstream tasks poses an interesting avenue of research [141].

In the next section, we summarise the discussed models in tabular format. That is, we break down each work by looking at its family of cross-lingual models, data used, resources and type of resource(e.g. bilingual lexicons, or parallel data) used, and the downstream task investigated in the work.

2.3.6 Summary of projection models

In Tables 2.1, 2.2, and 2.3, we summarise these cross-lingual works by looking at their category, pre-training datasets, monolingual embeddings, cross-lingual embedding method, type of extracted embedding (e.g. sparse or dense), evaluation tasks (intrinsic and/or extrinsic), and lastly, available resources.

Table 2.1: Summary of projection models

		Embedding Extraction					Evaluation of embedding method			
Projection Type	**Ref**	**Data**	**Bilingual signal**	**Monolingual embeddings**	**CL Model**	**Representation**	**Evaluation Data**	**Intrinsic**	**Extrinsic**	**Code**
	[2]	WMT-2011, Google News datasets	dictionary,	CBOW	Linear Projection	Dense	-	-	MT	code
	[43]	WMT-2011, WMT-2012,	word alignment counts[142]	LSA [114]	CCA	Sparse		SynRel,WordSim, SemRel	-	resource, resource, code
	[90]	WMT-2011	dictionary	Skip-gram	Normalize and Orthogonal transform model	Dense	WordSim353	WordSim	BLI	-
	[91]	Corpus1, Corpus2	dictionary	CBOW	Max-margin loss	Dense	Test Dictionary	-	BLI, Image labeling	-
Offline models	[92]	Europal	dictionary	ClusterID, LSA	multiCluster(Skip-gram), multiCCA	Dense and Sparse	MEN[143], RVC Corpus , Universal dependencies, MWS353, Stanford's Rare Words[144]	WordSim	BLI, MDC, MDP	resource
	[94]	Wikipedia Corpora	-	Skip-gram	Aversarial Auto Encoder	Dense	Reuters Corpora [44]	-	CLDC [44]	code
	[48]	Wacky crawling corpora	-	CBOW	Unsuperviced self-learning	Dense	[145, 146]	BLI	-	code
	[97]	-	-	FastText	Mini-Batch Cycle Iterative Closest Point	Dense	MUSE	-	Word translation	resource
	[98]	Wikipedia	-	FastText and Dinu[145]	Gromov Wasserstein distance	Dense	-	Qualitative analyses	-	-
	[99]	Wikipedia	-	FastText and Dinu[145]	Convex relaxation of CSLS	Dense	Wacky dataset	-		code
	[100]	ukWAC, Wikipedia, BNC,itWAC	dictionary	CBOW	Discriminative latent-variable model	Dense	WordSim353	BLI, WordSim, RG-65,	-	code
	[103]	-	dictionary	fastText	dubbed Procrastes model	Dense	multilingual XNLI corpus, TED CLDC	BLI	XLNLI, CLDC	resource

Table 2.2: Summary of projection models

		Embedding Extraction					Evaluation of embedding method			
Projection Type	**Ref**	**Data**	**Bilingual signal**	**Monolingual embeddings**	**CL Model**	**Representation**	**Evaluation Data**	**Intrinsic**	**Extrinsic**	**Code**
	[3]	-	Wiktionary	-	DNN	Dense	CoNLL	-	CL dependency parsing	-
	[106]	-	Wiktionary	-	BARISTA(CBOW)	Dense	Google universal tagset	-	POS tagging, CL super sense tagging	code
Pseudo CL	[107]	Wikipedia	dictionary	-	CBOW	Dense	Reuter RCV1/RCV2	Word similarity	BLI, CLDC	-
	[108]	-	-	document-aligned	Skip-gram	dense			bilingual lexicon extraction and suggesting word translations	
	[5]	Europarl	sentence aligned	Gaussian distribution	BiCVM	Dense	Reuters RCV1/RCV2[44]	-	CL document classification	code
	[109]	Europarl, TED corpus	document aligned	Gaussian distribution	DOC	Dense	Reuters RCV1/RCV2[44]	Qualitative analyses	CL document classification	code, resource
CL models	[110]	Europarl	document aligned	bag-of-words/TFIDF	Auto-encoder	Dense	Reuters RCV1/RCV2[44]	-	CL document classification	-
	[111]	[147]	FASTALIGN and parallel sentences	-	Distributed word alignment	Dense	Reuters RCV1/RCV2[44]	Qualitative analyses	CL document classification	-
	[112]	Europarl	parallel sentences	bag-of-words/TFIDF	Auto-encoder	Dense	Reuters RCV1/RCV2[44]	-	CL document classification	code, resource
	[113]	[25]	dictionary matrix	co-occurrence statistics	Translation invariant LSA	Sparse	[25]	Word Similarity[43]	CL dependency parsing	code
	[115]	Wikipedia	Wikipedia indexing/concepts	co-occurrence statistics	CBOW, Skip-gram	Dense	AMAZON, Google Universal Treebanks, CoNLL 2006/7	Word alignment	CL document, POS tagging, dependency parsing, classification	code resource
	[120]	OPUS	parallel sentences	BPE embeddings	BiLSTM	Dense	[31], MLDoc[148], BUCC[149, 150]	Taboeba: Similarity search	Bitext mining, XNLI, CL document classification,	code

Table 2.3: Summary of projection models

		Embedding Extraction					Evaluation of embedding method			
Projection Type	**Ref**	**Data**	**Bilingual signal**	**Monolingual embeddings**	**CL Model**	**Representation**	**Evaluation Data**	**Intrinsic**	**Extrinsic**	**Code**
	[44]	Europarl[151], RCV1/RCV2[152]	cooccurance statistics and parallel sentences	neural model[84]	Bengio neural[84] + Multitask objective	Dense	RCV1/RCV2[152]	-	CL document classification	code
	[26]	Gigaword corpus	MT word alignments, parallel text	[24]	Neural net(Translation Equivalent objective + collobert joint[24])	Dense	OntoNotes,NIST08	Vector matching alignment, Semantic similarity	Phrase based MT,NER	-
Joint CL models	[122]	Europarl V7[151]	parallel sentences	Skip-gram with negative sampling	Biskip	Dense	WordSim353[144], RCV1/RCV2[44], NN dataset[112]	Word Similarity, Nearest Neighbor	CL document classification	code
	[123]	Europarl	parallel sentences	Skip-gram with NS	BilBOWA	Dense	Setup in[44], WMT	-	Word translation, CL document classification	code
	[124]	Europarl[151]	parallel sentences	Embeddings[153]	Tran-Gram	Dense	RCV1/RCV2[44],WMT		Word translation, CL document classification	-
	[125]	Wikipedia, Gigaword	cooccurrence statistics	Glove [87]	Sparse bilingual vectors	Sparse	-	-	lexical entailment	-
	[126]	RCV1/RCV2[44]	co-occurrence statistics/ word alignments	Glove [87]	CLC-WA, CLSim	Sparse	RCV1/RCV2[44]	-	CL document classification	-
	[127]	Europarl[151]	parallel sentences/topic-aligned/label-aligned corpora	paragraph vector [86]	BRAVE	Dense	RCV1/RCV2[44], TED, CLSC[154]	-	CL document classification, multi-label classification, CL sentiment classification	-
	[21]	Wikipedia, BooksCorpus[155]	-	-	-	-		-	GLUE tasks	-
	[134]	Universal Dependency treebank [156], NER-WikiANN [157], XNLI[31], XQuAD[78]	-	-	-	-	-	-	POS tagging, Parsing, NER, XNLI, CL NLI, and CL Question Answering	-
MLM	[136]	MultiNLI[158], XNLI[31], and MultiQA[159]	-	-	-	-		-	XNLI and QA	-
	[138]	GePpeTto[160], Wikipedia, Articles[161]	-	-	-	-		GePpeTto[160], SonaR[162]	MT	-
	[79]	Wikipedia[2]	-	-	-	-		-	GLUE tasks	-
	[42]	WikiAnn NER[163], MasakhaneNER[164], Universal dependency treebank[165]	Panlex[3]	-	-	-		-	WikiNER, MasakhaneNER, POS tagging, and Parsing	-

With simple and sophisticated cross-lingual models discussed above, it is, therefore, necessary to discuss how these models have been adopted in both intrinsic and extrinsic downstream tasks. Section 2.5 discusses cross-lingual models and their adoption in application using a similar methodology adopted by Magueresse et al. [166].

2.4 Related Work

Research in NLP with the main focus of applying cross-lingual representations as a tool to mitigate the issue of limited resources has gained momentum over the years. Therefore, the need to keep track of advancements and limitations of existing methods and universally inscribe them for future research remains inevitable, especially for those languages where there is no work available. An earlier work attempting to achieve this is a systematic literature survey looking specifically at the trajectory of projection models used to achieve both cross-lingual and multilingual embeddings for low-resourced settings [167]. Ruder et al. [167] dates back to 2017 when multilingual language models(MLM) or contextual language models [21], for constructing joint representations were not in active cross-lingual embedding research. This means MLMs are not an active feature in their work. Magueresse et al. [166] on the other hand took an application-oriented approach to review existing works utilizing projection methods to cross-lingual embedding spaces. In their paper, they focused on applications such as part-of-speech tagging, named entity recognition, morphology induction, sentiment analysis, etc, to review past and recent work on projections and further provided insightful future directions that can be explored to improve performance in each application domain. Contrary to our work, their paper only discusses automatic alignment strategies used to create parallel data and not cross-lingual models. Our work can be viewed as an extension of Ruder et al. [167] for reviewing models that construct joint representations and Magueresse et al. [166] on how these models have been successfully applied in practice. This chapter assembles the aforementioned works and extends their arsenal with recent literature on cross-lingual works. In addition, a taxonomy (see tables 2.3.6, 2.3.6, and 2.3.6) of cross-lingual contribution to NLP research is highlighted.

The term “low-resourced” is a multi-faceted concept as discussed in [168]. As a

result, many works we will be reviewing here have reduced the facets to refer to only data scarcity. To begin placing cross-lingual models under scrutiny, it is immediately necessary to first discuss what embeddings are and what they bring to NLPs' playing field. This is what the next section aims to achieve.

2.5 Cross-lingual application in the wild

In this section, we discuss the scale of cross-lingual models to downstream tasks. We adopt a three-layered architecture for this section. First, we describe a downstream task. This is followed by identifying a cadre of cross-lingual models that have been used to tackle this task for low-resourced languages. Finally, we identify limitations that still need to be addressed in the downstream task and propose future works that can help alleviate some of these issues.

2.5.1 Cross-lingual dependency Parsing for low-resourced languages

Dependency parsing (DP) is a form of syntax parsing in which the main task is to map words from sentences to their dependency structure [6]. Formally, given a sequence of tokens $\{X = x_1, x_2, x_3, ...x_n\}$ representing a sentence $s \in S$ with n words, the task is to map each x_i in X to its dependency relations x_j in X following the rules of dependency grammar. The x_i, and x_j in this formulation can be described as head and dependent respectively. In Figure 2.5, we show a typical dependency structure of an English sentence, where dependencies are represented by arrows connecting the head word and its subordinate word. Moreover, each relation connection has a type. For example, subject(SBJ) is the type between a noun *news(dependent)* and the auxiliary verb '*had*' is shown in Figure 2.5.

Other types include noun phrases (NP), objects (OBJ), verb phrases(VP), etc. There are four main dependency representations: context-free dependency parsing, constraint dependency parsing, graph-based dependency parsing, and transition-based dependency parsing. We will not be discussing these dependency representations.

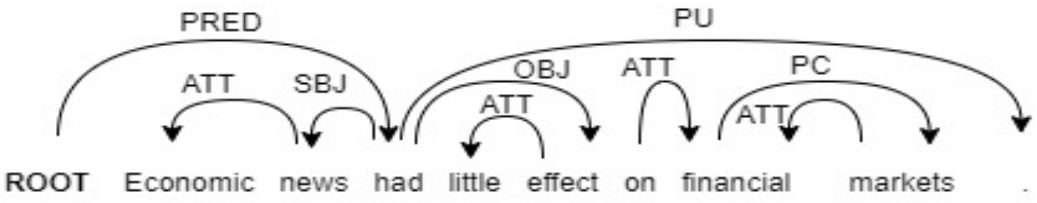


Figure 2.5: Dependency structure of an English sentence [6]

There has been an uneven distribution of research effort in exploring DP among NLP languages, with high proportions of effort dedicated to English and other high-resource languages [169, 170, 171]. For this reason, high quantities of supervised and annotated resources such as treebanks and Universal dependency datasets aid the advancements in research in DP for high-resourced languages. While this is encouraging, research progress (including data collection, and model development) for other unattended languages has witnessed a great stall. Leveling the playing field for these neglected languages proved to be labor-intensive and time-consuming; thus, researchers resorted to CL training. The main aim of learning cross-lingual embeddings is to preserve bilingual or multilingual translation equivalence in the learned shared space. The resulting shared space supplements the knowledge spaces of limited languages. Guo et al. [25] trained two linear-based projection models to induce a cross-lingual space for CL dependency parsing. Their projection approaches use monolingual embeddings extracted using CBOW [1] and parallel data.

2.5.1.1 Alignment-based projections with CL model

This approach uses the learned monolingual embedding space (S) from the source language and then projects this space to the target space (T). To achieve this, they first extract an alignment dictionary $A^{T|S} = \{(w_i^T, w_j^S, c_{i,j}), i = 1, 2, .., N_T; j = 1, 2, .., N_S\}$, where c is the number of times the words w_i^T, w_j^S are aligned in the parallel corpora and N_S, N_T is the vocabulary size of source and target respectively. They then used the bilingual dictionary D to find projections using the equation below denoting the

weighted sum of translation words, following the equation 2.26:

$$v(w_i^T) = \sum_{(i,j)\epsilon D^{T|S}} \frac{c_{i,j}}{c_i} * v(w_j^S) \tag{2.26}$$

To account for OOV words, they implemented a morphologically-driven approach to extract their embeddings. Specifically, they used edit distance C to extract a list of similar words and used the average vector of the words as a representation of OOV words (see equation 2.27):

$$v(w_{oov}^T) = \underset{w'\epsilon C}{\text{Avg(v(w}'\text{))}}, \tag{2.27}$$

where $C = \{w|EditDist(w_{oov}^T \leq \tau)\}$

2.5.1.2 Cross-lingual models using Canonical Correlation Analysis (CCA)

The implementation of Cross-lingual models using Canonical Correlation Analysis (CCA) [25, 43] is to learn continuous joint representations to compare with their proposed method. Their results demonstrate the superior performance of alignment-based projections over CCA projections. The study did not evaluate the cross-lingual embeddings induced by their projection models to investigate the knowledge (semantic, syntactic) ingrained in their space. Evaluating embeddings on semantic and syntactic level can be useful in inferring which downstream task is most likely to have high-performance gains. It is, however, worth mentioning that the study used insights from other papers to infer that their induced monolingual embeddings might be potentially useful for syntactic-related tasks.

Lynn et al. [172] converted the Irish dependency treebank into a universal dependency scheme and performed direct CL transfer parsing between Irish and ten languages from four family groups. The transfer was possible because: 1.) parser training is delexicalized; and 2.) input representations are encoded using the universal POS tagset and universal annotation scheme. Recent work in cross-lingual models is focused on large multilingual language models [21], but in the end, better performance gain is due to having access to monolingual data in the target language. This basically means that performance is limited when there is little to no monolingual data available. For this reason, it is important to mention the works in [42], where alternative strategies were proposed using bilingual lexicons for adapting multilingual models. The author proposed

an adaptation mechanism that uses monolingual data from the high source language X_{source} to generate pseudo monolingual text in the low resource target language X_{pseudo} using paired bilingual lexicons $D_{source,target}$ and adapted a BERT [21] model using the pseudo generated data. The method observed competitive results with the state-of-the-art (SOTA) results using dependency parsing, POS tagging, and named entity recognition.

Most works base their projections on English as the source language. For future work, it would interesting to devise a measure of language proximity and use one or several languages with the highest measure of proximity with the target language as the source language(s). An interesting direction in this thought would be to use the Wasserstein distance as proposed in [98] to measure language proximity. This argument is also highlighted in [166].

2.5.2 Part of speech (POS) tagging

POS tagging is the task of assigning words in a given sentence to a grammatical category - noun, verb, object, etc. POS tagging is the natural first step to many NLP applications such as information retrieval, sentiment analyses, entity recognition, etc [173, 174]. Many resource-rich languages such as English, and French have witnessed research attention for POS tagging resulting in a plethora of resources for these languages. However, this is not the case for low-resource languages. To address this, CL approaches have been widely adopted for POS tagging in low-resourced settings. The approach taken in Kim et al. [27] closely resembles joint objective learning, where the task is to optimize both the monolingual objective M_1, M_2 for languages L_1, L_2 respectively, and the CL objective. Their joint architecture for learning POS tags comprised a common Bidirectional LSTM for learning language-agnostic representations, a private Bidirectional LSTM for learning language-specific representations, and a softmax layer for tag prediction. The parameters of this architecture were optimized using the following objectives:

2.5.2.1 The loss function

The outputs of the common and private BiLSTM are summed and passed through the softmax layer for prediction. The error of the resulting prediction is propagated using the loss functions as follows (equation 2.28):

$$L_p = -\sum_{i=1}^{S}\sum_{j=1}^{N} p_{i,j} log(\hat{p}_{i,j}) \tag{2.28}$$

where S is the number of sentences in the given minibatch, N is the number of words for the given sentence, $p_{i,j}$ is the label of the j^{th} tag of the i^{th} sentence in the given minibatch, and $\hat{p}_{i,j}$ is the predicted tag.

2.5.2.2 Monolingual language modeling

Their paper used language modeling objectives to learn language-specific representation motivated by Rei [175]. They optimized the following language model represented in equation 2.29:

$$L_l = -\sum_{i=1}^{S}\sum_{j=1}^{N} log(P(w_{j+1}|f_i)) + log(P(w_{j-1}|b_j)) \tag{2.29}$$

where f_j and b_j are the j^{th} forward and backward output representations of the BiLSTM.

2.5.2.3 Cross-lingual training using language-adversarial models

Cross-lingual embeddngs are learned using language-adversarial training of Chen et al. [176]. To achieve this, the output of the common BiLSTM is passed to a CNN with Max pooling and the resulting vector representation serves as input to a language discriminator. Then, the error gradients of the language discriminator are propagated back to the BiLSTM with opposite signs to encourage the common BiLSTM to be language-agnostic using the following loss (see equation 2.30):

$$L_a = -\sum_{i=1}^{S} l_i log\hat{l}_i \tag{2.30}$$

where S is the number of sentences, l_i is the language of the i^{th} sentence and $\hat{l}_i$ is the Softmax output of the tagging.

The overall training optimizes the loss as follows (see equation 2.31):

$$L = w_s(L_p + \lambda L_a + L_l) \tag{2.31}$$

where λ is a parameter that stabilized training and was set to increase from 0 to 1, and w_s is a weighting mechanism to give different weights to the source and target language.

Many tasks such as dependency parsing build on the task of POS tagging. For this reason, gold standard resources specifically for this task are needed for ongoing research. Therefore, emphasis should be put on collecting gold standard resources for this task for low-resourced languages lacking them. In collecting such resources, Lignos et al. [177] suggests extensively involving native speakers for quality assurance.

2.5.3 Text Classification: propaganda detection, topic classification, etc

Text classification involves a model Φ, a pair of input and output training patterns $X_{train} = \{x_i^{train}, y_i^{train}\}$, and a test set $X_{test} = \{x_i^{test}, y_i^{test}\}$, where the task is to optimise the parameters of Φ on X_{train} to correctly classify unseen instances in x_i^{test} into categories or labels y_i^{test}. Specifically, x_i^{train} is a dataset of longer linguistic units such as sentences, paragraphs, or documents and the task is to understand x_i^{train} through optimizing Φ. For example, given $x_i^t \in x_i^{train}$, the goal would be to use Φ to be able to learn, e.g polarity (in the case of sentiment analysis), the overall theme of x_i^t (in the case of topic classification), useful content patterns (in the case of propaganda detection), etc. These models, Φ, vary from classical machine learning models such as SVM, linear regression, and Naive Bayes Models to deep learning models such as CNNs and the more popular Transformer models [130].

Artetxe et al. [78] investigated their MonoTrans adaption approach on four text classification tasks: cross-lingual natural language inference, cross-lingual document classification, Paraphrase identification, and cross-lingual question answering. Their results are competitive with the state-of-the-art in each task.

