2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)

2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom) | 978-1-6654-1485-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISPA-BDCLOUD-SOCIALCOM-SUSTAINCOM51426.2020.00157

Twi Corpus: A Massively Twi-to-Handful Languages Parallel Bible Corpus

Michael Adjeisah
School of Computer Science and Technology
Donghua University
Shanghai, China
madjeisah@hotmail.com

Guohua Liu
School of Computer Science and Technology
Donghua University
Shanghai, China
ghliu@dhu.edu.cn

Richard Nuetey Nortey
School of Information Science and Technology
Donghua University
Shanghai, China
rn.nortey@yahoo.com

Jinling Song*
School of Mathematics and Information Technology
Hebei Normal University of Science & Technology
Qinhuangdao, Hebei, China
songjinling99@126.com

Khalid Odartey Lamptey
School of Computer Science and Technology
Donghua University
Shanghai, China
khaliddraig@yahoo.com

Felix Nana Frimpong
School of Computer Science and Technology
Donghua University
Shanghai, China
frimpong.felix@yahoo.com

***Abstract*— This paper presents detailed modeling of massively parallel Bible corpus based on Twi, a common Ghanaian language, to a handful of languages. We discussed some of the common issues we encountered in obtaining, processing, converting, and formatting the corpus and the latent desire for success in NLP. We stored the sentence aligned data in various files based on the Twi to the selected language pairs with a tab-delimited separation. Verses with the same line number in a line pair are mappings of each other. It is often challenging to learn what a "clean" corpus looks like in lower-resource situations, especially where the target corpus is the only sample of the language's parallel text. We, therefore, performed unsupervised measurements on each sentence pair. We engage the squared Mahalanobis distances that predicted parallelism on the dataset. Eventually, we perform a statistical analysis of the corpora collected based on selected text categorization models for text classification by leveraging vector embedding (like Word2vec). Finally, we trained the Twi vocabs for a 2D representation. Similar words find their vectors closer by engaging t-Distributed Stochastic Neighbor embedding (t-SNE).**

Keywords—Natural language processing; Sentence-alignment; Twi corpus; Word embedding; Mahalanobis distances

I. INTRODUCTION

English is an official language for most African countries like Botswana, Ghana, Kenya, Nigerian, South Africa, and part of the francophone countries. It means that it is the dominant language used in businesses, education, and government. It serves as a medium through which we impart and evaluate knowledge. Authors, Webb [1] and Kamwangamalu [2], in their paper, argued that using English as a mode of teaching and learning instead of one's mother-tongue. It affects students' ability to learn and perform undeviatingly. There is a need for people worldwide to use their language to understand better, especially when using computers or accessing information on the Internet. While this task seems simple, it requires various applications, including software like local language spell-checkers, word processors, machine translation systems, and search engines. Simultaneously, the amount of work required to develop all natural language processing (NLP) aspects for a new language is enormous as far as Machine Translation (MT) is concerned.

Solving MT problem with corpus statistical and neural techniques is a rapidly growing field that is leading to better translation. The most relevant resource for linguistic research and NLP applications is the availability of parallel corpora: text paired with its translated version in a second language. A typical use of the bottommost is as training data for statistical machine translation (SMT) like Giza++ [3], *Moses* [4], and neural machine translation (NMT) like RNN [5], Convolutional Seq2Seq [6], [7], Transformer [8]. It applies customarily large amounts of aligned text to learn word alignment models in the lexica of two languages. Even though some systems do not necessitate text alignment [9], the most successful models use supervised sentence-aligned corpora, from a resource-rich language. It tutors the unsupervised learning algorithm in a target language.

Most parallel corpora exist in common language pairs like the English-French, Chinese, Danish, and German. There are, however, a few language pairs, especially languages in African. Meanwhile, to perform MT from the less popular language to the famous ones, access to the parallel corpus in a very widely translated text is needed and, from many diverse languages for multilingual translation. The acquisition of the Twi to a handful of languages (English, Chinese, German, French, and Spanish) parallel corpus, which we describe in this paper, is one keen contribution to this endeavor. It is the harvest of the Twi Bible and its translated version. The built corpus comprises about 30 million words for each of the eight selected official languages of the Bible: Chinese (zh), English (en), French (fr), German (de), Spanish (es), and the original written language, Greek (el) and Hebrew (he) [10]. The parallel corpus is freely available for use[1].