Niyongabo et al. [178] implement a linear CL model that relied on embeddings of resource-rich language and mutually intelligible languages. That is, author used embed-

dings of Kinyarwanda (high-resourced) to train a Kirudi (low-resourced) classifier using Kirundi data. This is possible due to the two languages being highly mutually intelligible (approximately 32% coverage as indicated in their paper). They further argued the efficacy of their embeddings on three tasks: propaganda detection, sentiment analyses, and multiclass classification of news articles.

Robnik-Sikonja et al. [179] experimented on LASER embeddings and cross-lingual embeddings extracted from the BERT language model and evaluated their performance on a sentiment analysis task. Using BERT and LASER embeddings for transferring between 13 low-resourced languages, the study observed outstanding results, advocating the full power of multilingual BERT models on extracting hidden joint representations. They observed great performance on languages belonging to the same family. The empirical result of their study demonstrated that embeddings extracted from BERT models are superior to other methods, however, this warrants further investigation.

Espinosa-Anke et al. [77] constructed joint representations between Welsh (low-resourced) and English using a supervised version of VecMap with Skip-gram and FastText as monolingual embeddings and evaluated their embeddings on a sentiment analyses task.

Learning finer-grained representations with high vocabulary overlap such as subwords through BPE [180] has demonstrated effectiveness in NLP downstream tasks. An interesting future research avenue would be to, instead of learning word embeddings and projecting them into a joint space, rather learn sub-word embeddings and project embedding spaces of sub-words into a joint space. The hope is that this space may cover more inter-lingual similarities as opposed to word-based spaces due to high sub-word overlap and may facilitate greater transfer performance. Finally, parallel signals such as word-to-word or sentence-aligned are generally difficult to get and often not available for low-resourced languages. For this reason, methods that massage this requirement by adopting weak supervision signals such as noun-noun-equivalences should be interesting research avenues [106]. However, simply adopting such shallow signals may lead to suboptimal results. Therefore, supplementing these approaches by mechanisms that explicitly learn the type of equivalence as well as how to exploit the equivalences may be beneficial to the methods. For example, a weighting strategy can be adopted to police

the contribution between common-noun-abstract-noun equivalences or common-noun-common-noun equivalences between languages and allow higher equivalences to have higher contributions to the learning of embeddings.

2.5.4 Word similarity (WordSim) and Simlex tasks

WordSim evaluation is an intrinsic evaluation technique that measures the similarity of representations using cosine measure [85]. For example, given a test set of words, the task is to project these words into the embedding space and extract similar words using cosine measure(see [85] for a detailed description). To assert that the extracted embedding space contains knowledge about the language, similar words must be extracted from this process. The distribution of extracted words is visualized using the t-SNE [181] tool. This approach can be simply extended to the cross-lingual case by projecting test words $\{X = x_1, x_2, ..x_n\}$ in language $l1$ into the shared embeddings space W and extracting words in languages $l2$ that are neighbors of X in the shared space. The objective is to observe if the extracted words are indeed similar to those in X. WordSim tasks are a great resource since they allow evaluating the embedding space directly. This is more useful for shared representations, where we are interested in understanding the content embedded in our shared spaces.

Approximately 9 out of the 11 official South African languages, with the exception of English and Afrikaans, are classified as low-resourced languages [168]. In this regard, Makgatho et al. [37] investigated CL transfer learning for Sepedi, and Setswana (Niger-Congo languages) using VecMap [78] for extracting joint representations. VecMap [78], requires monolingual embeddings for each observed language. For that, Makgatho et al. [37] extracted monolingual embeddings using the SG model and FastText [182] for comparison. They evaluated their embedding model on the Sepedi and Setswana versions of the WordSim task by translating the WordSim dataset to Sepedi and Setswana.

Future work for Makgatho et al. [37] could include investigating different transfer mechanisms as they used only zero-short transfer learning, evaluating their embeddings on a downstream task, expanding their approach to the multilingual case, using multiple language sources and investigating their impact, and perform comprehensive analyses and comparisons with recent models (if applicable).

2.5.5 Bilingual lexicon induction (BLI)

Bilingual lexicon induction (BLI) is a popular intrinsic evaluation task to suggest translation in word context and is predominantly used to evaluate linear-based projection models. BLI is the task of extracting translation pairs(X_{src}, X_{trg}) for source words X_{src} of the test corpus by projecting the source words X_{src} in the induced shared space and extracting neighbors of the projected words. CL models demonstrated the importance of bilingual lexicons as an indispensable tool for extracting shared embedding spaces. However, some languages present NLP practitioners with extreme scenarios where even shallow bilingual lexicons are not available. To address this, Vulic and Korhonen [183] proposed a hybrid method that first learns to induce seed bilingual lexicons using Pseudo-CL models and uses the extracted lexicons to train a more robust cross-lingual model for extracting shared embedding spaces. Their study deduced interesting conclusions: 1)., careful selection of lexicon pairs improves the performance of cross-lingual models, and 2)., the size of lexicons used together with an intelligent selection of lexicon pairs can further add performance gains.

2.5.5.1 Pseudo-CL model

Vulic and Korhonen [183] proposed a Pseudo-CL model by adopting a document-level cross-lingual embedding space proposed in Vulić and Moens [108] to extract equivalent words from different languages in this space by projecting words from a list of BNC words into this space and retrieving their nearest neighbors as translations. They then used the extracted pair of lexicons as seed lexicons to train a cross-lingual model.

2.5.5.2 $L2$ regularizer with cross-lingual model

Vulic and Korhonen [183] implemented the $L2$ regularizer with a cross-lingual model by combining it with the linear projection model proposed by Mikolov et al. [2] to translate source vectors into the shared target embedding space. To train the projection matrix, they used the seed lexicons extracted by the pseudo-CL model above. The goal is to minimise the $L2$ least-squares objective.

Numerous methods for selecting neighbors in a manner that avoids hubness have been

proposed and demonstrated to be effective. It would be interesting to observe future directions that focus on solving this issue from the embedding extraction viewpoint. That is, creating objective functions that avoid creating universal neighbor vectors.

2.5.6 Entity linking and Discovery

Entity linking, commonly known as Wikification is the task of using Wikipedia mentions (sub-strings in text) to identify their corresponding titles (entries) in Wikipedia or other related knowledge bases. The English language, similar to other downstream tasks, takes a lead in research coverage for this task [184, 185, 186]. However, such efforts and the associated groundbreaking results achieved in the aforementioned works have not been witnessed for other available non-English languages. The availability of such tasks for low-resourced languages opens up interesting avenues for exploration in both research and application such as developing information retrieval systems that serve as a value-add to these under-represented languages. Efforts towards achieving this have focused on using cross-lingual embeddings for disambiguating entities.

Interestingly, Lu et al. [29] took a rather peculiar route to find cross-lingual representations by using image searches/processing and multi-media techniques to find projections from high-resourced languages to low-resourced languages and evaluated their methods on entity linking and name tagging tasks. Similarly, Tsai and Roth [187] adopted a CCA[43] for projecting monolingual English and X (X is set of foreign languages) embeddings generated using the SG [1, 88] model. They evaluated these projected CCA embeddings on an Entity linking task and their results demonstrated promising avenues.

Cross-lingual Entity linking uses representation-similarity between mentions in a foreign language and titles in English as an indicator of relevant titles for a specific mention. It would be interesting to transform the same notion of similarity into the multilingual context where, instead of disambiguating X (foreign language)-English, we disambiguate X-[English, Y (other languages)]. The natural advantage of this is that if the English title is not available (which is often the case), other titles in languages can still be found and used. Furthermore, we can use the obtained content "Y" as a source and attempt to get the English titles using the same procedure.

2.5.7 Grammatical Error Correction (GEC)

GEC is the task of correcting incorrectly constructed sentences following the rules of grammar. This is interesting in the sense that, machine learning algorithms are learned to capture the morphology of a language instead of following case-based approaches. English is the most spoken language worldwide and has taken the lead in training GEC models [188, 189, 190]. Available resources such as data, models, and the wider research community have made it possible for the English language to see a rise in sophisticated and cutting-edge technology that can auto-correct incorrect sentences with exceptional accuracy such as Grammarly. This however is not possible for low-resourced languages such as Isixhosa (a South African Bantu language of the Nguni tribe), Shona (Zimbabwean language), Yoruba a Nigerian language), and many more prior insufficient resources including native researchers from these languages [177]. To remedy this, cross-lingual approaches have been adopted.

Yamashita et al. [28] used pre-trained Masked Language Model (MLM) / Translation Language Model (TLM) [135] to learn cross-lingual embeddings between four languages(English, Russian, Czech, and Japanese) and investigated its performance on a grammar error correction (GEC) task, where their main objective was to understand if the possibility of transferring grammatical knowledge across languages was possible. The study observed competitive results of using cross-lingual methods compared to only training on limited target training data. Their work curved interesting research directions for learning shareable grammatical knowledge between languages. It would be interesting for future work to observe how the transfer of the knowledge base of the grammar of one language is taking place and whether the rules of one language are applied as they are or modified according to the constraints of another language.

2.5.8 Speech

The aforementioned cross-lingual methods have also gained interest in other applications that involve speech data as opposed to text data. Schuster et al. [191] remedies the issue of limited speech data by employing cross-lingual transfer for multitask-oriented dialogue. Their language focus was English, Spanish, and Thai, with Spanish and Thai

being the target languages. Tu et al. [192] built an end-to-end text-to-speech model for low-resourced languages using cross-lingual techniques and proposed a Phonetic Translation Network model that learns transformations between language pairs to create the cross-lingual embeddings. The languages considered in this study were Mandarin, German, and French. Continuing with this line of speech data, Das and Hasegawa-Johnson [193] used a maximum likelihood goal to train a Gaussian mixture model (GMM)-based hidden Markov model (HMM) to learn how to represent phonemes across languages. The inclusion of speech-related works in this study was only to advocate the widespread use of cross-lingual transfer as a tool to mitigate limitations of data scarcity, availability, and quality. Thus, we were unable to investigate the speech data or models used to make bonded embedding spaces in this work. Instead, we talked about how to cross-reference joint embeddings in any situation based on the findings of these elicited works that share common ground. These are in the form of the following conclusions:

- Transferring models between languages of the same family can produce positive results [194], [195].
- The size of the pretraining corpus of the target language affects transfer performance.
- Embeddings (e.g. monolingual or cross-lingual) are generally better than baseline representations such as BOW, TF-IDF, etc for learning representations.

Despite the fact that these cross-lingual models were developed over 20 years ago, there has been insufficient theoretical and practical research in the NLP space of South African low-resource languages.

2.6 Conclusion

Cross-lingual projection models provide an easy and more computationally efficient alternative to other resource-collection methods such as machine translation for learning shared embedding spaces between languages. As such, these models play a vital role in facilitating progress in research and application for resource-unfortunate languages.

Furthermore, these models can serve as a basis for understanding how these under-represented languages function in isolation or in combination with other languages(either rich or poor). For these reasons and more, the evidence asserts that for researchers beginning research in this direction for their own endangered languages, to start with pioneer methods such as the ones proposed in [2, 43, 196] as demonstrated in this paper before adopting costly methods such as BERT [21] so that they can build up some level of understanding on their respective languages. Moreover, the Gromov-Wasserstein distance measure [101] could be a useful asset in solving the transfer language choice commonly mentioned in cross-lingual research. It would be interesting to observe what the future holds for these approaches as they are being slowly but effectively overwritten by large multilingual transformer models.

2.7 Summary

This chapter conducts a comprehensive survey of the evolution of cross-lingual models and their prevalence in application. To achieve this, we broke down the long history of NLP research that integrates cross-lingual models into tractable 6 tiers involving the introduction of language models, the discussion on monolingual embeddings, the coining of cross-lingual models, related works, and finally, the adoption of cross-lingual models in downstream application.

2.7.1 Introduction

Section 2.1 of this chapter provides a brief introduction to NLP and the trajectory of novelty in language processing. This is supported by strategically selected nomenclature that plays an integral protagonism role in the storyline.

2.7.2 Monolingual embeddings

Section 2.2 on the other hand discusses a revolutionary era of continuous representation of words as language models, referred to as monolingual word embeddings. In this section, we discussed how monolingual embeddings differ from their previous lan-

guage model counterparts in terms of language representation, processing implications, and downstream task learning. Furthermore, we also hinted at how these monolingual embeddings played a role in the establishment of cross-lingual models.

2.7.3 Coining of cross-lingual embeddings

This section provided an in-depth dive into cross-lingual models and discusses their mathematical formation, major drawbacks, and their mathematical implication for projecting representations to another space (see Subsection 2.3). Lastly, we created a taxonomy that groups cross-lingual models based on their characteristics such as resources used, training setup, end-representations of the models, evaluation metrics used, datasets, and downstream tasks addressed, etc (see Subsection 2.3.6).

2.7.4 Related works

Section 2.4 highlights similar works that provided captivating discussions on the formation of cross-lingual works and how our work added to this arsenal.

2.7.5 Cross-lingual models and application in downstream tasks

Section 2.5 discusses the application of cross-lingual models in various tasks such as POS, NER, GER, and Word Similarity. This literature provides evidence of the efficacy of cross-lingual models in extending the application of language technologies in low-resourced languages using cheap resources (e.g. bilingual lexicons), fewer expertise requirements, and less computing power.

2.7.6 Concluding remarks

Sections 2.6 provide the concluding remarks regarding the relevance of cross-lingual models in the NLP domain of AI.

Chapter 3

Zero-shot Transfer Learning Using Affix and Correlated Cross-lingual Embeddings

This chapter investigates two cross-lingual projection models: Canonical correlation analyses (CCA) [43] and VecMap [48] for creating cross-lingual embeddings and provides experimental analyses on four South African languages, namely, Sepedi, Sesotho, Setswana, and IsiXhosa. The experimental models use English as a high-resource language, thus all the projections use English as the source language. From the generated cross-lingual embeddings we shared insights by showing intrinsic evaluations of the embeddings using cosine similarities. Furthermore, we shared extrinsic evaluation insights on News headlines classification downstream task and Named Entity Recognition evaluation task. Interestingly, the overall study showed improved performance when training with cross-lingual embeddings compared to monolingual embeddings. Moreover, CCA performed better in all evaluation tasks compared to the popular VecMap model. The machine learning and deep learning models explored in this chapter included both single and hybrid combinations of Conditional Random Fields, Logistic Regression, XgBoost, Long Short Term Memory, Gated Recurrent Neural Networks, and Attention mechanisms. Although cross-lingual aided training showed empirical evidence of improved transfer performance, however, the significant and thought-provoking question – "Which properties, and characteristics are shared between languages, which aided to improved performance ?" remains unanswered and an open question.

3.1 Introduction

South African languages are not included in impactful and active NLP research. This is primarily due to the lack of digital data, standardised and annotated datasets, expertise focusing exclusively on local languages, and computing resources. As such, many of the South African languages can be categorized as left-behind or scrapping based on the system of classification outlined in [168]. As a result, most South African languages are classified as low-resource languages, and cross-lingual models are the strategies developed to remedy low-resource languages by leveraging resources from high-resource languages. These cross-lingual models aim to improve NLP research for underrepresented languages by transferring knowledge and resources from more widely studied languages. This approach helps bridge the gap between low-resource and high-resource languages in the field of natural language processing.

Cross-lingual models as exhaustively described in the previous chapter (Chapter 2) involve learning intra-semantic and inter-translation equivalent representations of words from two or more languages by projecting the embeddings space of one language into another or mapping both monolingual embeddings into a shared embedding space. That is, creating an entangled space where vector representations of similar words within a language preserve language properties (e.g semantics, syntax, temporal dependencies) and vector representations of translation pairs between observed languages are entangled (i.e. are trained to be similar) [2, 5, 26, 48, 106]. The application of these cross-lingual models has shown eminence in low-resourced downstream tasks such as information retrieval [197], entity linking [187], text mining [198], and natural language inference [199]. However, with over 2 decades of active NLP research of cross-lingual models, South African languages are the only visitors of this history of impactful research and application. Additionally, major drawbacks commonly raised in cross-lingual research such as that these models are often evaluated on language pairs where there are either available parallel and comparable corpora, availability of supervision signals such as bilingual lexicons, translation tools such as Google translate, and/or large enough raw non-aligned text corpus are not extensively investigated and contextualized for South African languages [95]. This means, still, languages with extreme scenarios unable to

meet these prerequisites such as South African languages continue to remain unalleviated from the threat of exclusion from language technology development progress.

From the aforementioned low-resourced standpoint of South African languages, there is a need to conduct thorough experiments using cross-lingual models. For this, we investigated two linear projection (see Chapter 2) techniques, namely, Canonical correlation analyses and VecMap for entangling monolingual embeddings spaces between four agglutinative South African languages – Sepedi, Sesotho, Setswana, and IsiXhosa. We considered English as the source language in all experiments. To evaluate the generated entangled spaces, we considered intrinsic and extrinsic approaches. The instinct evaluation we explored included cosine similarity measures of cross-lingual embeddings. A higher cosine similarity score between the observed vector representation of a word w_i^{L1} from language L_1 and a word w_k^{L2} from language L_2, where w_i^{L1} and w_k^{L2} are translation pairs, means the projection technique used was able to pull similar word closer together. Simultaneously, a lower score is expected for words that are not translation pairs. Intrinsic evaluation primarily extracts insights on the robustness and capability of the projection technique in creating entangled representations. Extrinsic evaluation studies considered included training classification models for two tasks, namely, News headlines classification and Named entity recognition. The objective is to observe if there are any variations from training with monolingual embeddings and cross-lingual embeddings. Better performance on cross-lingual embeddings would imply that the observed cross-lingual embeddings contain properties useful for training classifiers compared to their monolingual counterparts. In hindsight of this, it then means cross-lingual models can support low-resourced settings of the observed under-resourced languages. The unique contribution of this study to the cross-lingual and overall NLP community is three-fold:

1. The development and evaluation of entangled embedding spaces for Southern African languages: Sepedi, Sesotho, Setswana, and IsiXhosa.

2. We introduced a pre-training dataset for the 4 South African languages: we collected raw datasets from different open-sources for pre-training with the cross-lingual embedding model.

3. We presented a thorough experimental evaluation and comparison of the tradi-

tional machine learning models (Logistic Regression, XgBoost, and CRF) and deep learning models (LSTM, GRU, and Attention) on monolingual and cross-lingual representations.

3.2 Literature Review

The intervention curved by cross-lingual models championed recent progress in many NLP downstream tasks such as Part of Speech tagging, Named entity recognition, Document classification, sentiment analyses, etc, for low-resourced languages [94, 106, 200, 201]. These cross-lingual models proposed the novelty of exploiting vector representations learned with suitable conditions to supplement representations learned with insufficient resources. That is, vector representations such as GloVe [87], FastText [182], Word2Vec [83] from the English language, learned with millions of tokens (i.e. with considerable quantitative and qualitative resources) can be entangled with representations learned for a low resourced language (Urdu [202]) for knowledge transfer. The intuition is, with enough resources, we can effectively learn useful language properties such as semantics, word relations, e.t.c from a high-resourced language as demonstrated by many downstream tasks and transfer (using projection techniques) these properties to embedding spaces where there are not enough resources for training to capture them.

A continuum of works including low-resourced languages in active and impactful NLP research using cross-lingual models emerged as a consequence. For example, Rehman et al. [202] investigated Orthomap, VecMap supervised and unsupervised for projecting representations of English and the low-resourced language Urdu into the same cross-lingual embedding space. Their study extrinsically evaluated their entangled representations using topic classification, propaganda identification, and sentiment analyses and came to the conclusion that increased token coverage aided in optimized transfer performance. However, their study did not intrinsically evaluate their proposed cross-lingual embeddings.

Makgatho et al. [37] used an unsupervised VecMap projection model to create cross-lingual embeddings between Sepedi and Setswana and validated their performance on Word Similarity tasks. However, their work did not investigate extrinsic downstream

tasks and only covered 2 (out of 11) official South African languages. Niyongabo et al. [178] investigated two mutually intelligible languages, namely, Kinyarwanda, and Kirundi for creating cross-lingual embeddings and evaluating them on multi-class news classification. Ngomane et al. [38], created cross-lingual embeddings using VecMap between English and Isizulu for news headlines classification. Notably, their work only considered 1 of the 11 official languages, and no intrinsic evaluation was performed on the proposed embeddings.

Recently, sophisticated architectures such as mBERT [21] and XMLR [32] in the cross-lingual regime with implicit entanglements, are more prominent. For example, [28] used pre-trained masked language models to create implicit cross-lingual representations and evaluated them on Grammatical Error Correction (GER). Robnik-Sikonja et al. [179], performed a systematic comparison of sentence-based cross-lingual embeddings (Language-Agnostic sentence representations [120]) with embeddings extracted from BERT on sentiment analyses. Lauscher et al. [134], investigated cross-lingual embeddings for closely related languages and concluded by highlighting a correlation between transfer performance and language intelligibility. Nooralahzadeh et al. [136], supplemented pre-trained models with meta-learning and arrived to the same conclusion as Lauscher et al. [134]. Myoya et al. [39] contrasted the recently accepted multilingual models (AfriBerta, AfroXmlr, and AfroLM) for news classification. However, just one South African language (IsiZulu) is addressed in the study. Ifeoluwa Adelani et al. [13], Muhammad et al. [36] and the recent Dione et al. [14] works considered only three South African languages, namely, IsiXhosa, IsiZulu, and Setswana on NER, Sentiment Analyses, and POS downstream tasks respectively, using pre-trained multilingual models. Just recently, Lastrucci et al. [40], created a multilingual corpus covering all South African languages and benchmarked their dataset through MT using a multilingual M2M100 model. However, their crawled dataset may be susceptible to limitations imposed by their domain-specificity. Regardless, such anticipations remain open research questions. Clearly, more work needs to be done for South African low-resourced languages with varying states of low-resourcedness. Additionally, a comparative analysis of available projection techniques on South African languages proves to be important as 1) most works consider one projection technique, and 2) most available works do not

evaluate both intrinsically and extrinsically for more informative insights on the transfer performance. To partially address these concerns, in this study, we explore two projection techniques: CCA, and VecMap for creating entangled representations for four South African languages: Sepedi, Sesotho, Setswana, and IsiXhosa, with English as the source language. Additionally, we evaluate both intrinsically(using cosine similarities) and extrinsically (using NHC and NER), the resulting cross-lingual embeddings. A detailed description of our approach is given in the following Methodology section below.

3.3 Methodology

Figure 3.1 describes the end-to-end approach adopted in this study for creating and evaluating cross-lingual embeddings. The initial step includes collecting monolingual data for the four considered languages resulting in a monolingual repository of size 742.7 MB (details of source described in table 3.1). The data collection step is followed by a preprocessing step that utilizes language identification using bi-gram models. This step is proceeded by the creation of monolingual embeddings using the FastText API. These monolingual embeddings are then used to create cross-lingual embeddings using CCA or VecMap projection techniques. And finally, the evaluation of cross-lingual embeddings using News Headlines Classification and Named Entity Recognition downstream task concludes the methodology.

3.3.1 Data Collection

Table 3.1, describes the data collected and sources for four languages: Sepedi, Sesotho, Setswana, and IsiXhosa. This study only considered these 4 languages because the work aimed to build on the work of [37], and also these datasets were the only dataset we had access to or could curate online. Our collection pipeline does not include any ethical considerations or declarations as the datasets used in this study are from publicly available resources such as news websites or government repositories. The next, subsection 3.3.2 details the preprocessing steps taken for data sanitation.

[1] https://drive.google.com/drive/folders/1jYwpxEdRxqXlB7BSmE6JxDar61U91xfI

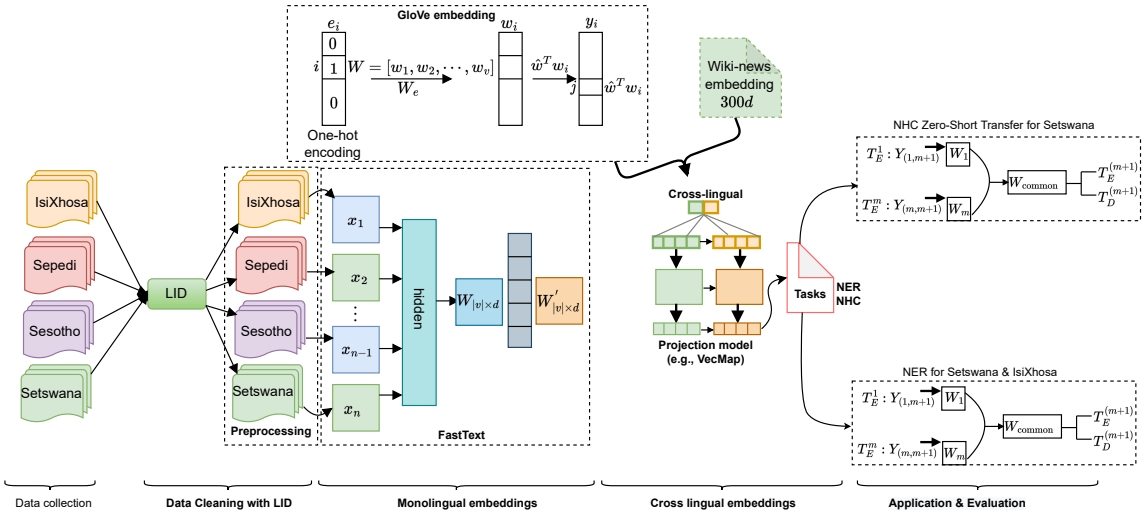


Figure 3.1: End-to-end approach for creating and evaluating cross-lingual embeddings

Table 3.1: This table shows dataset sources used for developing monolingual embeddings. The total number of sentences is shown for each corpora.

Sources	IsiXhosa	Sepedi	Sesotho	Setswana
African webcrawl datasets [203]	✓	✗	✗	✗
government proceedings[1]	✓	✓	✓	✓
WMT2022 [204]	✓	✓	✗	✓
NCHLT [205]	✓	✓	✓	✓
Common Crawl 100(CC100) [206, 206]	✗	✗	✗	✓
MC4 [207]	✓	✗	✓	✗
Opus [208]	✓	✗	✗	✗
Xtreme [209]	✓	✗	✗	✗
Sotho Webcrawl 2015, 2017, 2018 [210]	✗	✗	✓	✗
Sepedi community 2017 , 2018, 2020, and 2021 [210]	✗	✓	✗	✗
Flores [211]	✗	✓	✓	✗
Total	1.9m	693k	1.5m	1.06m

3.3.2 Primitive Language Identification

Manual inspections of the collected dataset demonstrated that datasets coming from sources such as WMT2022, MC4, etc. contained foreign tokens within sentences. To remedy this, we used preliminary bi-gram models trained with the NCHLT dataset for language identification. The NCHLT dataset is a trusted source and we used it to develop two bi-gram models: Sepedi bi-gram and IsiXhosa bi-gram model. The Sepedi bi-gram

model is used to filter Sepedi, Sesotho, and Setswana as these languages are mutually intelligible languages and belong to the same language group. And, IsiXhosa bi-gram model was used to filter the IsiXhosa corpus. In total, we remained with approximately 83K sentences for Setswana, 102K sentences for IsiXhosa, 60K for Sepedi, and 126K for Sesotho. The next subsection 3.3.3 explains how this corpus was used to generate continuous representations of words for each language using the FastText[182] model that exploits subwords for creating continuous representations.

3.3.3 Generating Monolingual Embeddings

Our experiments explored FastText embeddings of dimension d={200,300}. That is, our monolingual embeddings were created using the FaceBook FastText model [182] with parameters: the minimum and maximum character n-gram set to 2 and 5 respectively, the model was set to iterate through examples 10 times (epoch=10, chosen due to resource constraints - where manual inspection showed improved word similarity search as epochs are increased, whereas, iterations closer to 10 increased training time), and the remaining parameters were set to default. These parameter choices were set for all four languages. This study used English as the pivot or source language. As such, we did not have to train English monolingual embeddings from scratch, rather, we used available and popular embeddings such as Glove [87], and FastText Wikipedia embeddings [212]. The next subsection 3.3.4 discusses how these monolingual embeddings are projected into the same embeddings spaces.

3.3.4 Generating Entangled Embeddings

To create entangled representations with intra-semantic and inter-translation equivalences properties intact, we used VecMap and Canonical correlation analyses projection models. VecMap and CCA are explained in detail in Chapter 2, Subsection 2.3.1. For extracting representations with CCA, we used default parameters as suggested in Faruqui and Dyer [43]. Additionally, CCA requires some form of supervision signals, in this case, bilingual lexicons, and therefore we used lexicon pairs provided in [37] from their Simlex Tasks. IsiXhosa was not included in the study conducted by [37], and therefore

we had to collect our own lexicon pairs from official government websites[2]. VecMap on the other hand used semi-supervised training, self-learning, vector normalization set to {unit, centeremb center} to create the entangled space between English and the individual four languages. Semi-supervised training of VecMap utilized the same lexicon pairs as CCA. Collectively, we generated a total of 24 cross-lingual embedding sets (i.e. 6 for each language).

Only Setswana extracted cross-lingual embedding from CCA and VecMap are used to perform zero short transfer learning on the task of NHC due to NHC being available for Setswana only and other embeddings are used to train a NER model from scratch for the task of NER. The cross-embeddings are further compared with monolingual embeddings (as baselines) on the task of NER.

3.4 Intrinsic and Extrinsic Evaluation on Downstream Tasks

This section discusses the models, model parameters, evaluation metrics, and the use or non-use of annotated data for downstream stream tasks. We discuss, NHC downstream task and NER downstream task.

3.4.1 NHC Model

The considered task for this section is a classification task proposed by Marivate et al. [213] as a proof of concept that considers using a model to render the category of news headline extracted from a news article of a low-resourced language. That is, the task is to attribute a news headline NH_i^{lr} to belonging to one of the classes $C = \{politics, entertainment, sport, art-and-culture, e.t.c\}$. To evaluate cross-lingual embeddings, our study considered zero-shot transfer learning. Zero-shot transfer learning is a machine learning technique where a model is trained in one domain and it is used directly as is (i.e. without further training or adaption) in another domain. In our case, we trained a cohort of models (Logistic regression, XGBoost, Feed-Forward Neural

[2] za-natural-science-and-technology-term-list

Netowork) on the English data (BBCA [3]) and performed zero-shot on all of them on the Setswana NHC datasets [213]. Our study model is a Feed-forward Neural Network (FFNN) model with parameters selected using keras-tuner API[4]). The final set of hyper-parameters is reported in Table 3.2. Table 3.3 shows the constraints imposed on the search space for optimal parameters for NHC. All other parameters such as the training epochs and batch size, were set to 120 and 64 respectively based on cross-validation.

Table 3.2: Hyper-parameters for NHC Model.

Embeddings	Model	spatial dropout	Hidden layer size	dropout rate	learning rate
CCA 300	FFNN	0.1	16	0.5	0.0001
VecMap 300	FFNN	0.3	12	0.1	0.01

Table 3.3: Searching optimal parameters for NHC.

Parameter	Restriction
Spartial dropout	sp dr ∈ [0.1 , 0.5]
Number of hidden layers	1
Number of neurons	n ∈ [8 , 16]
Dropout rate	dr ∈ [0.1 , 0.5]
Learning rate	lr ∈ { 0.01, 0.001, 0.0001 }
maximum trials	5
Optimizer	Adam optimizer
Loss	Categorical cross-entropy

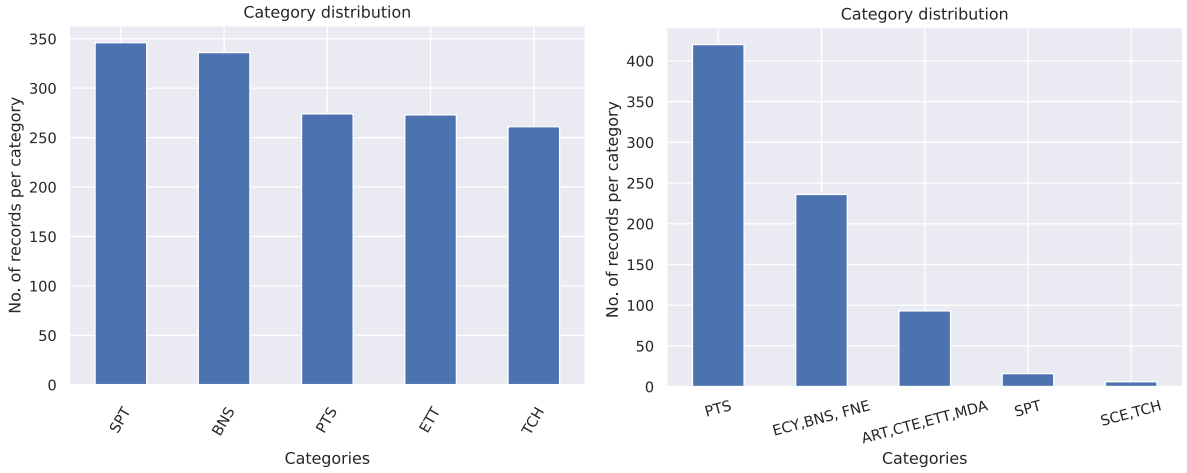
3.4.1.1 Zero-shot Transfer Learning

The BBCA English dataset was used as the training dataset and was sourced from Kaggle. This dataset contains over 17M tokens with sentences categorized into 5 classes (see Figure 2a for distribution of classes). For preprocessing the BBC dataset we removed URLs and punctuations using the python regular expressions module. For stemming and lemmatization, we used the python NLTK toolkit. For testing, we used the Setswana NHC dataset from Marivate et al. [213]. This dataset was preprocessed in the same manner as the training BBCA dataset. From the original dataset from Marivate et al. [213], we adapted it to match the training BBCA dataset for ease of inference (see Figure

[3] kaggle bcc dataset

[4] kerastuner

3.2). The acronyms in Figure 3.2 are denoted as SPT = Sport, BNS = Business, PTS = Polotics, ETT = Entertainment, TCH = Technology, ECY = Economy, FNE = Finance, ART = Art, CTE = Culture, MDA = Media, and SCE = Science.



(a) Distribution of classes on BBC datasets (b) Dataset after class reduction [213]

Figure 3.2: Category distribution of Engish(left) and Setswana(right).

3.4.1.2 NHC Model Evaluation

We will evaluate our NHC model on the Setswana dataset from Marivate et al. [213], as our test data, using accuracy(the number of correctly classified classes over a total number of examples) and weighted F1-score for a fair comparison with [213].

3.4.2 NER Model

In this section, we discuss a variety of encoder models for learning latent representation useful for downstream tasks. The encoders include classical deep learning models such as Bi-directional Long Short Term Memory (BLSTM), Gated Recurrent Units (GRU), and Attention mechanisms, and traditional machine learning models such as Conditional Random Field. Additionally, various combinations of these models are explored. All the hyper-parameter choices of the models were selected using Keras-tuner hyper-parameter search algorithm. Hyper-parameters for all models are reported in Table 3.4. This

table highlights all the optimal parameters found using monolingual training. The exact parameter setting was used to train models with cross-lingual embeddings for fair and direct comparison.Similalarly, the CRF model was trained with monolingual embeddings and the same model with initialised with the same parameters is trained with cross-lingual embeddings [214]. Following Adelani et al. [164], ADAMW [215] was used as the training optimizer, and due to limited training examples, we trained using k-fold cross-validation with k-10. As a baseline of CRF, we also investigated Feature Engineering (FE) as input to CRF. These features include subwords, word length, stemming, the beginning of sentence identifiers, end of sentence identifiers (EOS) e.c.t. We report the results of our models in the NER results in section 3.5. The following section discusses the train and test data distribution.

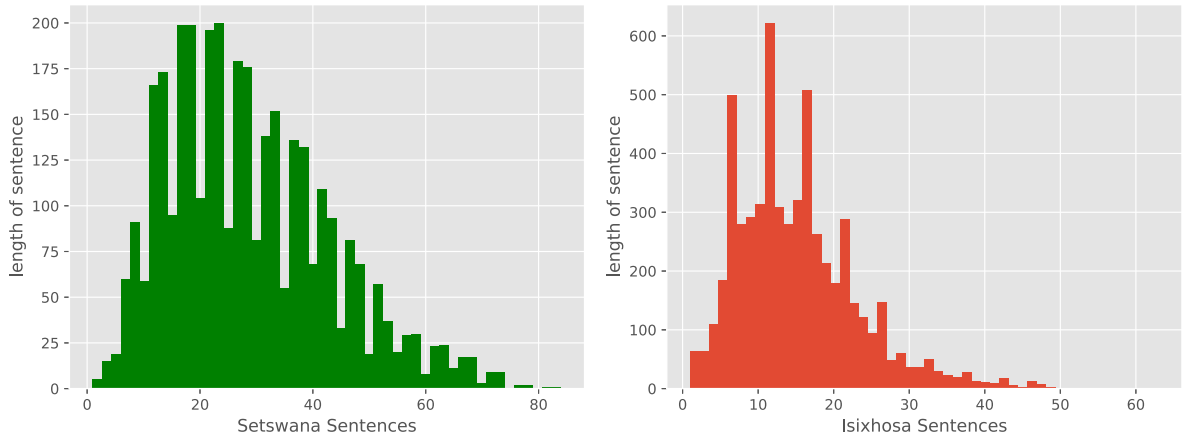
Table 3.4: Hyper-parameters for NEC[a] [9]

Embeddings	sp dr	Model	Units	FFNN	lr	weight decay	dr
Setswana FastText 300	0.2	BiLSTM+Attention	150	352	0.01	0.0001	0.1
	0.2	BiGRU + Attention	125	64	0.01	0.001	0.1
	0.1	BiLSTM+BiGRU+Attention	25	32	0.01	0.01	0.1
IsiXhosa FastText 300	0.5	BiLSTM+Attention	100	224	0.01	0.001	0.4
	0.5	BiGRU + Attention	100	128	0.01	0.001	0.4
	0.3	BiLSTM+BiGRU+Attention	50	224	0.01	0.001	0.3

[a] Named Entity Classifiers

3.4.2.1 NER with Training and Test Dataset

MasakhaNER2.0 is the most recent and standardized Named Entity Recognition dataset. NER is a text classification task that involves grounding tokens or phrases within a sentence based on the entity they belong to, such as a person, organization, location, e.t.c. From this dataset, we extracted the IsiXhosa and Setswana datasets as they are the only two datasets available for embeddings considered in this study. The Setswana data is divided into 3489, 499, and 996 training, development, and test data respectively. There are 9148, 2812, and 3489 unique tokens in the training, development, and testing data respectively. Similarly, the IsiXhosa data is divided into 5718, 817, and 1633 training, development, and testing data respectively. The training, and development data contains 26221, and 6223 unique tokens. Figure 3.3 shows the sentence-length distributions of the languages: Sestwana (left) and IsiXhosa (right). The next section discussed the evaluation metric for NER dataset.



(a) Setswana (b) IsiXhosa

Figure 3.3: Distribution of words and sentences for NER datasets.

3.4.2.2 NER Model Evaluation

Our NER models will be evaluated using Precision, Recall, and F1-scores on the train, as well as the test data following Adelani et al. [164]. The next section discusses the results of training the aforementioned models with monolingual and cross-lingual embeddings as well as zero-shot transfer learning. Additionally, we provide results on the intrinsic evaluation results of cross-lingual embeddings.

3.5 Results

This section discusses the main results collected from NHC and NER downstream tasks using monolingual and cross-lingual embeddings. This is presented in two-fold. The first subsection discusses zero-shot transfer learning results for NHC task. The second subsection presents results for NER downstream task.

3.5.1 NHC model results

In this subsection, we discuss the main results obtained for each projection technique: CCA and VecMap.

3.5.1.1 Results and Analyses for CCA Embedding

The main results obtained using CCA entangled embeddings are reported in Table 3.5. From this, we observe that using the best hyper-parameters obtained using keras-tuner (see Table 3.2) the highest average accuracy achieved for zero-shot transfer is 54.5% from CCA embeddings of dimension $d = \{300\}$.

These embeddings were obtained from GloVe pre-trained English monolingual embeddings and FastText Setswana monolingual embeddings. The remaining two cross-lingual embeddings between English and Setswana yielded poor results (see figure 3.4), which shows the FFNN accuracy using CCA cross-lingual embeddings of 200 dimensions created with Glove and FastText monolingual embeddings (CCA300 G & F), 200 dimensions created with Glove and FastText monolingual embeddings (CCA 200G & F), and 300 dimensions created with FastText and FassText monolingual embeddings. Figure 3.5 presents a confusion matrix of these high-performing CCA embeddings as a means of additional analyses on their performance. We observed that the majority of the models' incorrect predictions happened for class "Science and technology", class "Sport", and class "arts, culture, entertainment and Media". This means the models were unable to learn robust representations given the low number of examples used for training with the BBC dataset. We hypothesize that this performance can then be attributed to the small training data as well as partly class imbalance of the training and test datasets. Class "art, culture, entertainment, and media" 'spoor prediction on the other hand can be argued to be caused by poor representation inclusion. This is because, in the training data, the associated class ("entertainment") only includes entertainment content, while in the test data, we have other content coming from arts, culture, and media. For this, the two contained varying contents (See Figure 3.6 of side-by-side word plots of the two classes from the two datasets). As shown in Figure 3.6, the left-hand side (LHS) represents the word cloud of all train sentences with "Entertainment" as a label from the English datasets, while the right-hand side (RHS) represents the word cloud of all train sentences with "Art, culture, entertainment, and media" as a label from the Setswana dataset. Figure 3.6 shows that the entertainment content in our available English dataset differs significantly from that in our Setswana dataset, negatively influencing our transfer performance. This exemplifies the "comparable corpora" issue we mentioned earlier in the introduction (Section 3.1). However, these observations need further investigations with better conditions (e.g data, models, etc). Regardless, the model was able to perform at an average accuracy of over 54% from the available classes in the test dataset. This is promising and impressive given the nature of the task, the limited resources used in this work, and the mere simplicity of our model.