Most multilingual corpus to our knowledge, are currently available in the Bible in 100 languages [11], the OPUS collection [12], [13] which contains 90 languages in

* This is to indicate the corresponding author.

[1] https://github.com/Madjeisah/tw-parallel-lg-corpus

978-1-6654-1485-2/20/$31.00 ©2020 IEEE
DOI 10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00157

Authorized licensed use limited to: Zhejiang Normal University. Downloaded on November 25,2022 at 08:59:37 UTC from IEEE Xplore. Restrictions apply.

various parallel corpora, the Europarl [14], the WMT corpus [15], and other, non-parallel but comparable corpora (like Wikipedia with more than 10,000 articles in 121 languages) [9]. However, comparatively few of the possible language pairs are available with sentence-aligned in African languages. Recently, Zeljko [16] introduced JW300, a parallel corpus of over 300 languages with around 100 thousand parallel sentences per language pair on average, which include the Twi language. Nonetheless, statistics were performed on common language pairs and also their focus was on popular languages.

II. Overall Architecture

Our system shares the same skeleton system architecture as the NRC supervised and unsupervised submissions [17] and parallelism of sentence pairs estimation [18]. The system consists of initial filtering to eliminate accessible noise and prevent selections from an extensive collection of short sentences. Finally, feature scoring for measuring parallelism, based on ratios of squared Mahalanobis distances[2], was engaged

A. Overview of the Mahalanobis Ratios

We briefly introduced the Mahalanobis distance for the parallelism assessment proposed by Littell [18]. The probability of a draw from a univariate normal distribution can be related to its distance to the mean in terms of standard deviations (the *z*- score). Notwithstanding, measuring the Euclidean distance to the mean can lead to inaccurate conclusions, especially in a multivariate normal distribution. Preferably, the Mahalanobis distance square [19] is the appropriate measurement for correlating distance to the probability of a vector x from distribution X with correlation Σ and mean μ:

$$d^2(x) = (x-\mu)^T \sum^{-1} (x-\mu) \tag{1}$$

It is similar to dissociating and rescaling to unit variance in all dimensions, through correlation matrix's transposed square root. It finally measures the squared Euclidean distance to the mean in the outcome space.

$$\begin{aligned} d^2(x) &= (x-\mu)^T \sum^{-\frac{1}{2}} \sum^{-\frac{1}{2}} (x-\mu) \\ &= \left(\sum^{-\frac{1}{2}} (x-\mu)\right)^T + \left(\sum^{-\frac{1}{2}} (x-\mu)\right) \\ &= || \sum^{-\frac{1}{2}} (x-\mu) ||_2^2 \end{aligned} \tag{2}$$

Authors compared a figure distribution side by side in their paper with an observation on a distribution transformed by $\sum^{\frac{1}{2}}$. It further proved that the squared magnitudes could be used to compute the probabilities. However, in practice, the probabilities were so related in higher dimensional spaces as to be identical. There remains the possibility, however, that the magnitudes themselves remain sufficiently informative [18].

In computing and presentation, the center distribution has zero means which transform the resulting matrix by $\sum^{\frac{1}{2}}$ especially on high-dimensional vectors of a trained mono-lingual of the Twi and the handful-language sentences. We were able to directly concatenating their vectors with regards to their joint distribution.

We also consider three vectors in transformed space based on each sentence vector pair $\langle l_1, l_2 \rangle$ after re-centering as proposed in the original paper.

- the vector e_1 corresponding only to l_2's contribution to the concatenated and transformed vector (as if $l_2 = \vec{0}$)
- the vector e_2 corresponding only to l_2's contribution (as if $l_1 = \vec{0}$)
- the vector e corresponding to the transformation of the concatenation of l_1 and l_2'

$$\begin{aligned} e_1 &= \sum^{-\frac{1}{2}} (l_1, \vec{0}) \\ e_2 &= \sum^{-\frac{1}{2}} (\vec{0}, l_2) \\ e &= \sum^{-\frac{1}{2}} (l_1, l_2) = e_1 + e_2 \end{aligned} \tag{3}$$

The estimation m we are interested in the squared magnitude of the combined vector, divided by the sum of the squared magnitudes of only e_1 and e_2.

$$m = \frac{||e||_2^2}{||e_1||_2^2 + ||e_2||_2^2} \tag{4}$$

Results of these experiments are presented in section IV. We also performed sanity check on random sentence pairs from the target corpus, according to whether we judged them to be parallel or not.