Comparatively, our model (i.e., FeedForward Neural Network) was able to perform comparably well to results obtained in Marivate et al. [213] in terms of weighted F1-score (see Figure 3.5). Figure 3.5 shows the prediction accuracy of the zero-shot models trained with CCA and VecMap embeddings on the NHC dataset. Both figures suggest that the models are more effective at transferring political data than other categories. This requires further investigation into why this is the case. Our advantage is that our model has not been trained on the target dataset and is still competitive with the benchmark. For this, we believe that with better training conditions (e.g., more bilingual lexicons, a robust cross-lingual model, aligned train and test datasets, etc.), there is room to surpass the performance of traditional ML models on downstream tasks with transfer models aided by cross-lingual embeddings.

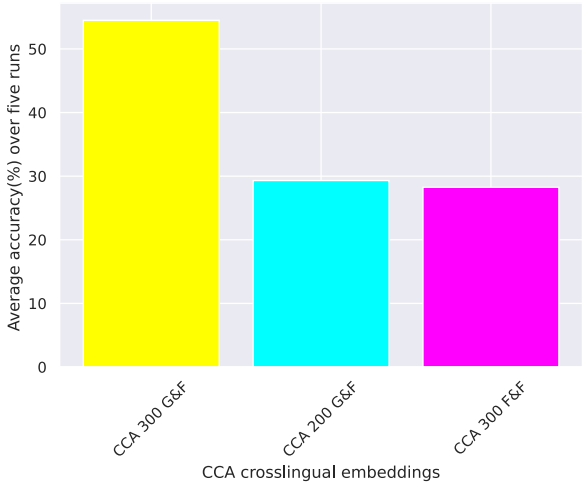
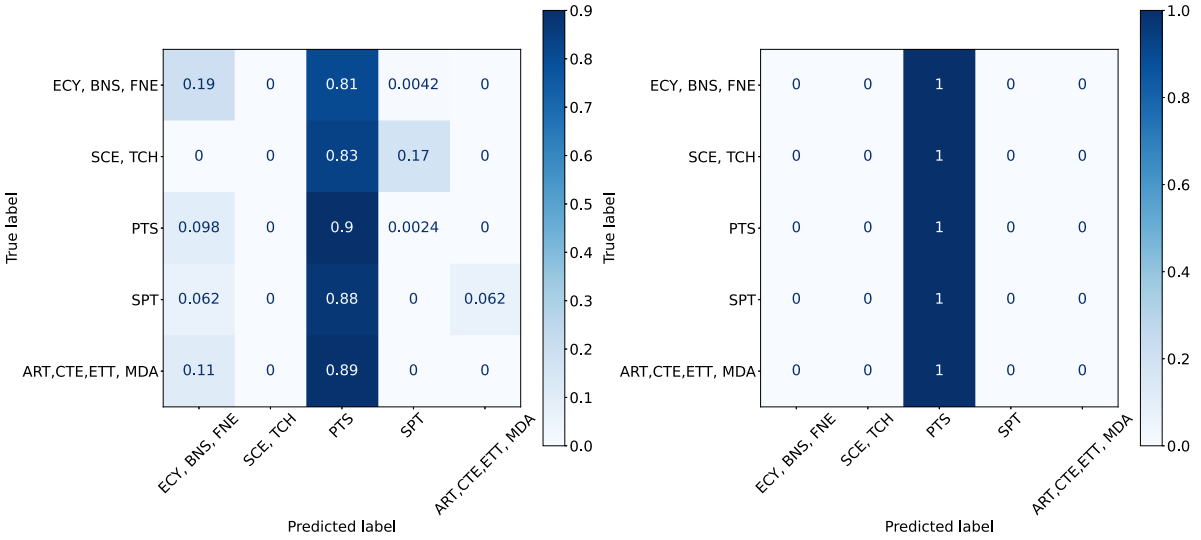


Figure 3.4: FFNN accuracy using CCA cross-lingual embeddings.

3.5.1.2 Results and Analyses for VecMap Embeddings

Similar to CCA cross-lingual embeddings, the final hyper-parameters for the VecMap model are reported in Table 3.2. With entangled representations created using VecMap, the average accuracy achieved on the NHC zero-shot transfer was 54.48%. The reported accuracy was achieved using entangled embeddings of dimension $d = \{300\}$, normalization set to {unit, centeremb, center} in that order. These entangled representations were created from Glove English monolingual embeddings and Setswana FastText monolingual embeddings. The remaining 200d (from Glove and FastText) and 300d (from Glove and FastText) produced unsatisfactory results for all experimented parameters and were left out for the remainder of this study. Additional inspection of predictions using the high-performing embedding space shows similar trends as the CCA embeddings: 1) comparable corpora affected model prediction capacity; 2) more training data is needed for better representation learning (see Figure 3.5b), etc. However, VecMap-borne representation did not learn a representative input feature space since the model obtained with them was overfitting to one category of the available classes: politics.

Our zero-shot results indicate that the CCA projection technique is able to project representations of English and Setswana robustly compared to the VecMap projection technique. In support of this, we explored the intrinsic evaluation of the projected representation using cosine similarity scores. That is For this, we explore a few visualizations of the best cross-lingual embeddings. We observe that CCA was able to extract representations that are more similar compared to VecMap, which then may have contributed positively to transfer performance (see Figure 3.7) illustrating the cosine similarity between English and Setswana words from the best performing VecMap (Figure 3.7a) and CCA (Figure 3.7b) embeddings. Figure 3.7) clearly shows the cosine similarity of words projected into the same cross-lingual embedding space. Nine words from Setswana and their corresponding English translations were



(a) CCA (b) VecMap

Figure 3.5: Confusion matrices for CCA and VecMap using zero-short test data from Setswana.

randomly selected, and their embeddings were retrieved from VecMap and CCA embedding spaces. The cosine similarity of the retrieved embeddings is measured and plotted in the matrix above for VecMap and CCA embeddings. Clearly, CCA produces more similar representations (seen by high scores on the diagonals of the matrices), which may have contributed positively to transfer performance. Furthermore, the confusion matrices (Figure 3.5) indicate that CCA embeddings can generalize better compared to VecMap embeddings. This, however, needs further analysis and is left for future work with favorable conditions. In the next subsection, we discuss our findings on the NER downstream task.

3.5.2 Results and Analyses on the NER Model

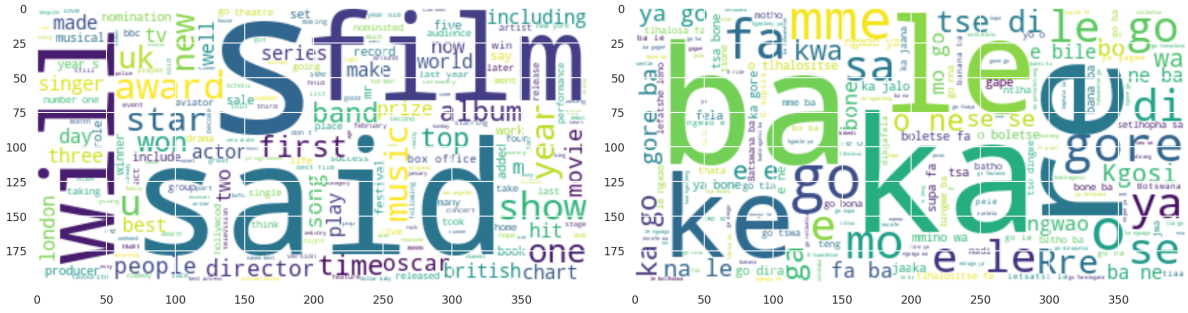
This section presents NER results for monolingual and cross-lingual embeddings using deep learning and machine learning models - LSTM, GRU, Attention, and CRF. For Setswana NER, we only considered best-performing cross-lingual embeddings in section 3.4.1.2 (see Figure 3.5a and 3.5b). That is, our experiments are based on the entangled representation of dimension $d = \{300\}$ projected using CCA technique from Glove English monolingual and Setswana monolingual embeddings. For IsiXhosa NER, we investigated entangled representations of the same dimension created using the same projection technique. Setswana and IsiXhosa monolinguals serve as our baseline models and are compared with their cross-lingual counterparts. Table 3.6, presents the main NER results of this study using precision (P), recall (R) and F1-score (F1) as the evalution matrics. Our experimental results show that training with cross-lingual embeddings consistently performs better compared to their monolingual counterpart.

Table 3.5: Zero-shot models were trained on the NHC dataset with CCA and VecMap embeddings.

Embeddings	Model	Accuracy	Fl-score (weighted)
Marivate et al. [213]	XGBoost/LR	-	± 60%
CCA 300	FFNN	54.5%	46.05%
	XGBoost	54.08%	38%
	LR	54.73%	39%
VecMap 300	FFNN	54.48	38.42
	XGBoost	-	-
	LR	54.47%	38 %

This means, the CCA project technique shares properties between the languages that are useful for the NER downstream task. Our best-performing machine learning model is the CRF model with an average F1-score, precision, and recall of 96.3%, 96.6%, and 96.4 respectively for Setswana. With CRF models, we observed an increase in performance when using cross-lingual embeddings and better performance gains when using both feature engineering and cross-lingual embeddings compared to other feature representations (e.g., only FE or FE + monolingual embeddings). For this, we observed an improvement of 0.35%, 0.36%, and 0.44% on precision, recall , and F1-score respectively on average compared to other models. The best CRF model used FE combined with cross-lingual embeddings (see Table 3.6). We observe contradicting results on IsiXhosa experiments. That is, while the SKlearn CRF with FE and monolingual embeddings showed the best performance, the cross-lingual embeddings with FE were on par with the best model.

For other latent representation learning models such as GRU, BiLSTM, and Attention, cross-lingual embeddings either out-perform or are on par with monolingual embeddings in terms of precision, recall and F1-score gains (see Table 3.6). Our stacked GRU with Attention temporal representation learning model is our second best-performing model achieving 83%, 64%, and 72% on precision, recall, and F1-scores respectively for Setswana. For IsiXhosa, the performance in terms of recall, precision, and F1-score of BiLSTM and BiGRU were on par. TF in Table 3.6 below indicate transfer learning with Few-shot learning. Few-short learning differs from the explained zero-shot learning in that, before testing in a new domain, a few examples are used to adapt the transfer model. Intuitively, this is supposed to have better performance. TF model in Table 3.6 is trained on English NER data from [164] using cross-lingual embeddings and transferring the trained model to target (Setswana and IsiXhosa) test data. The NER development data in [164] was used as the few-shot examples. We fine-tuned the parent model for 10 epochs. As expected, TF obtained better results compared to zero-shot transfers on the test data with the results for Setswana and we only report Few-shot results as zero-shot results are extremely low. Even though our results are not state-of-the-art, they still show that cross-lingual embeddings consistently add value to a variety of language encoders for NER downstream tasks. This means the same results could possibly be observed in other low-resourced downstream tasks such as POS tagging or related sequence labeling tasks.



(a) English word cloud: Entertainment category

(b) Setswana word cloud: Entertainment, art, media

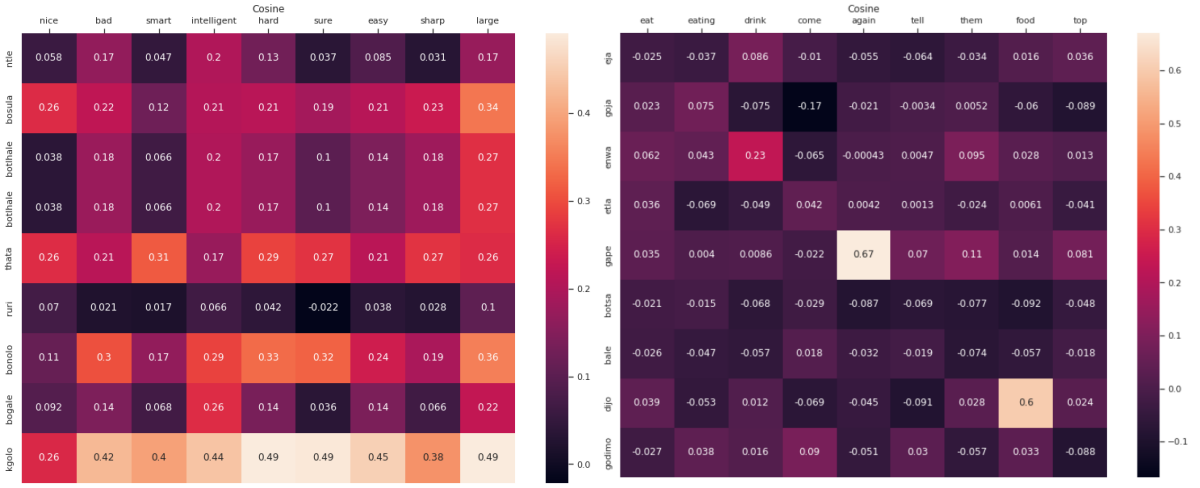
Figure 3.6: Word Clouds on Setswana dataset.

3.5.3 Conclusion

We collected and primitively cleaned datasets using Bi-gram LID for four under-resourced languages of South Africa: Setswana, Sesotho, Sepedi, and Isixhosa. Additionally, we created and evaluated both monolingual and cross-lingual embeddings for these languages on two downstream tasks: News Headlines Classification (zero-shot evaluation) and Named Entity Recognition (traditional train, validate, and test pipeline evaluation). For zero-shot NHC, cross-lingual embeddings created using CCA outperformed embedding created using VecMap evaluated using accuracy. We believe a further increase in zero-short performance can be realized by increasing bilingual lexicons used as signals for pooling languages together, which is cheaper to collect as opposed to other language signals or resources. This capability has great potential because classification, prediction, recommendation, and translation models can be created for low-resourced languages without training data as this is currently an albatross of NLP progress. The NER models also show promising results for entity recognition using cross-lingual embeddings. This advocates the use of supplemented embeddings in scenarios where training data is available and asserts the need to explore and investigate cross-lingual embeddings as a tool to support language learning.

3.5.4 Future Works

Future prospects that may lead to a potential increase in performance gains for extracting better morphologically supplemented embedding space include increasing bilingual lexicons used as signals for embedding projection. Increasing raw data and using more robust language identification models for filtering unclean tokens can also aid better performance for future works. Cleaning data at the token level and increasing the English or source training data to learn more robust models is another perspective for future work. A deeper analysis of transfer learning amongst all South African languages



(a) VecMap embeddings cosine similarities. (b) CCA embeddings cosine similarities.

Figure 3.7: VecMap and CCA embedding on Setswana dataset.

and their language families to capitalize on their relatedness can also improve future works. Developing inter-language-specific projection models leveraging on known relatedness between languages can also aid better performance for future works. Lastly, comparing and contrasting different traditional models with modern machine learning classification algorithms(e.g. deep neural nets) for zero-shot transfer on multiple downstream tasks can be useful to uncover baselines and insights for processing these South African languages. Additionally, it would be interesting to observe to what limit can these embedding projection methods supplement languages with respect to morphology and how this is scalable to other tasks such as machine translation, sentiment analyses, etc. Furthermore, it would be interesting to compare cross-lingual models with large multilingual models and gain an understanding of what advantages can be accumulated from each of the two modes of learning joint representations and for which task of South African languages a particular mode is most suitable to adopt.

3.6 Summary

This section summarizes Chapter 3 on using entangled representations from two projection techniques, namely, canonical correlation analysis [43] and VecMap [48] for the downstream task. The downstream tasks include News Headlines Classification and Named Entity Recognition. This chapter is divided into the Introduction 3.1, Related works 3.2, the Methodology 3.3, the Results 3.5, the Conclusion 3.5.3 and Future works 3.5.4 succinctly summarized in the following subsections 3.6.1, 3.6.2, 3.6.3, 3.6.4, and 3.6.5 respectively.

Table 3.6: Monolingual and cross-lingual embeddings for NER token classification.

Embeddings	Model	Decoder	Setswana			IsiXhosa		
			P	R	F1	P	R	F1
FE	Sklearn CRF	-	96.1%	96.4%	96.1%	94.1%	94.1%	94%
FText Mono	Sklearn CRF	-	95.6%	95.8%	95.5%	89.5%	89.8%	89.4%
FE+FText Mono	Sklearn CRF	-	96.2%	96.5%	96.2%	94.2%	94.2%	94.2%
CCA300 CL	Sklearn CRF	-	95.6%	95.9%	95.6%	89.9%	90.2%	89.3%
FE+CCA300 CL	**Sklearn CRF**	-	96.3%	**96.6%**	96.4%	94.2%	94.2%	94.1%
	LSTM+ATN	Softmax	78%	64%	70%	67%	48%	55%
	LSTM+ATN	CRF	41%	47%	43%	44%	41%	41%
FText Mono	GRU+ATN	Softmax	78%	63%	69%	66%	48%	54%
	LSTM+GRU+ATN	Softmax	72%	61%	65%	67%	48%	55%
	LSTM+ATN	Softmax	78%	65%	70%	67%	48%	55%
	LSTM+ATN	CRF	-	-	-	44%	41%	41%
CCA300 CL	**GRU+ATN**	**Softmax**	**83%**	**64%**	**72%**	66%	48%	55%
	GRU+ATN	CRF	48%	48%	48%	-	-	-
	LSTM+GRU+ATN	Softmax	77%	61%	68%	65%	46%	53%
TF (Few short)	LSTM+ATN	Softmax	69%	49%	57%	48%	13%	20%
	GRU+ATN	Softmax	66%	50%	56%	46%	13%	20%

3.6.1 Introduction

This section introduces the reader to the field of cross-lingual Natural Language Processing in the context of South African languages. Additionally, we highlight some of the shortcomings of cross-lingual models as well as the current remedies in the field. Finally, in this section, we highlight our unique contribution and value-add to the research community as well as developing resources for South African low-resourced languages.

3.6.2 Related Works

This section discusses related works on cross-lingual works for South African languages and the general landscape of low-resourced languages. In this section, we discuss how our work supplements the related work and builds on the existing research foundations.

3.6.3 Methodology: Intrinsic and Extrinsic Evaluation of Downstream Tasks (NHC and NER)

This section outlines the methodology adopted in this study for collecting data, preprocessing the data, projection model used, algorithms used including machine learning and deep learning models, hyperparameter choices, evaluation decisions, and downstream tasks considered.

3.6.4 Results

This section discusses the analyses of experimental findings and results for each projection technique (CCA and VecMap) on each downstream task (NHC and NER). In this section, our experimental results support the use of cross-lingual embeddings in elevating the state-of-the-art in extremely low-resourced settings. That is, through downstream evaluation, we observe that cross-lingual embedding consistently improves performance on NER and illustrates promising results on NHC zero-transfer learning.

3.6.5 Conclusion and Future Works

The Concluding remarks and future directions for cross-lingual models for South African languages is provided in this section. This section highlights a potential research avenue that uses cheap resources such as bilingual lexicons for improving projection techniques in creating cross-lingual embeddings. Our next Chapter explores this research avenue extensively.

Chapter 4

Point of Pivot: Calibration of Cross-lingual Embeddings for Southern Nguni and Niger-Congo Low-Resourced Languages

Analytics for transfer learning on low-resourced languages are still grounded by the English-to-X outset even though recent research shows that English is not always the best pivot language. Causality can be traced back to accessibility, availability, trend, and most importantly expertise of Natural Language Processing resources in the English language. However, with some constraints gradually being loosened, such as the availability of raw internet corpora, and annotated data, the need to explore other pivot alternatives becomes indisputable despite linguistic insights on language intelligibility being sometimes available. In this study, we extensively investigate point-of-pivot using cross-lingual embeddings for all permutations of Southern Nguni and Niger-Congo languages of South Africa. We explore three different techniques for generating cross-lingual embeddings, namely, VecMap, Muse, and Canonical correlation analyses (CCA) on intrinsic and extrinsic downstream tasks. We evaluate our approach across multiple tasks: Word similarity, Machine Translation, News Headlines Classification (NHC), Part of Speech Tagging (POS), and Named Entity Recognition (NER) and discovered unexpected best transfer source languages (e.g Xitsonga) across multiple tasks, thus supporting the need for a better source selection (i.e point-of-pivot). We further showcase the efficacy of the extension of linguistic-based tasks such as POS and NER and text classification tasks such as NHC to other South African languages through Machine Translation, with encouraging transfer performance on ground truth datasets (E.g NER – (98.55, 98.77), and POS – (91.21, 91.63), and NC – (59.09,60.32) average accuracy, on IsiZulu and IsiXhosa available gold test sets respectively), and therefore supporting the similarity of the generated data and the ground truth data.

4.1 Introduction

The Constitution of the Republic of South Africa recognizes 12 official languages with eleven spoken languages [216] and one sign language [217]. Some share origins while others are typologically diverse.

Figure 4.1 shows the distribution of the different languages according to first (L1) and second (L2) language speakers. Where L1 refers to the speaker's first language, and L2 is the second additional language. Many of the South African languages have a small digital footprint which hinders the development of

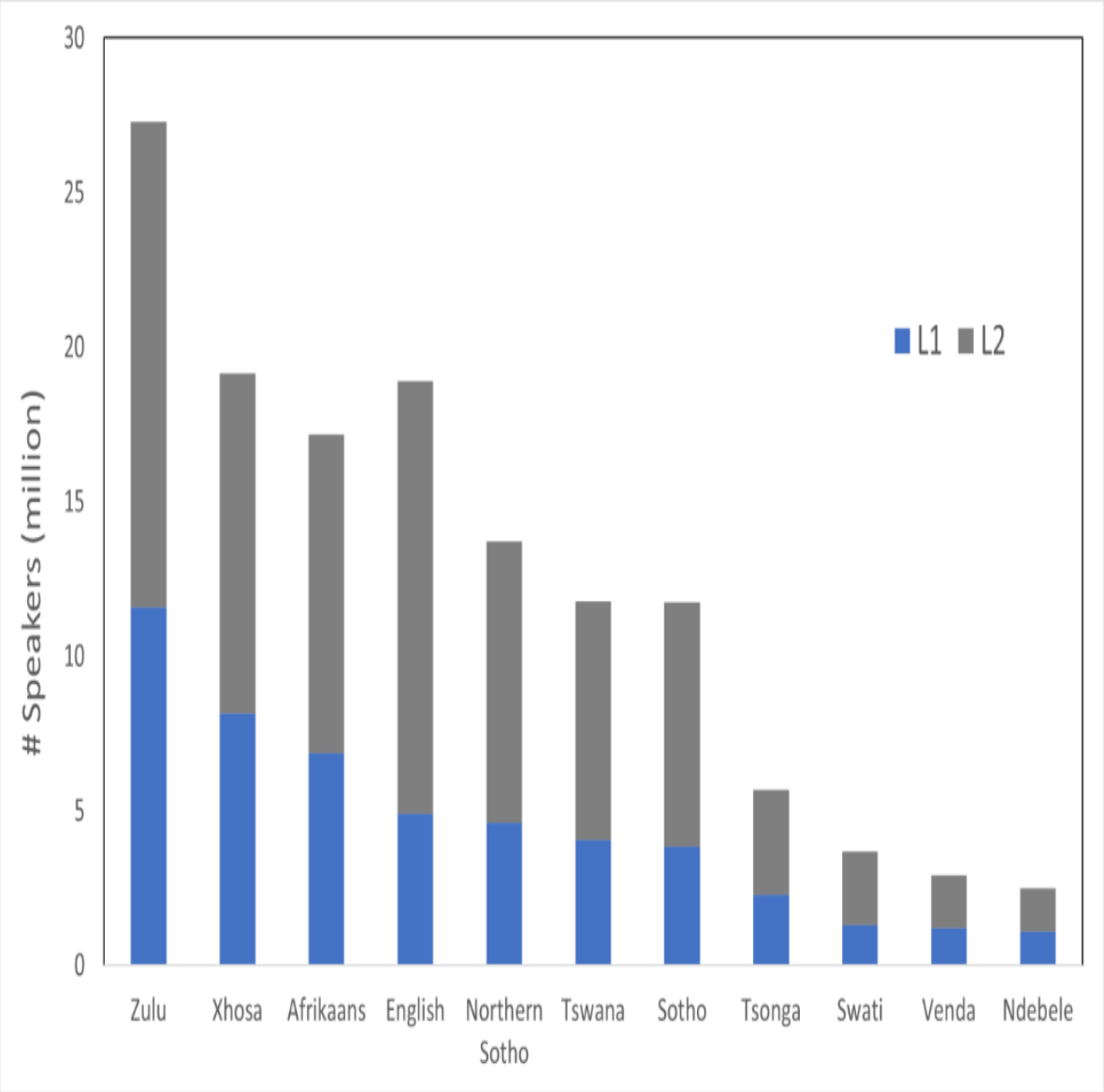


Figure 4.1: Number of first and Second additional language speakers for South African languages [7]

language technologies. Recent remedial initiatives focus on cross-lingual transfer learning. This process leverages advancements from high-resourced languages to learn language representations that can sup-

plement low-resourced language representations [2]. Languages such as English, French, and Spanish are often used as pivot languages for learning distributed representations [5, 26, 44, 92, 106, 123, 126]. However, due to typological diversity and other linguistic disparities, these high-resourced languages have not always proven effective as pivot languages [14, 218]. The inability to capture linguistic idiosyncrasies translated to poor performance in downstream tasks. From this observation, the question of point-of-pivot continues to persist in cross-lingual NLP research [166].

As fundamentals supporting both theoretical and practical soundness of cross-lingual models continue to evolve and advance, the accompanying insights regarding the best choice of source language lag behind [167]. In some cases, mutual intelligibility from a linguistics lens is often grounds for point-of-pivot [178]. In most cases, the availability of resources, advancement of technologies, and a large community of experts worldwide form the basis for point-of-pivot (i.e. the English language). English-pivoted cross-lingual representations have been evaluated on part-of-speech tagging [106], dependency parsing [25], entity linking [187], sentiment analysis [77] and speech recognition [219]. In all the aforementioned works, it is observed that the choice of point-of-pivot was based on the high availability of English resources.

Recent advanced models such as the Large Language Models have also ascribed to the narrative of English (and other high-resourced languages) as the point-of-pivot. For example, de Vries and Nissim [138] trained a generative language model on English as a means to initialize the parameters of the models and later adapted the model to low-resourced languages. Gogoulou et al. [79] applied the approach proposed in de Vries and Nissim [138] on a BERT model and also used English as point-of-pivot. Contrary to these, the work of Gogoulou et al. [79] argues that there is no correlation between transfer performance and language similarity.

For South African languages, Figure 4.2 shows all known language families[1], and therefore the languages expected to transfer well in theory. Makgatho et al. [37], explored cross-lingual embeddings within the same language family (Setswana and Sepedi) on a Word Similarity downstream task. However, this assumption of using language similarity as grounds for the choice of point-of-pivot has not been examined for South African low-resourced languages. To the best of our knowledge, it hasn't been tested on any other language pairs either. We therefore pose the following research question: Are results obtained with linguistically-determined pivot languages consistent with the transfer capabilities of the target languages relative to other languages? We investigate three different techniques for generating entangled representations between all 11 South African languages: (i) VecMap [48] – a model that exploits structural similarities based on the isometry assumption and self-learning to iteratively improve embeddings projection; (ii) Muse [49] – a model that reconstructs representations from latent space using a sentence denoising objective function, and (iii) CCA [43] – a technique that learns correlated projection vectors by maximizing the correlation between the projected vectors. In summary, our major contributions are succinctly organized as follows:

[1] https://southafrica-info.com/arts-culture/11-languages-south-africa/#sources

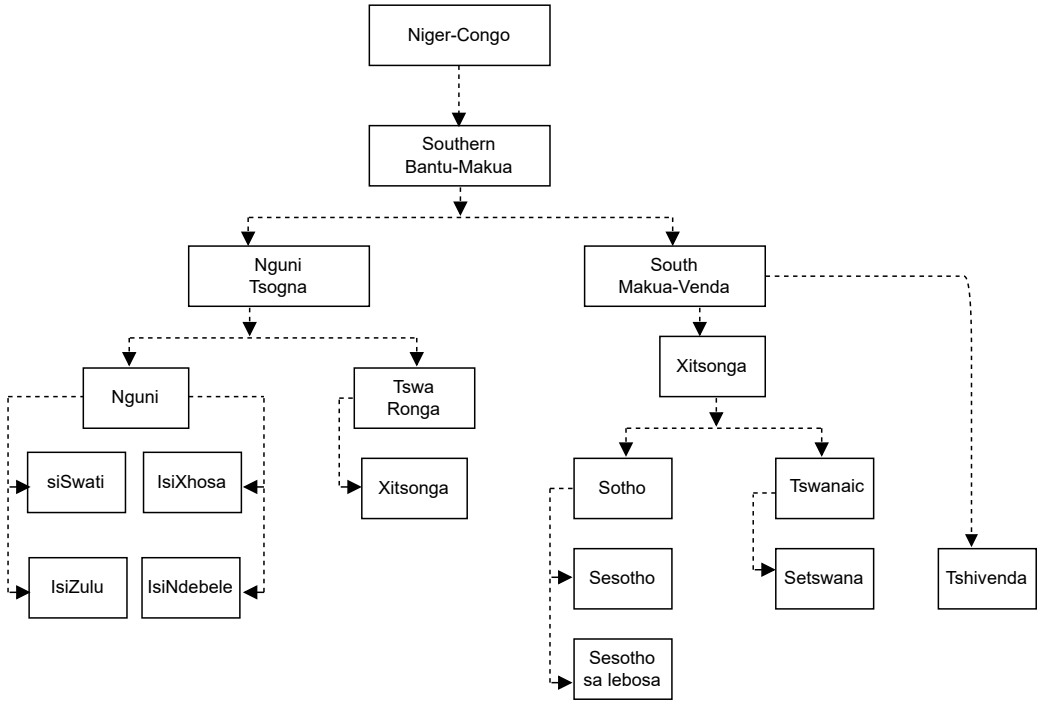


Figure 4.2: South African language and origins [8].

- We investigate a pivot language calibration-study of South African languages from three projection models: VecMap, Muse, and CCA.
- We extensively evaluate 110 permutations of cross-lingual embeddings using cosine similarity (intrinsic) and four extrinsic evaluation tasks: Machine Translation (MT), News Headlines Classification (NHC), Named Entity Recognition (NER), and Part of Speech (POS) Tagging.
- We build the first topology of transfer performance across all South African languages.

This work aims to highlight South African languages that transfer well and that can be seamlessly exploited to develop mutual resources for enhancing NLP research.

4.2 Related Work

Generating entangled representations using cross-lingual mathematical models has shown practical immanence, especially when supplemented by accompanying resources such as bilingual lexicons [2, 92, 103]. However, the research on the choice of pivot language from which to base the entanglements, for

optimal transfer performance together with language closeness evaluations and determination, remains unattempted [166]. Recent works rely on linguistic insights such as mutual intelligibility for source language selection. For example, [178] created cross-lingual embeddings between mutually intelligible languages, namely, Kinyarwanda and Kirundi, and evaluated them on a news classification task. In their paper, the source language choice is based on high lexical overlap and similarity and not on empirically informed source language choice. Lately, a new wave of research has emerged embracing a need for better source language selection (referred to as the closeness index of languages in Magueresse et al. [166]).

To demonstrate that typological agreement between two languages is not necessarily a positive predictor of transfer performance, Buys and Botha [220], experimented using MaltPaser model on the task of morpholical parsing to highlight optimal transfer performance for unrelated languages. Cardenas et al. [221] also highlighted special cases where the choice of pivot language is essential for optimal transfer capabilities on the task of part-of-speech tagging. These two studies demonstrated that multilingual learning was needed to decypher high-order patterns essential to better transfer capabilities. Similarly, [222] observed that cross-lingual training grounded by careful selection of the source language can result in improved transfer robustness on dependency parsing tasks. While [223] argue, instead of a single source search, incremental value can be attained by transferring multi-source derived delexicalized parsers to low-resourced languages and witnessing improved performance compared to single-source transfer learning.

In parallel with adapting pre-trained multilingual models for low-resourced settings, [28] investigated pivot language selection for optimal transfer performance on the Grammatical Error Correction (GEC) task. Gogoulou et al. [79], investigated the impact of source monolingual models adaptation to target downstream tasks. They concluded that knowledge learned independently of the source language enhances transfer performance. Regardless, finding a language with such abstractions to foster optimal performance requires some selection mechanism. Adelani et al. [224], showcased improved F1-score performance in a zero-shot setting as a product of informed pivot language selection. In [14], authors explored source language selection for zero-shot transfer of part-of-speech tagging using pre-trained multilingual models. A commonality in these works arises from the pressing need to select an optimal source language for a specific target language for improved transfer capabilities. Regardless, practical evidence of mutually intelligible languages being the best choice for point of pivot remains sparse. In this work, we explore a brute-force approach to investigate language permutations with better projection capabilities and later transfer performance to subsequent downstream tasks: NER, POS, NHC, and Machine Translation. Our exploratory study considers all 11 South African languages, resulting in 110 pairs of forward and backward projections.

4.3 Methodology

Bilingual-lexicons aided cross-lingual representation learning has shown better results compared to unsupervised learning counterparts. For this reason, all our cross-lingual representation learning focuses on the use of bilingual lexicons for training our three projection techniques: Muse, VecMap, and Canonical correlation analysis. In this section, we discuss the datasets collected (bilingual lexicons and monolingual data), projections models, and evaluation strategies.

4.3.1 Data

This section discusses various types of datasets collected in this study. The language codes are as follows: Afrikaans (af), IsiNdebele (nr), Sepedi (nso), IsiSwati (ssw), Sesotho (st), Setswana (tsn), Xitsonga (tso), Tshivenda (ven), IsiXhosa (xho) and IsiZulu (zul).

4.3.1.1 Multilingual Lexicons

Our bilingual lexicons are collected from the government public school repositories[2], CPUT[3], and Open Education Resource Term Bank (OERTB) [225]. Figure 4.3 shows the distributions of all lexicons collected for all combinations of 11 languages in South Africa. The English-Zulu and Zulu-English lexicons are the largest with ∼17 000 translation pairs. From these bilingual lexicons, we created our multilingual lexicons, where each language is mapped to all other languages. Our multilingual lexicons are processed and formatted into a machine-readable format and made available online[4].

4.3.1.2 Monolingual Data

Monolingual data used to train our FastText monolingual embeddings are sourced from multiple repositories including Flores[5], WMT[6], MC4[7], NCHLT [205], and African Crawl Dataset[8]. Data description and sources are obtained from [226]. We excluded retraining English monolingual embeddings and used GloVe embeddings instead [87].

[2] https://www.dsac.gov.za

[3] https://mlg.cput.ac.za/

[4] https://github.com/dsfsi/za-bilingual-lexicons

[5] https://github.com/facebookresearch/flores

[6] https://huggingface.co/datasets/allenai/wmt22_african

[7] https://huggingface.co/datasets/mc4

[8] https://github.com/pavanpankaj/Web-Crawl-African

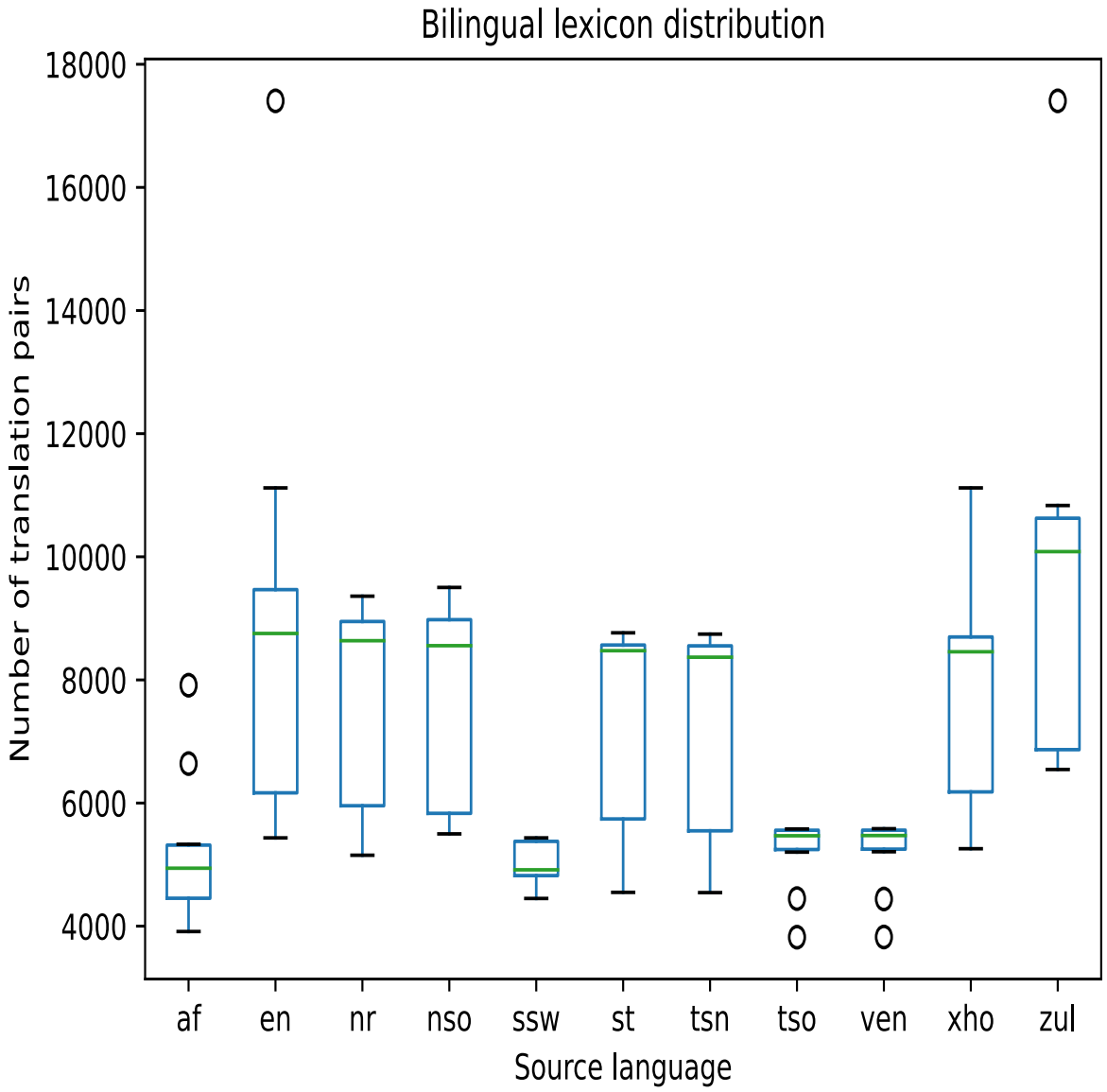


Figure 4.3: Number of translation pairs for each source language and the other 10 languages.

4.3.1.3 Machine Translation Data

Our MT experiments used a combination of multilingual lexicons with the Machine Translation test set released in Federmann et al. [227]. Each translation pair in their paper contains roughly 1500 examples. We divided the data into the train (80% of the total set), the development (20% of the train set), and the test set. The multilingual lexicons were divided according to the same split ratios and appended to

the translation set. Adding multilingual lexicons expanded our dataset.

4.3.1.4 News Headlines Classification (NHC) Data

The availability of annotated and quality resources such as datasets for low-resources poses a pressing challenge to the NLP community. We used Sepedi and Tshivenda quality annotated datasets from Adelani et al. [228], Madodonga et al. [229], Marivate et al. [230]. To expand to other low-resourced languages we used Google Deep Translate (GDT)[9], to translate the available gold datasets to af, nso, sot, xho, tso, and zul low-resourced languages and retained the original label for each translated sentence.

4.3.1.5 Named Entity Recognition Data

Similar to NHC and other downstream tasks, NER, is no exception to the shackles of low-resourcedness. As a result, we adopted the same methodology using MT and available gold datasets to expand datasets for other South African languages. Since GDT is limited to only Sepedi, Sesotho, IsiZulu, IsiXhosa, and Xitsonga, our generated datasets are bounded by that capability. For NER, we performed token-token translation to translate each token in the corpus, and the source-token label is used as the target-token label directly. For multiple-word translations, each new token is given the label of the source token. As source language data, we used MasakhaNER2.0 datasets [224]. No further quality control was done on the generated dataset, instead, we evaluated the quality of the datasets by training sequance2sequance models with translated examples and testing them on gold datasets.