III. Acquiring and Converting Corpus

Koehn [20], stated that the acquisition of a parallel corpus for the MT system use typically takes five steps and is further discussed in detail. Our system shares the same general steps for the Twi corpus collection. It differs primarily in the crawling, as we used a different tool for mining and sentence-alignment (section *B*). Having identified our potential sources for parallel text, we develop a python script using the BeatifulSoup HTML. A parsing library to crawl and harvest additional text from HTML tags. This approach turned out to be an efficient approach for our purpose because the first step generated equally human-readable text files. It makes cleaning effort required a relatively small amount of time for each language. We scraped the Bible verses for each language until we obtained about 30 million words and about 31 to 130K sentence pairs.

A. Data Collection

Machine-readable versions of the Holy Bible in multiple languages is accessible on websites to the public domain. Example sources are the Bible Gateway website, GospelGo, the YouVersion, and the Unbound Bible. Each one offered

[2] https://github.com/aboSamoor/pycld2

Authorized licensed use limited to: Zhejiang Normal University. Downloaded on November 25,2022 at 08:59:37 UTC from IEEE Xplore. Restrictions apply.

the Bible in different formats, some containing HTML and others in plain text. We finally concluded to harvest the entire corpus from the Version website. It is enough and contains the Bible in multiple languages and different versions, as shown in Table I. We scrapped four versions for English, French, and Chinese even though there are many versions of about 61, 14, and 6, respectively. Two versions of German and Spanish and a single version of the Twi (new language in NLP) [10], Greek, and Hebrew.

TABLE I. DISPLAYS LANGUAGE PAIRS IN TWI TO SELECTED HANDFUL-LANGUAGES

Language pairs	Versions
Mfitiaseɛ no Onyankopɔn bɔɔ ɔsoro ne asase.	***twi-eng***
	In the beginning, God created heaven and earth. In the beginning, when God created the universe, In the beginning, when God created the earth and sky. In the beginning, God created the heavens and the earth.
	twi-zh
	起初 神 創造 天地 起 初 神 創 造 天 地 元始 上帝創造天地 元 始 上 帝 創 造 天 地
	twi-fr
	Au commencement Dieu créa le ciel et la terre Au commencement, Dieu créa le ciel et la terre. Au commencement DIEU créa les cieux et la terre Au commencement Dieu créa les cieux et la terre
	twi-de
	Am Anfang schuf Gott Himmel und Erde. Am Anfang schuf Gott Himmel und Erde.
	twi-es
	En el principio creó Dios los cielos y la tierra En el principio creó Dios los cielos y la tierra
	twi-el
	Εν αρχη εποιησεν ο Θεος τον ουρανον και την γην.
	twi-he
	הארץ׃ ואת השמים את אלהים ברא בראשית

B. *Sentence Alignment*

Sentence alignment is usually complicated when mining text for MT work. Fortunately for us, it is simplified because each Bible is formatted as an XML file, with elements corresponding to books and chapters, and or <p> elements corresponding to verses. Each verse is marked with class "label" and the verse's content as "content" with the verse on the same line except for verses that require referencing from other verses. Sentence alignment becomes more flexible since each verse corresponds to its number, as shown in Figure 1.

Fig. 1. Easy harvesting and processing for sentence alignment

An inconsistency in the online sources' formatting was the major problem encountered during the corpus's conversion and formatting. Mostly incorrect HTML: an unclosed <p> or tags, multiple uses of without class or ID, use of for referencing which ends up sifting part of the verse to the next line, errors in the numbering of verses (notably missing ones). Also, some verses were just stated as "ref" (reference) to the previous verse instead of the text, especially in Chinese. Fortunately, systematic errors can be corrected by finding multiple sources of the same translation. If neither option was available, we do a human translation. The English Bible's amplified version is more complicated to mine and process. It comes with a bunch of referencing and dictionary definitions of some keywords. E.g.,

"1In the beginning God (Elohim) created [by forming from nothing] the heavens and the earth."

Authorized licensed use limited to: Zhejiang Normal University. Downloaded on November 25,2022 at 08:59:37 UTC from IEEE Xplore. Restrictions apply.

"6And God said, "Let there be an expanse [of the sky] in the midst of the waters, and let it separate the waters [below the expanse] from the waters [above the expanse]."

The definition of "*created [by forming from nothing]*" as presented in Figure 2, is not vital for our work.