4.3.1.6 Part of Speech Tagging Data

Annotated POS datasets for low-resourced languages are very scarce. For this reason, similar to NHC and NER, we also used the recently released annotated POS datasets from Dione et al. [14] in conjunction with Machine Translation to generate pseudo-annotated datasets for af, nso, sot, xho, zul, and tso low-resourced languages. That is, we used GDT to perform token-to-token translation of the available zul and xho POS datasets. Similar to NHC, and NER, quality control for the generated data was done by training a model on the generated data and evaluating on the existing gold datasets.

4.3.2 Models

4.3.2.1 Cross-lingual Models

We investigated three supervised projection strategies for creating cross-lingual embeddings, namely, Muse [49], VecMap [48], and Canonical correlation analysis [43] using our multilingual lexicons. Our

[9] https://pypi.org/project/deep-translator/

continuous vector representations were all of 200 dimensions. In each of the three models, we conduct a calibration study where we extract 110 permutations (see equation 4.1, where n = 11 is the number of languages, r = 2 represents the pairs) of cross-lingual embeddings in order to discover favorable transfer pairs between the 11 languages.

$$
\begin{aligned}
{}_nP_r &= \frac{n!}{(n-r)!} = \frac{11!}{(11-2)!} \\
&= \frac{11 \times 10 \times 9!}{9!} = 110
\end{aligned}
\tag{4.1}
$$

4.3.2.2 Classification Models

We evaluated and benchmarked multiple sequential models such as Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Gated Recurrent Units (GRU), and Bidirectional LSTM (BiLSTM) for tasks - POS and NER. For a more complex machine translation task, we used a hybrid GRU + Attention mechanism architecture preceded by monolingual or cross-lingual embeddings.

4.3.3 Evaluations

In this section, we compare our findings for cosine similarity as an intrinsic feature and the downstream NLP task as an extrinsic feature.

4.3.3.1 Cosine similarity as an intrinsic feature

In the first tier, we intrinsically evaluate our projected representations using cosine similarities. We expect sampled translation pairs to have a high cosine similarity score. Likewise, non-translation pairs are expected to have a low similarity score. This is done to evaluate if the projection model was able to project similar embedding to the same space and dissimilar embeddings to separate spaces. Equation 4.2 gives the calculation of cosine similarity.

$$
\textit{Cosine Similarity} = cos(\theta) = \frac{\vec{a}}{\|\vec{a}\| \cdot \|\vec{b}\|}
\tag{4.2}
$$

A value closer to 1 implies that the vectors are more similar. This approach is applied to determine whether words and/or sentences are similar. For the various language pairs, the bilingual lexicons were converted to vectors using cross-lingual word embeddings. Where either the source or target comprised of more than one word, the average of the word vectors was used to represent a one-to-one vector for each translation pair. The average of all the cosine similarities was used to compare the effectiveness of the various cross-lingual embedding approaches as well as to identify which languages transfer better.

4.3.3.2 Downstream NLP is an example of an extrinsic task

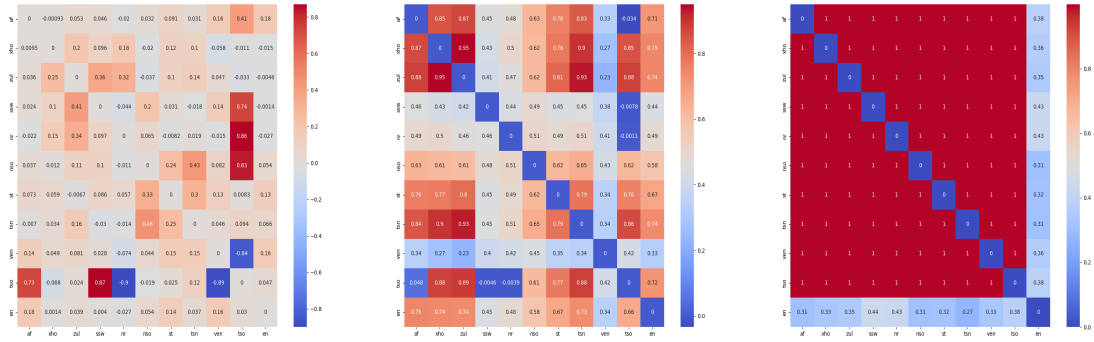
The second tier of evaluation looks at it from an extrinsic perspective where cross-lingual embeddings are evaluated on the following four downstream tasks: NER, POS, News Classification, and Machine Translation.

4.4 Results

This section discusses the results obtained from the word-vector similarity measure with cosine similarity, Machine Translation, News News Classification, Named Entity Recognition, and Part of Speech experiments.

4.4.1 Cosine Similarities

Figures 4.4a, 4.4b, and 4.4c show the cosine similarity scores (Section 4.3.3.1) for 110 cross-lingual embeddings from three models: CCA, VecMap, and Muse. Cosine scores in Figure 4.4b show clusters of cosine concentrations for languages from the same language family. This is in support of the persisting ideology that linguistically informed mutual intelligence is a strong source for point-of-pivot choices. However, these clusters of Figure 4.4b are not the performance values, which motivates the need for informed point-of-pivot insights. Additionally, IsiNdebele (a Southern African Nguni language) transfers better to the Sotho-Tswana language family. Locally, isiNdebele is known as being closer to the Sotho-Tswana language family, specifically, the Sepedi (Northern Sotho) language, while on paper it is recorded as belonging to the Nguni language family. In hindsight of this, we hope the language-family assignment for IsiNdebele can be revisited or clearer distinctions for the language can be made based on the demographics of the language. Regardless, cosine scores produced by CCA from Figure 4.4a indicate conflicting results with languages {Northern Sotho (nso), IsiNdebele (nr), IsiSwati (ssw), and Afrikaans (af)} showing the best transfer performance while the majority of the point-of-pivots underperform. Regardless, even higher concentrations manifest in different language groups. For example, the Sotho-Tswana family transfers well to IsiXhosa, IsiZulu, and Afrikaans, and vice versa. On the other hand, cosine scores produced by Muse outperform all other projection models except for any English-related projections. Figures 4.4a, 4.4b, and 4.4c indicate the progression of cross-lingual projection models, with CCA coming first, then VecMap, and followed by Muse (i.e., the model with the highest cosine scores). With these clearly distinguishing performances from the three models, the mathematical integrity underpinning the projections may range greatly, necessitating the application of specific theoretical assumptions and mathematical modelling, which requires additional exploration. Unravelling this information may help in making the best model selection decision for projections, depending on available resources and intentions.



(a) CCA cosine scores (b) VecMap cosine scores (c) Muse cosine scores

Figure 4.4: Comparison of cross-lingual transfer models.

4.4.2 Machine Translation

Table 4.1 reports the performance of our cross-lingual + attention mechanism + Gated recurrent units encoder-decoder model for translating sentences across the 110 permutations. Overall, English highlights the best average (08.52%) accuracy score across all languages. This could be because GloVe embeddings learned with sufficient data encapsulated sufficient semantic, syntactic, or morphological patterns necessary for transfer. Regardless, we leave this for future work to decipher. However, an individual inspection of each transfer source language highlights that each target language has its optimal source language which is not English. For example, af transfers well from nso (8.48%) with (0.12%) points higher than English, sot transfers well from nr (98.97%), and tso transfers well from zul, and nso with 98.31% and 98.30% respectively. The trend of results signifies the need to properly select a point of pivot for optimal transfer performance.

Our second analysis of our MT model compares the results obtained using cross-lingual embeddings (Table 4.1), with the results of MT model trained with monolingual embeddings (Table 4.2). Each language code maps in both forward and backward directions to represent cases where it was the source and when it was the target language. The gray-colored values are results from the development set and not the test set. This is due to computing limitations, where our model could not run end-to-end. The goal is to observe any disparities in model performances of entangled representations and their monolingual counterparts. Indeed, there is a slight improvement when training with cross-lingual embeddings compared to their monolingual counterparts on average. Additionally, pair-based analysis shows that entangled representation can improve the accuracy of the trained low-resourced models. Noticeably, English (as the target) in monolingual training outperformed cross-lingual embedding training in almost all languages. This means projection models struggled to improve further the already sufficiently learned GloVE embeddings. Interestingly, our results also highlight a noticeable difference in the backward and

↕↔	en	af	xho	zul	ssw	nr	nso	tsn	sot	tso	ven	Avg.
ven	98.47	97.56	97.86	97.92	97.53	97.84	97.76	97.86	97.11	98.14		97.81
tso	98.32	97.28	97.85	97.96	97.62	97.84	96.84	97.27	96.87		97.29	97.51
sot	98.74	97.51	98.04	98.32	97.73	97.95	98.30	98.11		98.24	97.55	98.05
tsn	98.70	98.28	98.08	98.20	97.74	97.94	98.40		98.34	98.18	97.58	98.14
nso	98.85	98.48	98.27	98.20	97.69	97.97		97.67	98.59	98.30	98.00	**98.20**
nr	99.03	98.36	98.10	98.16	97.68		97.71	98.16	98.87	98.00	97.15	98.12
ssw	98.28	98.20	97.53	97.43		97.46	98.00	97.33	98.73	97.86	96.86	97.77
zul	-	97.62	-		97.64	-	98.23	98.24	98.15	98.31	97.51	97.96
xho	98.00	97.80		98.16	97.85	97.96	98.31	97.41	97.70	98.22	97.37	97.89
af	97.80		97.55	96.77	96.87	96.80	95.36	94.65	95.87	95.66	94.67	96.2
en		98.36	-	-	97.91	98.48	98.82	98.80	98.85	98.29	98.71	**98.53**
Avg.	98.47	97.95	97.91	97.90	97.63	97.80	97.77	97.55	97.91	97.92	97.27	97.87

Table 4.1: Performance accuracy (in %) using cross-linguals (CLs) embedding with attention mechanisms and gated recurrent units (GRU)

forward transfer performance of language pairs in both monolingual and cross-lingual training. This means that one language benefits more from the entanglement (in the case of cross-lingual learning) and translation (in the case of monolingual) than the other language. This could stem from one language having effective vocabulary representation while the other does not, word embedding quality, and or varying training data sizes [78, 231]. This means, valid cross-lingual mappings exist for one direction while not often available in one direction to assist with the translation. It would be interesting to investigate further why this is the case, however, we leave it for future work. Our results are not intended to be state-of-the-art but shed light on the inherent need to properly select pivot language currently persisting in research for South African languages.

4.4.3 News Headlines Classification

Table 4.3 represents the cross-linguistic weighted accuracy of the best sequence-sequence model (BiLSTM) on a news headline classification dataset. The diagonal values show each language's monolingual performance, and the values above the diagonal line signify that the language in the column is the source language. The results from Table 4.3 indicate that the Tswana-Xhosa (tso-xho) language pair exhibits the highest performance with a test F1 score of 71.05%, and accuracy of 70.80% when employing LSTM models. However, not all indigenous language pairs fared as well; for example, the Tswana-Xhosa (tsn-xho) pair and the Zulu-Tswana (zul-tsn) pair lag considerably behind, with accuracies of 37.9% and 47.7%, and F1 scores of 37.2% and 48.5%, respectively. Setswana (tsn) and Northern Sotho (nso), despite being in the same Sotho language family, exhibit only moderate cross-lingual transferability with an accuracy of approximately 58.0% and an F1 score of 58.0% when employing LSTM models. This could suggest that shared family lineage is not the sole determinant of successful cross-lingual transfer, and other factors like corpus-specific nuances or social proximity may be more important. The findings also challenge the common assumption that linguistic and typographical similarity alone guarantees

↕↔	en	af	xho	zul	ssw	nr	nso	tsn	sot	tso	ven	Avg.
ven	99.16	97.87	98.16	97.95	97.60	97.86	97.95	97.82	96.87	98.12		97.94
tso	97.18	97.18	97.88	97.94	97.56	97.86	97.64	96.54	98.09		97.39	97.53
sot	98.10	97.44	98.12	98.16	97.78	98.00	98.36	97.29		98.30	98.39	97.99
tsn	99.43	99.39	98.13	98.16	97.72	97.95	97.93		97.31	98.26	97.59	**98.19**
nso	99.43	98.28	98.02	98.19	97.77	97.96		97.96	98.09	97.27	-	98.11
nr	98.03	97.11	98.07	98.20	97.74		98.34	97.37	98.30	98.21	96.77	97.81
ssw	99.10	97.34	97.40	97.60		97.53	97.73	97.27	96.03	97.95	96.85	97.48
zul	-	97.62	-		98.00	-	98.24	97.76	98.58	98.36	98.63	**98.17**
xho	99.55	97.59		98.14	97.73	97.93	98.36	97.67	98.69	98.21	97.48	98.16
af	98.21		97.75	97.26	97.19	97.10	98.99	95.80	96.74	97.48	95.98	97.25
en		97.61	-	-	97.56	97.98	98.51	97.36	98.33	98.10	97.35	97.85
Avg.	98.69	97.74	97.94	97.96	97.67	97.80	98.21	97.28	97.70	98.03	97.38	97.86

Table 4.2: Performance accuracy (in %) using monolingual embedding with attention mechanisms and gated recurrent units (GRU)

better cross-lingual transferability.

On the other hand, Setswana (tsn) and Venda (ven) pair also show good transferability under the LSTM models, albeit slightly lower than the nso-ven pair, with accuracies around $54-59\%$. The strong transferability observed between Northern Sotho (nso, also known as Sesotho sa Leboa) and Venda (ven) can be initially surprising, given that they belong to different language families. However, the high transfer performance could be attributed to other factors such as the geographic proximity of the Pedi and VhaVenda people. This hypothesis is further supported by Venda (ven) and Tsonga (tso) pairing which shows strong cross-lingual transferability, with an accuracy of approximately 63.3% and a corresponding F1-score of approximately 63.5%.

From the Nguni family, we observe F1-scores in the range of approximately 58% to 60% for isiZulu (`zul`) and isiXhosa (`xho`) when paired together, signaling strong cross-lingual transferability. Moreover, both isiXhosa and isiZulu exhibit strong transferability with Northern Sotho (nso), with F1 scores of approximately 66.8% and 66.2%, and accuracies exceeding 66.5% and 65.8%, respectively, when using LSTM models. While the isiZulu (`zul`) and Venda (`ven`) pair exhibits moderate potential for cross-lingual transfer learning with an accuracy of about 55.6% and an F1 score of approximately 58.4%, it does not reach the high compatibility levels observed in other language pairs such as nso-ven, which has an F1 score and accuracy of over 65%, or nso versus the Nguni languages.

On the other end of the spectrum, we observe that the performance of English with languages like Zulu (eng-zul) and Northern Sotho (eng-nso) suggests that English could be a viable pivot language for certain indigenous languages, with accuracies of 68.9% and 69.4%, and F1 scores of 70.3% and 69.5%, respectively. Notably, Afrikaans also demonstrates plausible performance, with an accuracy of 59.2% with Zulu (afr-zul) and a notably higher accuracy of 66.7% with Northern Sotho (afr-nso), thereby indicating that Afrikaans could also serve as a potential pivot language, albeit not as effectively as English in this case.

Upon closer examination, it is evident that English and Afrikaans possess a remarkably promising capacity for cross-lingual transfer with various Bantu languages. The performance metrics, particularly

	Target Language								
Source Language	afr	eng	nso	tsn	tso	ven	xho	zul	**Avg.**
afr	63.67	76.7	66.67	56.29	53	64.25	63.0	59.24	**62.85**
eng	-	77.29	69.36	58.84	78.8	69.26	69.00	68.94	**70.21**
nso	-	-	66.56	58.38	69.34	65.34	65.12	65.83	**65.1**
tsn	-	-	-	40.18	55.03	59.01	47.01	45.83	**49.41**
tso	-	-	-	-	65.60	63.29	70.80	64.93	**66.12**
ven	-	-	-	-	-	52.49	53.88	52.42	**52.93**
xho	-	-	-	-	-	-	53.40	58.67	**56.04**
zul	-	-	-	-	-	-	-	56.82	**56.82**
Avg.	**63.67**	**77**	**67.53**	**53.42**	**64.35**	**62.27**	**60.32**	**59.09**	**59.94**

Table 4.3: Cross-lingual weighted accuracy of best sequence-sequence model (BiLSTM) on a News headline classification dataset.

the F1 scores and accuracies exceeding 64%, highlight this trend. Such findings are particularly revelatory, given the conventional preconception about the challenges of cross-lingual transfer involving low-resource languages and languages from Indo-European language families. The results defy the conventional expectations that low-resource languages would inherently perform poorly when paired with languages from different families for transfer learning. It opens up intriguing possibilities for leveraging English and Afrikaans as pivot languages in NLP tasks involving Bantu languages.

Finally, our results indicated that incorporating back-translated English articles may not be the optimal approach for augmenting annotated news classification resources in indigenous languages like Zulu, Xhosa, or Sotho through GDT. While this result may come as a surprise, it likely, the translations from English into indigenous Bantu languages were not precise enough. We hypothesize that utilizing sophisticated, pretrained cross-lingual transformer models like XLM-Roberta could offer a more promising solution. However, this is not investigated here. We do however emphasize that there are untapped opportunities to exploit the linguistic relationships between indigenous languages, and Indo-European languages.

4.4.4 Part of Speech Tagging

Table 4.4 shows the best sequence-sequence model trained with cross-lingual embeddings on our GDT POS-generated data with the best pivot language for each target language. Additionally, xho and zul values (in the shared columns xho(gdt)/xho and zul(gdt)/zul) in this table show the evaluation of GDT datasets (left value) and the zero-shot transfer of the best model on the original dataset for available gold datasets (xho and zul) for the task POS. Cross-representation training demonstrated potential performance compared to their monolingual counterparts as shown in Table 4.6 evaluated with the Accuracy metric. Xitsonga performed well compared to all pivot languages with an average of 92.83%. It would be interesting to investigate why this language performed better compared to all source languages by deeply inspecting its source embedding space in relation to the task. Interestingly, Figure 4.4b highlights high cosines of tso to other languages. However, we leave this for future work.

Src	af	xho(gdt)/xho	zul(gdt)/ zul	nso	st	tso	Avg.
af		92.55/91.59	91.97/91.53	92.64	92.74	92.35	92.45
xho	93.97		91.84/90.63	92.56	92.78	92.73	92.78
zul	93.72	92.66/91.59		92.53	92.91	92.22	92.81
nso	93.91	92.55/92.01	91.76/91.52		92.76	92.50	92.70
st	94.00	92.80/91.24	91.72/91.02	92.57		92.78	97.77
tso	93.64	93.05/91.72	91.92/91.33	92.53	93.00		**92.83**
Avg.	93.85	92.72/91.63	91.84/91.21	92.5	92.84	92.52	93.56

Table 4.4: Accuracy scores of cross-lingual embeddings using BiLSTM sequence-sequence model on GDT POS data.

Table 4.5 on the other hand shows extensive experimental results for models RNN, LSTM, GRU, and BiLSTM with BiLSTM consistently outperforming the other 3 models.

4.4.5 Named Entity Recognition

Similar to GDT-generated POS data, we evaluated cross-lingual embeddings on GDT-generated NER data and reported the best model in Table 4.7. The full cross-lingual results and monolingual trained models are reported in Table 4.9, and Table 4.8 respectively. A consistent pattern emerges, where cross-trained model outperforms their monolingual counterpart. Despite this, a major highlight of the NER results is the reliability of GDT-generated NER datasets observed during the transfer of GDT-trained models to original datasets as reported in Table 4.8 and Table 4.7, respectively.

4.5 Conclusion

We created 110 × 3 (VecMap, Muse, and CCA) permutations of cross-lingual embeddings from 11 South African languages and evaluated them on both intrinsic tasks (Word Similarity with cosine similarity scores) and extrinsic downstream tasks (Machine Translation, News Headlines Classification, NER and POS tagging). Moreover, our results shed light on the need to explore pivot language selection for optimal transfer performance and the use of entangled representations for improved learning compared to monolingual embeddings. In addition to showcasing the efficacy of shared representation learning, we highlighted the potential of expanding linguistics-based datasets such as POS, NER, and text classification datasets: News Headlines Classification using Machine Translation that still preserves high similarities with the gold datasets. The link to the source repository and available material is available here for the implementation of this method and adaptation for future work[10],[11].

[10] https://github.com/dsfsi/thapelo-sindane-msc-public.git

[11] https://github.com/dsfsi/za-bilingual-lexicons

Src.	Model	af	xho(gdt)/xho	zul(gdt)/zul	nso	st	tso	Avg.
af	RNN		92.54/91.12	91.52/90.06	91.21	91.58	91.15	91.31
	LSTM		92.82/91.37	91.79/90.47	91.50	91.90	91.64	91.64
	GRU		92.80/91.27	91.48/90.32	91.57	92.17	91.42	91.58
	BiLSTM		92.55/91.59	91.97/91.53	92.64	92.74	92.35	92.2
xho	RNN	93.03		91.31/90.19	91.26	91.07	91.28	91.36
	LSTM	93.29		91.58/90.85	91.28	91.58	91.63	91.7
	GRU	93.17		91.43/90.72	91.65	92.16	91.66	91.8
	BiLSTM	93.97		91.84/90.63	92.57	92.78	92.73	92.42
zul	RNN	92.82	92.23/91.00		91.54	91.01	90.77	91.56
	LSTM	93.35	92.70/91.30		91.60	91.69	91.81	92.08
	GRU	93.38	92.28/91.40		91.84	91.70	91.51	92.02
	BiLSTM	93.71	92.66/91.59		92.53	92.92	92.22	**92.61**
nso	RNN	92.74	92.42/91.10	91.16/90.61		91.35	91.13	91.50
	LSTM	93.34	92.92/91.23	91.69/90.68		91.62	91.63	91.87
	GRU	93.15	92.65/91.00	91.42/90.47		92.02	91.71	91.77
	BiLSTM	93.91	92.55/92.01	91.76/91.52		92.76	92.50	92.43
st	RNN	92.91	92.31/91.18	91.44/90.57	91.31		91.26	91.57
	LSTM	93.26	92.70/90.92	91.77/90.44	91.68		91.72	91.78
	GRU	92.57	93.00/91.14	91.42/90.54	91.67		91.57	91.70
	BiLSTM	94.00	92.80/91.24	91.72/91.02	92.57		92.76	92.30
tso	RNN	92.96	92.21/90.62	91.00/90.48	91.34	91.51		91.45
	LSTM	93.30	93.01/91.21	91.66/90.36	91.70	92.04		91.90
	GRU	93.32	92.30/91.31	91.21/90.98	91.87	91.96		91.85
	BiLSTM	93.64	93.05/91.72	91.92/91.33	92.53	93.00		**92.46**
ssw	RNN	92.61	91.89/91.11	91.10/90.28	91.45	91.34	91.17	91.37
	LSTM	93.32	92.75/91.22	91.86/90.44	91.60	91.64	91.86	91.84
	GRU	93.31	92.92/91.23	91.68/90.57	91.70	92.18	91.40	91.87
	BiLSTM	93.99	92.77/91.76	91.68/90.99	92.69	93.03	92.56	92.43
nr	RNN	92.80	92.33/90.94	91.31/89.94	91.21	91.42	90.78	91.34
	LSTM	93.65	92.45/91.38	91.30/90.21	91.61	91.75	91.49	91.73
	GRU	93.34	92.67/91.29	91.49/90.56	91.73	92.05	91.48	91.83
	BiLSTM	93.58	92.79/91.89	91.75/90.44	92.87	92.86	92.69	92.36
ven	RNN	93.04	92.15/91.00	91.17/90.01	91.04	91.57	91.24	91.40
	LSTM	93.32	92.76/91.42	91.76/90.98	91.82	91.94	91.61	91.95
	GRU	93.31	92.50/91.27	91.17/90.17	91.64	91.65	91.76	91.68
	BiLSTM	93.95	92.54/91.53	91.71/91.38	92.12	92.44	92.74	92.30
en	RNN	91.96	92.15/91.16	90.66/90.09	90.68	90.58	90.43	90.96
	LSTM	92.85	92.49/91.12	91.51/90.18	91.31	91.73	91.39	91.57
	GRU	92.71	92.44/91.20	91.34/89.85	91.09	91.52	91.47	91.45
	BiLSTM	92.98	92.26/91.77	91.59/90.93	91.91	92.02	92.18	91.96

Table 4.5: Accuracy scores of RNN, GRU, LSTM, BiLSTM sequence-sequence models on GDT POS data using cross-lingual embeddings

4.6 Limitations

- **Optimization**
  This study did not conduct hyper-parameter tuning to determine optimal parameters on cross-lingual models or our sequence-to-sequence models for training and inference. From this, optimal settings would be extracted to determine the best-case scenarios for learning representations together with downstream tasks training. This would provide, not only a clear and fair indication of model comparisons but also a good estimate of practical implications in the case of cross-lingual models. However, we leave this for future works to exploit rather than explore (as

Models	af	xho(gdt)/xho	zul(gdt)/ zul	nso	st	tso	Avg.
CRF	78.66	80.46/60.60	78.39/64.40	76.02	76.06	76.92	77.42
RNN	92.16	92.22/91.38	91.32/90.44	91.42	91.26	91.25	91.61
LSTM	92.94	92.53/91.29	91.65/90.88	91.65	92.02	91.46	92.04
GRU	93.00	92.69/90.95	91.17/90.57	91.96	92.03	91.42	92.05
BiLSTM	93.92	92.40/91.79	91.74/90.99	92.33	92.79	92.44	**92.60**
Avg.	90.14	90.06/85.20	88.85/85.46	88.68	88.83	88.70	89.14

Table 4.6: Accuracy scores of best sequence-sequence model on GDT POS data using monolingual embeddings

done in this study) representative samples of this study to uncover useful insights for practical implementations of this work. For example, one study can look at a few cross-lingual representations with best-performing source-target languages, and investigate hyper-parameter search for best training models (e.g BiLSTM and more) to uncover favorable parameters with promising practical implications.

- **Evaluation metrics**
  This study consistently adopted accuracy to compare types of projection models, capabilities of traditional machine learning models and deep neural networks, and efficacy of expanding datasets with machine-generated datasets. However, accuracy as a metric is sometimes not the best reflector of performance. As such, it would be interesting to observe a representative sample of this study adopting other robust metrics such as F1-score, Precision, and recall to evaluate the aforementioned variables of this study. More importantly, it would be interesting to see if the conclusions are consistent, if not, why. However, we leave this for future work.

- **Benchmarking**
  This study did not compare our findings with any existing benchmark on the available resources or datasets we used. In this study's defense, the work is intended to showcase the need to select an optimal source language for the best transfer performance. And due to space limitations, we were unable to perform additional experimentation's outside the 110×3 observed cross-lingual embeddings. We leave this for future works.

- **Large Language Models (LLM)**
  The recent trend of multilingualism adopts large language models that implicitly learn shared representations between languages for optimal transfer performance. This study did not investigate any large language model to compare with traditional cross-lingual models. It would be interesting to investigate the difference in transfer performance between traditional cross-lingual models and large language models and what leads to these disparities. Additionally, it would be important to uncover which model type (i.e cross-lingual or LLM-based) performs well under which tasks, and why this is the case. Regardless, we leave this for future work.

Src	af	xho(gdt)/xho	zul(gdt)/ zul	nso	st	tso	Avg.
af		98.99/98.84	98.65/98.47	98.61	98.95	98.88	98.82
xho	99.08		98.71/98.52	98.55	99.09	99.00	98.89
zul	99.13	98.96/98.81		98.62	99.13	98.87	98.94
nso	99.03	99.01/98.81	98.64/98.61		99.08	98.95	98.94
st	99.00	98.98/98.67	98.63/98.59	98.46		98.93	98.80
tso	99.20	98.95/98.72	98.69/98.58	98.55	99.10		**98.90**
Avg.	99,09	98,98/98,77	98,66/98,55	98,56	99,07	98,93	98,88

Table 4.7: Accuracy scores of BiLSTM sequence-sequence model on GDT NER data using cross-lingual embeddings

4.7 Summary

This section summarizes the works of this Chapter. The main theme of this work is to advocate the need to explore and investigate optimal source language selection for possibly optimal transfer performance. We further wanted to showcase the limitations of the persisting ideology that linguistic mutual intelligibility may lead to optimal transfer performance. To achieve this, we gave a detailed overview of the field in Section 4.1 (summarized in 4.7.1 below), we sourced related works in Section 4.2 (summarized in 4.7.2, gave a detailed description of our approach in Section 4.3 (summarized in 4.7.3), we shared our finding in Section 4.4 (summarized in 4.7.4), and finally highlighted our conclusions in Section 4.7.5 (summarized in 4.7.5).

4.7.1 Introduction

Cross-lingual representation learning has shown significant improvement over the last decade. With a better cross-lingual model gradually being developed, transfer capabilities are proportionally improved. However, the impact of source language selection on transfer performance has not been extensively investigated for South African languages or any other languages. From this, our introduction shares a brief literature landscape on this issue and how we intend to contribute to the field to address it.

4.7.2 Related Works

The literature section scopes out works that share the same sentiments as this work in terms of highlighting the need to have a mechanism to select a source language with enough shareable and useful properties for transferring to the target language.

4.7.3 Methodology

The methodology adopted in this study involved training 110 permutations of cross-lingual embeddings for each projection technique – CCA, VecMap, and Muse resulting in a collection of 3×110 cross-

Models	af	xho(gdt)/xho	zul(gdt)/ zul	nso	st	tso	Avg.
CRF	96.44	94.43/85.06	92.59/90.53	95.05	96.10	96.29	95.15
RNN	98.87	98.58/98.32	98.23/97.99	98.21	98.83	98.66	98.58
LSTM	98.77	98.58/98.32	98.41/97.78	98.23	98.87	98.73	98.60
GRU	98.76	98.65/98.38	98.47/98.26	98.33	98.90	98.67	98.63
BiLSTM	99.18	98.98/98.67	98.55/98.59	98.49	99.07	98.99	**98.88**
Avg.	98.40	97.84/95.75	97.25/96.63	97.66	98.35	98.27	97.97

Table 4.8: Accuracy scores of best sequence-sequence model on GDT NER data using monolingual embeddings

lingual embeddings. From these, embeddings we conducted some experiments on multiple tasks such as to observe three things – which projection technique reveals the best-shared representations, - two – are there any improvements when training models with cross-lingual embeddings compared to monolingual embeddings, and three – which languages serve as the best source languages on which tasks.

4.7.4 Results and Findings

Our results support the need to first discover a source language that can result in optimal transfer performance, and do away with relying on linguistic-extracted mutual intelligibility as a good indicator of transfer performance. For example, an unexpected language Xitsonga (tso), consistently outperformed most languages on most downstream tasks.

4.7.5 Conclusion

Our conclusion suggests the need to develop a mechanism for source language selection for optimal transfer performance. We further indicate the importance of shared representations compared to do their monolingual counterpart on multiple downstream tasks. Finally, this study also shows the importance of scaling practical work to other low-resource languages with the use of Machine translation due to the lack of quality annotated datasets.

Src.	Model	af	xho(gdt)/xho	zul(gdt)/zul	nso	st	tso	Avg.
	RNN		98.34/98.45	98.19/97.71	98.12	98.82	98.67	98,33
af	LSTM		98.52/98.34	98.01/98.20	98.25	98.84	98.76	98,42
	GRU		98.76/98.43	98.45/98.16	98.30	98.85	98.63	98,51
	BiLSTM		98.99/98.84	98.65/98.47	98.61	98.95	98.88	98,77
	RNN	98.81		98.28/98.18	98.09	98.52	98.72	98,43
xho	LSTM	98.64		97.79/98.12	98.26	98.89	98.66	98,39
	GRU	98.90		98.33/98.06	98.25	98.94	98.72	98,53
	BiLSTM	99.08		98.71/98.52	98.55	99.09	99.00	**98,83**
	RNN	98.83	98.70/98.31		98.20	98.78	98.66	98,58
zul	LSTM	98.72	98.36/97.83		98.28	98.91	98.64	98,46
	GRU	98.80	98.71/98.51		98.27	98.68	98.59	98,59
	BiLSTM	99.13	98.96/98.81		98.62	99.13	98.87	**98,92**
	RNN	98.87	98.58/98.42	98.29/97.31		98.87	98.53	98,41
nso	LSTM	98.78	98.58/98.18	98.19/98.27		98.75	98.58	98,48
	GRU	98.91	98.64/98.49	98.33/98.18		98.87	98.67	98,58
	BiLSTM	99.03	99.01/98.81	98.64/98.61		99.08	98.95	98,88
	RNN	98.84	98.65/98.49	98.17/98.19	98.26		98.66	98,47
st	LSTM	98.75	98.69/98.30	98.18/98.34	98.22		98.69	98,45
	GRU	98.85	98.64/98.46	98.44/98.32	98.20		98.71	98,52
	BiLSTM	99.00	98.98/98.67	98.63/98.59	98.46		98.93	98,75
	RNN	98.85	98.72/98.22	98.36/98.18	98.19	98.45		98,42
tso	LSTM	98.70	98.54/98.45	98.17/98.10	98.27	98.94		98,45
	GRU	98.90	98.73/98.41	98.24/98.12	98.27	98.90		98,51
	BiLSTM	99.20	98.95/98.72	98.69/98.58	98.55	99.10		**98,83**
	RNN	98.65	98.64/98.42	98.10/98.14	98.15	98.87	98.62	98,45
ssw	LSTM	98.79	98.72/98.49	98.28/97.89	98.24	98.88	98.62	98,49
	GRU	98.91	98.76/98.29	98.32/98.29	98.27	98.76	98.74	98,54
	BiLSTM	99.10	98.98/98.83	98.56/98.56	98.60	99.04	98.76	98,8
	RNN	98.90	98.58/98.48	98.26/98.21	98.29	98.84	98.63	98,52
nr	LSTM	98.83	98.63/98.34	98.10/98.05	98.26	98.90	98.66	98,47
	GRU	98.76	98.56/98.52	98.41/98.01	98.27	98.93	98.68	98,52
	BiLSTM	99.14	98.93/98.77	98.66/98.35	98.57	99.07	99.00	98,81
	RNN	98.76	98.52/98.28	98.39/98.02	98.24	98.77	98.65	98,45
ven	LSTM	98.79	98.59/98.09	98.24/98.01	98.19	98.86	98.75	98,44
	GRU	98.79	98.77/98.37	98.46/97.78	98.26	98.81	98.77	98,5
	BiLSTM	99.12	98.94/98.56	98.61/98.62	98.59	99.01	98.81	98,78
	RNN	98.67	98.54/98.34	98.26/98.24	98.16	98.87	98.59	98,46
en	LSTM	98.65	98.41/98.35	98.27/98.25	98.30	98.79	98.62	98,46
	GRU	98.82	98.66/98.47	98.26/98.27	98.11	98.86	98.71	98,52
	BiLSTM	99.10	98.97/98.76	98.58/98.58	98.59	99.03	98.97	98,82

Table 4.9: Accuracy scores of sequence-sequence models on GDT NER data and best transfer source using cross-lingual embeddings.

Chapter 5

Conclusions

This chapter summarizes key findings, our main contribution, conclusions, and highlights future prospects for our work in the area of addressing marginalized languages with cross-lingual shared representations. Retrospectively, our Chapter 2 discussed the history of cross-lingual models from inception to the current evolution of large multilingual pre-trained models. Additionally, we hinted at the value-add of these models on downstream tasks and provided a taxonomy that highlights research gaps and potential future directions. This literature work provided a comprehensive survey of the history of cross-lingual techniques and how they have been adopted to accelerate research in various Natural Language Processing domains. Chapter 3, on the other hand, provided a stepping stone towards the implementation of cross-lingual models in practice. In the aforementioned practical chapter, we provided empirical evidence supporting the efficacy of cross-lingual representation learning on tasks such as News Headlines Classification (NHC) and Named Entity Recognition (NER). The unfolded evidence supported the concept of shared representation by showcasing disparities in model performances on downstream tasks between cross-lingual representations learning and their monolingual counterparts. This is achieved by showcasing that training and evaluating models with the use of cross-lingual embeddings often show increased classification performance (e.g News Headlines Classification, Named Entity Recognition) over the monolingual counterparts, evaluated with several evaluation metrics such as accuracy, F1-score, recall, and precision. The experimental results obtained in this study, justify the need to explore cross-lingual embeddings even further as cheaper alternatives and build on their shortcomings.

Chapter 4, detailed a broader analysis of practical language transfer capabilities by exploring all transfer outcomes of all permutations of South African languages. That is, this study highlights languages that transfer well across multiple tasks such as Machine Translation (MT), Part of Speech (POS) Tagging, Named Entity Recognition (NER), and News Headlines Classification (NHC). In this study, we explored three algorithms for generating cross-lingual embeddings – Canonical Correlation Analyses (CCA), VecMap, and Muse for generating cross-lingual embeddings. We observed that, Muse-generated embeddings perform well intrinsically compared to other embeddings generated by the remaining two

methods. Furthermore, this study, covered tasks - NER, POS, NHC, and MT to extrinsically evaluate and compare performance gains between cross-lingual embeddings and monolingual embeddings. Evaluation on these tasks showed that choosing the best source language improves downstream task performance, as indicated by Xitsonga outperforming other languages (i.e a language not in active NLP research). Finally, Chapter 5 summarizes the contribution of each chapter 2, 3, and 4. The main contributions of this works are summarised as points below:

1. Our contribution to the scientific community:

    - We developed a comprehensive survey detailing various traditional cross-lingual models organized in a taxonomy that categorized these models based on their use, the resources they require, their formulation, e.t.c. This enables future research to be able to make decision on which cross-lingual techniques are most relevant depending of the state of resources available. We incision this to be extremely useful for future research, especially in low-resource languages, intending to explore cross-lingual models.
    - We have centralized monolingual datasets for more than 4 low-resourced languages and we are in the process of making these datasets publicly available.
    - We have developed and will be releasing both monolingual and cross-lingual embedding resources for all South African languages. Which to our understanding is still not publicly available or accessible.
    - We have shown empirically that cross-lingual transfer between English and South African languages is a possibility in Chapter 3. We have conducted various experiments comparing multiple traditional machine learning models such as XGBoost, Feed Forward Neural Networks, Long Short Term Memory (LSTM), Gated Recurrent Units (GRU), e.t.c on multiple downstream tasks such as NHC, NER, and Part of Speech Tagging.

2. What is new in this study?

    - This study developed both monolingual and cross-lingual embeddings for all South African languages. These resources will be made publicly available.
    - We have compared three cross-lingual techniques (Canonical correlation analyses, VecMap, and Muse) in one paper, for South African languages and evaluated these models intrinsically (word similarity - see figure) and extrinsically (NHC, and NER downstream tasks).
    - We have shown that the choice of the source language is important for South African languages, and discovered Xitsonga as an interesting case with improved transfer gains for other languages not in the same language family such as the Sotho-Tswana languages.

The next Section 5.1, covers the future directions that would be interesting to pursue for the presented work.

5.1 Future Work

This sections covers interesting future research avenues not considered in this thesis to support the literature and scientific community of cross-lingual embeddings with the aim of improving the status of low-resourced languages.

- It would be interesting for future works to explore evaluation techniques that can explain the innate linguistic patterns transferred from one monolingual embeddings space to the next.
- Future works should invest in the building of diverse dataset to expand the scope of evaluation for South African low-resourced languages.
- With availability of bilingual lexicons for South African languages, future works should consider works that include the creation, and analyses of pseudo-monolingual-corpora for adapting pre-trained cross-lingual language models. This extends to scaling models to all South African languages.
- It would also be interesting in the future to conduct a comprehensive exploratory study comparing the adaptation of different Afri-centric multilingual language models: mBERT, AfriBerta, and AfriLM using pseudo-monolingual-data created using bilingual lexicons.

Bibliography

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

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

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

[4] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[5] Karl Moritz Hermann and Phil Blunsom. Multilingual distributed representations without word alignment, 2014. URL https://arxiv.org/abs/1312.6173.

[6] Joakim Nivre. Dependency parsing. *Language and Linguistics Compass*, 4(3):138–152, 2010.

[7] Pali Lehohla. Census 2011: population dynamics in south africa. *Statistics South Africa*, 83, 2015.

[8] Jan Alewyn Nel, Velichko H Valchev, Sebastiaan Rothmann, Fons JR Van de Vijver, Deon Meiring, and Gideon P De Bruin. Exploring the personality structure in the 11 languages of south africa. *Journal of personality*, 80(4):915–948, 2012.

[9] Mohamad Zaim Awang Pon and Krishna Prakash KK. Hyperparameter tuning of deep learning models in keras. *Sparklinglight Transactions on Artificial Intelligence and Quantum Computing (STAIQC)*, 1(1):36–40, 2021.