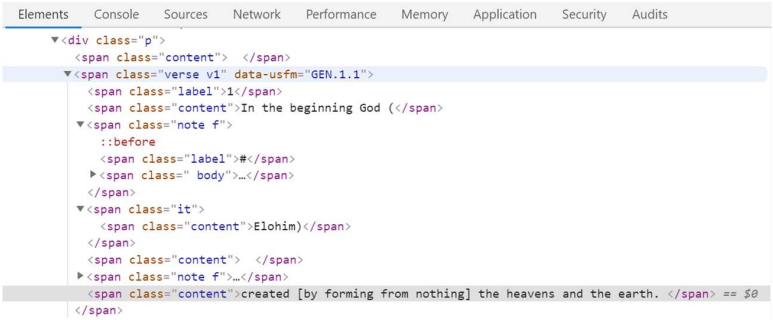

Fig. 2. Source page with an inconsistency HTML format

The processing of these formats comes with a lot of complications due to the inconsistency. Consider processing by deleting/eliminating all words or phrase in the square brackets; the two verses would be;

"1In the beginning God (Elohim) created the heavens and the earth."

"6And God said, "Let there be an expanse in the midst of the waters, and let it separate the waters from the waters."

While verse 1 sounds meaningful, such processing makes verse 6 ambiguous. Hence, we limit the English versions to the King James Version (KJV), Good News Bible (GNB), Easy to Read Version (ERV), and New International Version (NIV).

Overall, the entire process took two persons within three weeks of active concentration. In scraping, we scraped the text in their book's form (e.g., the book of Genesis, Exodus, …, Revelation) and saved separately. It gave room to make sure each verse with the same line number in a file is mappings of each other. We finally merged all the 66 books of the Bible into one file in all the handful languages. We stored the sentence-aligned data in one file based on the Twi to the selected language pairs with a tab-delimited separation. Verses with the same line number in a line pair are mappings of each other. We present statistics of the released corpus in Table II below. The number of words and sentences presented is after sentence-alignment with Twi to the listed handful of languages and the maximum length of sentences.

TABLE II. STATISTICS OF THE RELEASED CORPUS

Language	Vocabs	Sentences	Max Length
Twi	23162	31,100	93
English (en)	17979	124,400	118
Chinese (zh)	52780	124,400	96
French (fr)	22042	124,400	91
German (de)	19645	62,200	83
Greek (el)	37284	31,100	56
Hebrew (he)	55287	31,100	49
Spanish (es),	27077	62200	85

C. The Akan Language (Twi)

The Akan Language, known as Twi, is a spoken language in Ghana's southern and central parts by several people, mainly of the Akan tribe. It is the biggest of about 18 major tribes in Ghana and forms about 70% of the population as a first and second language. Twi is a common name for the Akan language's two ancient literary dialects; Asante (Ashanti) and Akuapem, which are mutually intelligible. There are over 9 million Twi speakers, mainly originating from the Ashanti Region. About 17–18 million Ghanaians are either first or second Twi language speakers. The Twi alphabet contains 22 letters, as shown in Table III. Letters C, J, V, and Z are also used, but only in loanwords.

Both SMT and NMT in recent times have demonstrated groundbreaking results in many MT tasks. However, comparatively few of the possible language pairs are available with sentence-aligned in popular languages like English, French, Chinese, and some African languages but none in Twi. It eventually energizes the research into such a language pair.

TABLE III. THE TWI ALPHABET

Majuscule forms (uppercase or capital letters)																				
A	B	D	E	Ɛ	F	G	H	I	K	L	M	N	O	Ɔ	P	R	S	T	U	W	Y
Minuscule forms (lowercase or small letters)																				
a	b	d	e	ɛ	f	g	h	i	k	l	m	n	o	ɔ	p	r	s	t	u	w	y

IV. EXPERIMENTAL RESULTS

Ultimately, we began by training monolingual sentence embedding using sent2vec [21], on all available corpus. We also implement an analysis of the corpora collected based on selected text categorization models for text classification by leveraging vector embedding (word2vec). We finally illustrate a 2D representation of 10% words of the Twi corpus.