[10] Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. What is your data worth to gpt? llm-scale data valuation with influence functions, 2024. URL https://arxiv.org/abs/2405.13954.

[11] Stylianos Mystakidis. Metaverse. *Encyclopedia*, 2(1):486–497, 2022.

[12] Felix Wong, Aarti Krishnan, Erica J Zheng, Hannes Stärk, Abigail L Manson, Ashlee M Earl, Tommi Jaakkola, and James J Collins. Benchmarking alphafold-enabled molecular docking predictions for antibiotic discovery. *Molecular Systems Biology*, 18(9):e11081, 2022.

[13] David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Joyce Nakatumba-Nabende, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. 2022.

[14] Cheikh M Bamba Dione, David Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, et al. Masakhapos: Part-of-speech tagging for typologically diverse african languages. *arXiv preprint arXiv:2305.13989*, 2023.

[15] Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*, 2022.

[16] Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. *arXiv preprint arXiv:2004.14911*, 2020.

[17] Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational ai. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1371–1374, 2018.

[18] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[19] Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. *arXiv preprint arXiv:2210.14712*, 2022.

[20] Mary Alexander. The 11 languages of south africa. *South Africa Gateway*, 2018.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[22] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*, 2020.

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[24] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

[25] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 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)*, pages 1234–1244, 2015.

[26] Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1393–1398, 2013.

[27] Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838, 2017.

[28] Ikumi Yamashita, Satoru Katsumata, Masahiro Kaneko, Aizhan Imankulova, and Mamoru Komachi. Cross-lingual transfer learning for grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4704–4715, 2020.

[29] Di Lu, Xiaoman Pan, Nima Pourdamghani, Shih-Fu Chang, Heng Ji, and Kevin Knight. A multimedia approach to cross-lingual entity knowledge transfer. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 54–65, 2016.

[30] Dmitry Karpov and Mikhail Burtsev. Monolingual and cross-lingual knowledge transfer for topic classification. *arXiv preprint arXiv:2306.07797*, 2023.

[31] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

[32] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[33] Kelechi Ogueji. Afriberta: Towards viable multilingual language models for low-resource languages. Master's thesis, University of Waterloo, 2022.

[34] Bonaventure FP Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages. *arXiv preprint arXiv:2211.03263*, 2022.

[35] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

[36] Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, et al. Afrisenti: A twitter sentiment analysis benchmark for african languages. *arXiv preprint arXiv:2302.08956*, 2023.

[37] Mack Makgatho, Vukosi Marivate, Tshephisho Sefara, and Valencia Wagner. Training cross-lingual embeddings for setswana and sepedi. *arXiv preprint arXiv:2111.06230*, 2021.

[38] Derwin Ngomane, Rooweither Mabuya, Jade Abbott, and Vukosi Marivate. Unsupervised cross-lingual word embedding representation for English-isiZulu. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 11–17, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.rail-1.2. URL https://aclanthology.org/2023.rail-1.2.

[39] Rozina Lucy Myoya, Fiskani Banda, Vukosi Marivate, and Abiodun Modupe. Fine-tuning multilingual pretrained african language models. In *4th Workshop on African Natural Language Processing*, 2023.

[40] Richard Lastrucci, Isheanesu Dzingirai, Jenalea Rajab, Andani Madodonga, Matimba Shingange, Daniel Njini, and Vukosi Marivate. Preparing the vuk'uzenzele and za-gov-multilingual south african multilingual corpora. *arXiv preprint arXiv:2303.03750*, 2023.

[41] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1): 4839–4886, 2021.

[42] Xinyi Wang, Sebastian Ruder, and Graham Neubig. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, 2022.

[43] Manaal Faruqui and Chris Dyer. 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, 2014.

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

[45] Katherine Yu, Haoran Li, and Barlas Oguz. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 175–179, 2018.

[46] Mram Kahla, Zijian Győző Yang, and Attila Novák. Cross-lingual fine-tuning for abstractive arabic text summarization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 655–663, 2021.

[47] Chia-Hsuan Lee and Hung-Yi Lee. Cross-lingual transfer learning for question answering. *arXiv preprint arXiv:1907.06042*, 2019.

[48] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*, 2018.

[49] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only, 2018. URL https://arxiv.org/abs/1711.00043.

[50] W John Hutchins. Warren weaver memorandum july 1949, 1949.

[51] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[52] Noam Chomsky. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124, 1956.

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

[54] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1):39–59, 1994.

[55] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[56] T De Heer. Experiments with syntactic traces in information retrieval. *Information Storage and Retrieval*, 10(3-4):133–144, 1974.

[57] Andrija Tomović, Predrag Janičić, and Vlado Kešelj. n-gram-based classification and unsupervised hierarchical clustering of genome sequences. *Computer methods and programs in biomedicine*, 81 (2):137–153, 2006.

[58] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[59] Youngjoong Ko. A study of term weighting schemes using class information for text classification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1029–1030, 2012.

[60] Harold Somers. Ebmt seen as case-based reasoning. In *Workshop on Example-Based machine Translation*, 2001.

[61] Sivic and Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings ninth IEEE international conference on computer vision*, pages 1470–1477. IEEE, 2003.

[62] Nikolaos Passalis and Anastasios Tefas. Entropy optimized feature-based bag-of-words representation for information retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 28(7): 1664–1677, 2016.

[63] Elena M Zamora, Joseph J Pollock, and Antonio Zamora. The use of trigram analysis for spelling error detection. *Information Processing & Management*, 17(6):305–316, 1981.

[64] Janusz L Wiśniewski. Effective text compression with simultaneous digram and trigram encoding. *Journal of Information Science*, 13(3):159–164, 1987.

[65] Richard C Angell, George E Freund, and Peter Willett. Automatic spelling correction using a trigram similarity measure. *Information Processing & Management*, 19(4):255–261, 1983.

[66] John C Schmitt. Trigram-based method of language identification, October 29 1991. US Patent 5,062,143.

[67] Victor V Solovyev and Kira S Makarova. A novel method of protein sequence classification based on oligopeptide frequency analysis and its application to search for functional sites and to domain localization. *Bioinformatics*, 9(1):17–24, 1993.

[68] William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175, page 14. Las Vegas, NV, 1994.

[69] J Stephen Downie. *Evaluating a simple approach to music information retrieval: Conceiving melodic n-grams as text*. Faculty of Graduate Studies, University of Western Ontario London, Ont., 1999.

[70] Madhavi Ganapathiraju, Deborah Weisser, Roni Rosenfeld, Jaime G Carbonell, Raj Reddy, and Judith Klein-Seetharaman. Comparative ngram analysis of whole-genome sequences. 2002.

[71] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264, 2003.

[72] Ji Qi, Hong Luo, and Bailin Hao. Cvtree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic acids research*, 32(suppl_2):W45–W47, 2004.

[73] Adnan El-Nasan, Sriharsha Veeramachaneni, and George Nagy. Handwriting recognition using position sensitive letter n-gram matching. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 577–582. IEEE, 2003.

[74] Betty Yee Man Cheng, Jaime G Carbonell, and Judith Klein-Seetharaman. Protein classification based on text document classification techniques. *Proteins: Structure, Function, and Bioinformatics*, 58(4):955–970, 2005.

[75] Abiodun Modupe, Turgay Celik, Vukosi Marivate, and Oludayo O Olugbara. Post-authorship attribution using regularized deep neural network. *Applied Sciences*, 12(15):7518, 2022.

[76] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[77] Luis Espinosa-Anke, Geraint Palmer, Padraig Corcoran, Maxim Filimonov, Irena Spasić, and Dawn Knight. English–welsh cross-lingual embeddings. *Applied Sciences*, 11(14):6541, 2021.

[78] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2019.

[79] Evangelia Gogoulou, Ariel Ekgren, Tim Isbister, and Magnus Sahlgren. Cross-lingual transfer of monolingual models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 948–955, 2022.

[80] James L McClelland, David E Rumelhart, and Geoffrey E Hinton. The appeal of parallel distributed processing. *MIT Press, Cambridge MA*, pages 3–44, 1986.

[81] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[82] Tomáš Mikolov. *Language modeling for speech recognition in czech*. PhD thesis, Masters thesis, Brno University of Technology, 2007.

[83] Tomas Mikolov, Jiri Kopecky, Lukas Burget, Ondrej Glembek, et al. Neural network based language models for highly inflective languages. In *2009 IEEE international conference on acoustics, speech and signal processing*, pages 4725–4728. IEEE, 2009.

[84] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13, 2000.

[85] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[86] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

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

[88] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[89] Alisa Zhila, Wen-tau Yih, Christopher Meek, Geoffrey Zweig, and Tomáš Mikolov. Combining heterogeneous models for measuring relational similarity. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1000–1009, 2013.

[90] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, 2015.

[91] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Zong C, Strube M, editors. 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); 2015 Jul 26-31;*

Beijing, China. Stroudsburg (PA): Association for Computational Linguistics; 2015. p. 270-80. ACL (Association for Computational Linguistics), 2015.

[92] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*, 2016.

[93] Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054, 2015.

[94] Antonio Valerio Miceli Barone. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *arXiv e-prints*, pages arXiv–1608, 2016.

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

[96] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, 2017.

[97] Yedid Hoshen and Lior Wolf. Non-adversarial unsupervised word translation. *arXiv preprint arXiv:1801.06126*, 2018.

[98] David Alvarez-Melis and Tommi S Jaakkola. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

[99] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. *arXiv preprint arXiv:1804.07745*, 2018.

[100] Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. A discriminative latent-variable model for bilingual lexicon induction. *arXiv preprint arXiv:1808.09334*, 2018.

[101] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

[102] Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: Hlt*, pages 771–779, 2008.

[103] Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *arXiv preprint arXiv:1902.00508*, 2019.

[104] Carolin Müller-Spitzer, Sascha Wolfer, and Alexander Koplenig. Observing online dictionary users: Studies using wiktionary log files. *International Journal of Lexicography*, 28(1):1–26, 2015.

[105] Fitrotul Maulidiyah. To use or not to use google translate. *Jurnal Linguistik Terapan*, pages 1–6, 2018.

[106] Stephan Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *HLT-NAACL*, pages 1386–1390, 2015.

[107] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1285–1295, 2016.

[108] Ivan Vulić and Marie-Francine Moens. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994, 2016.

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

[110] Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*, 2014.

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

[112] Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. An autoencoder approach to learning bilingual word representations. *Advances in neural information processing systems*, 27, 2014.

[113] Kejun Huang, Matt Gardner, Evangelos Papalexakis, Christos Faloutsos, Nikos Sidiropoulos, Tom Mitchell, Partha Talukdar, and Xiao Fu. Translation invariant word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1084–1088, 2015.

[114] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[115] Anders Søgaard and Barbara Plank Bernd Bohnet. Inverted indexing for cross-lingual nlp. 2015.

[116] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. *Advances in neural information processing systems*, 28, 2015.

[117] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, 2017.

[118] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174, 2018.

[119] Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. Learning general purpose distributed sentence representations via large scale multi-task learning, 2018. URL https://arxiv.org/abs/1804.00079.

[120] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

[121] Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. 2006.

[122] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 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, 2015.

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

[124] Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. Trans-gram, fast cross-lingual word-embeddings. *arXiv preprint arXiv:1601.02502*, 2016.

[125] Yogarshi Vyas and Marine Carpuat. Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1187–1197, 2016.

[126] Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 567–572, 2015.

[127] Aditya Mogadala and Achim Rettinger. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 692–702, 2016.

[128] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[129] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

[130] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[131] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, 2019.

[132] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, 2019.

[133] Shijie Wu and Mark Dredze. Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, 2020.

[134] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*, 2020.

[135] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

[136] Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. Zero-shot cross-lingual transfer with meta learning. *arXiv preprint arXiv:2003.02739*, 2020.

[137] Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*, 2018.

[138] Wietse de Vries and Malvina Nissim. As good as new. how to successfully recycle english gpt-2 to make models for other languages. *arXiv preprint arXiv:2012.05628*, 2020.

[139] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[140] Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. What's so special about bert's layers? a closer look at the nlp pipeline in monolingual and multilingual models. *arXiv preprint arXiv:2004.06499*, 2020.

[141] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020. URL https://arxiv.org/abs/2003.10555.

[142] Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Türe, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 7–12, 2010.

[143] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47, 2014.

[144] Minh-Thang Luong, Richard Socher, and Christopher D Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the seventeenth conference on computational natural language learning*, pages 104–113, 2013.

[145] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014.

[146] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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)*, pages 451–462, 2017.

[147] Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, 2013.

[148] Holger Schwenk and Xian Li. A corpus for multilingual document classification in eight languages. *arXiv preprint arXiv:1805.09821*, 2018.

[149] Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. Overview of the second bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, 2017.

[150] Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th workshop on building and using comparable corpora*, pages 39–42, 2018.

[151] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86, 2005.

[152] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[153] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*, 2013.

[154] Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1118–1127, 2010.

[155] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

[156] Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, et al. Universal dependencies 2.1. 2017.

[157] Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for ner. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, 2019.

[158] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang. Enhancing and combining sequential and tree lstm for natural language inference. *arXiv preprint arXiv:1609.06038*, 2016.

[159] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, 2020.

[160] Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, Marco Guerini, Aptus AI, and Fondazione Bruno Kessler. Geppetto carves italian into a language model. *Computational Linguistics CLiC-it 2020*, page 136, 2020.

[161] Roeland Ordelman, Franciska de Jong, Arjan Van Hessen, and Hendri Hondorp. Twnc: a multifaceted dutch news corpus. *ELRA Newsletter*, 12(3/4):4–7, 2007.

[162] Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. The construction of a 500-million-word reference corpus of contemporary written dutch. *Essential speech and language technology for Dutch: Results by the STEVIN programme*, pages 219–247, 2013.

[163] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, 2017.

[164] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

[165] Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, et al. Universal dependencies 2.2. 2018.

[166] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.

[167] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019.

[168] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*, 2020.

[169] Ryan McDonald and Joakim Nivre. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 122–131, 2007.

[170] Yue Zhang and Joakim Nivre. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 188–193, 2011.

[171] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

[172] Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. Cross-lingual transfer parsing for low-resourced languages: An irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, pages 41–49, 2014.

[173] P Kalarani and S Selva Brunda. Sentiment analysis by pos and joint sentiment topic features using svm and ann. *Soft Computing*, 23(16):7067–7079, 2019.

[174] Kefei Cheng, Yanan Yue, and Zhiwen Song. Sentiment classification based on part-of-speech and self-attention mechanism. *IEEE Access*, 8:16387–16396, 2020.

[175] Marek Rei. Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2121–2130, 2017.

[176] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570, 2018.

[177] Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. Toward more meaningful resources for lower-resourced languages. *arXiv preprint arXiv:2202.12288*, 2022.

[178] Rubungo Andre Niyongabo, Hong Qu, Julia Kreutzer, and Li Huang. Kinnews and kirnews: Benchmarking cross-lingual text classification for kinyarwanda and kirundi. *arXiv preprint arXiv:2010.12174*, 2020.

[179] Marko Robnik-Sikonja, Kristjan Reba, and Igor Mozetic. Cross-lingual transfer of sentiment classifiers. *arXiv preprint arXiv:2005.07456*, 2020.

[180] Toan Q Nguyen and David Chiang. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, 2017.

[181] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[182] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[183] Ivan Vulic and Anna-Leena Korhonen. On the role of seed lexicons in learning bilingual word embeddings. 2016.

[184] Razvan Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In *11th conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, 2006.

[185] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1375–1384, 2011.

[186] Xiao Cheng and Dan Roth. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, 2013.

[187] Chen-Tse Tsai and Dan Roth. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, 2016.

[188] Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138*, 2019.

[189] Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. An empirical study of incorporating pseudo data into grammatical error correction. *arXiv preprint arXiv:1909.00502*, 2019.

[190] Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. *arXiv preprint arXiv:2005.00987*, 2020.

[191] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*, 2018.

[192] Tao Tu, Yuan-Jui Chen, Cheng-chieh Yeh, and Hung-Yi Lee. End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508*, 2019.

[193] Amit Das and Mark Hasegawa-Johnson. Cross-lingual transfer learning during supervised training in low resource scenarios. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[194] Ryan Cotterell and Georg Heigold. Cross-lingual, character-level neural morphological tagging. *arXiv preprint arXiv:1708.09157*, 2017.

[195] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[196] Mikel Artetxe, Gorka Labaka, and Eneko 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*, pages 2289–2294, 2016.

[197] Ivan Vulic and Marie-Francine Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, volume 2, pages 719–725. ACL; East Stroudsburg, PA, 2015.

[198] Jose Camacho-Collados, Yerai Doval, Eugenio Martínez-Cámara, Luis Espinosa-Anke, Francesco Barbieri, and Steven Schockaert. Learning cross-lingual word embeddings from twitter via distant supervision. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 72–82, 2020.

[199] Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 294–304, 2018.

[200] Siqi Chen, Yijie Pei, Zunwang Ke, and Wushour Silamu. Low-resource named entity recognition via the pre-training model. *Symmetry*, 13(5):786, 2021.

[201] Kamil Kanclerz, Piotr Miłkowski, and Jan Kocoń. Cross-lingual deep neural transfer learning in sentiment analysis. *Procedia Computer Science*, 176:128–137, 2020.

[202] Shujah Ur Rehman, Bilal Tahir, and Muhammad Amir Mehmood. Investigating cross-lingual transfer learning techniques for urdu text using word embeddings. In *2021 15th International Conference on Open Source Systems and Technologies (ICOSST)*, pages 1–6. IEEE, 2021.

[203] Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna Kumar K R, and Chitra Viswanathan. Anvita-african: A multilingual neural machine translation system for african languages. In *Proceedings of the Seventh Conference on Machine Translation*, pages 1090–1097, Abu Dhabi, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.106.

[204] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

[205] Roald Eiselen and Martin Puttkammer. Developing text resources for ten south african languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3698–3703, 2014.

[206] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation*

Conference, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494.

[207] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.

[208] Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*, 2020.

[209] Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, et al. Xtreme-s: Evaluating cross-lingual speech representations. *arXiv preprint arXiv:2203.10752*, 2022.

[210] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, 2012.

[211] James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. No language left behind: Scaling human-centered machine translation. 2022.

[212] Tomáš Mikolov, Édouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[213] Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. *arXiv preprint arXiv:2003.04986*, 2020.

[214] Anna Mosolova, Ivan Bondarenko, and Vadim Fomin. Conditional random fields for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 121–123, 2018.

[215] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[216] James Fowkes. Founding provisions. *Constitutional law of*, 2014.

[217] SABC News. Sign language to become SA's 12th official language - SABC News - Breaking news, special reports, world, business, sport coverage of all South African current events. Africa's news leader., July 2023. URL https://www.sabcnews.com/sabcnews/sign-language-to-become-sas-12th-official-language/.

[218] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*, 2019.

[219] Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947, 2017.

[220] Jan Buys and Jan A Botha. Cross-lingual morphological tagging for low-resource languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1954–1964, 2016.

[221] Ronald Cardenas, Ying Lin, Heng Ji, and Jonathan May. A grounded unsupervised universal part-of-speech tagger for low-resource languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2428–2439, 2019.

[222] Yingting Wu, Hai Zhao, and Jia-Jun Tong. Multilingual universal dependency parsing from raw text with low-resource language enhancement. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 74–80, 2018.

[223] Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 62–72, 2011.

[224] David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, et al. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, 2022.

[225] University of Pretoria. Open educational resource term bank. URL https://www.up.ac.za/african-languages/news/post_2728581-open-educational-resource-term-bank-pg2.

[226] Abiodun Modupe, Thapelo Sindane, and Vukosi Marivate. Zero-shot transfer learning using affix and correlated cross-lingual embeddings. 2023.

[227] Christian Federmann, Tom Kocmi, and Ying Xin. Ntrex-128–news test references for mt evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, 2022.

[228] David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, sana al azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Tshinu Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. Masakhanews: News topic classification for african languages, 2023. URL https://arxiv.org/abs/2304.09972.

[229] Andani Madodonga, Vukosi Marivate, and Matthew Adendorff. Izindaba-Tindzaba: Machine learning news categorisation for Long and Short Text for isiZulu and Siswati. *Dhasa*, 4, Jan 2023. doi: 10.55492/dhasa.v4i01.4449. URL https://upjournals.up.ac.za/index.php/dhasa/article/view/4449.

[230] Vukosi Marivate, Moseli Mots'Oehli, Valencia Wagner, Richard Lastrucci, and Isheanesu Dzingirai. Puoberta: Training and evaluation of a curated language model for setswana. In *SACAIR 2023 (To Appear)*, 2023.

[231] Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. Unsupervised cross-lingual transfer of word embedding spaces. *arXiv preprint arXiv:1809.03633*, 2018.