A. Mahalanobis Distances Results

We test the measurements of the sentence vectors purely on the English-Twi dataset. Two sets of random normal

Authorized licensed use limited to: Zhejiang Normal University. Downloaded on November 25,2022 at 08:59:37 UTC from IEEE Xplore. Restrictions apply.

vectors *L1* and *L2*, were engaged. We estimate that some proportion *p* of vectors in *L1* corresponded to *L2* by a linear transformation *T* and checked some proportion of vectors that did not correspond to each other. Finally, we engaged some Gaussian noise σ to each of *L1* and *L2* making this transformation not ideal as most real-world data come with noise. We then diversified the proportion of "true" pairs, and additive noise, to examine the robustness of these measurements in various noise conditions. Results were compared with linear distribution, as shown in Table IV.

TABLE IV. ACCURACY OF DISTINGUISHING PARALLEL BASED ON THE PROPORTION OF PARALLEL PAIRS IN THE DATASET

p	0.1	0.2	0.3	0.4	0.5
Mahalanobis	0.984	0.981	0.978	0.977	0.977
Linear	0.948	0.947	0.944	0.932	0.924

The table above is the accuracy of distinguishing parallel associated with a translation matrix *T* against non-parallel vectors from the English-Twi dataset. It contains 124K pairs of 50-dimensional vectors, with standard normal additive noise. As stated, the *p* represents the proportion of parallel pairs in the dataset. Next in line (Table V and VI) is the accuracy of distinguishing parallel associated with a translation matrix *T* against non-parallel vectors with the same 50-dimensional vectors and "true" proportion $p = 0.1$ and 0.3 (random selection), with varying degrees of additive noise. Also, as mentioned, σ represents the standard deviation of the additive noise added to each of *L1* and *L2*.

TABLE V. ACCURACY OF DISTINGUISHING PARALLEL BASED ON PROPORTION $P = 0.1$ AND NOISE ADDITIVE σ

σ	1.0	2.0	3.0	4.0	5.0
Mahalanobis	0.984	0.826	0.775	0.643	0.610
Linear	0. 951	0. 791	0. 689	0. 639	0. 598

TABLE VI. ACCURACY OF DISTINGUISHING PARALLEL BASED ON PROPORTION $P = 0.3$

σ	0.1	0.2	0.3	0.4	0.5
Mahalanobis	0.978	0. 798	0. 675	0. 619	0.598
Linear	0.926	0.734	0.670	0.607	0.590

We also perform sanity checking (Table VII) on 1000 annotated sentences for the Mahalanobis estimation on various dimensions. From 10, 50, 100, 300, 500, and 1000-dimensional sentence vectors. It is unclear why 100-dimensional vectors perform more badly than 50-dimensional and 500-dimensional dropped a 0.003 accuracy after 300-dimensional vectors. We presumed that an increase in dimension would increase accuracy. However, we used 100 samples of the dataset, and accuracy on 1000-dimensional vectors is encouraging. We do not want to emphasize much stock in the results.

TABLE VII. SANITY CHECK RESULTS ON 1000 ANNOTATED SENTENCES

Dimensionality	10	50	100	300	500	1000
Accuracy	0.467	0.562	0.491	0.658	0.655	0.746

B. Word embedding

The most common methods to convert sentences to machine-readable code are TF-IDF and Word Embedding [22] are the backbone in performing efficient NLP models. The algorithms require the input features as a fixed-length feature vector. It is where word embedding (word2vec) [21] comes to play. It is mapping a text or words to real value fixed-size vectors or converting text into semantic vectors. The mapping preserves the semantic relationships so that the distance between vectors corresponds to the meaning of words. While word embedding presents a multi-dimensional vector which attempts to capture the relationship of a word to another word [22]–[24], the TF-IDF matrix presents a sparse matrix where each word maps to projects a single value and capture no meaning.

To understand the corpus's semantic relationship, we implement an embedding vectorizer model to feed into a text classifier. We throw in a version of the classifiers like NB, linear kernel SVM, and word2vec-based Extra Trees that use the TF-IDF weighting scheme for good measure. In other terms, to extend the word vectors and generate document level vectors, we fed the vectors into the listed classifiers. We used an average of all the words in the document. Because the word2vec algorithm for learning a vector representation for each word [25], [26] is, however, not enough to be used as features for text classification.

To allow a comparison of the classifiers, we define a standard training and test set. We suggest to split and use 20% as a test set, and the rest of the corpus as training data for this particular work. We benchmarked the model's results on 5-fold cross-validation on the dataset of about 248K. Each version contains 31100 lines of text and labeled with eight languages as categorical variables. For the scores for the models, see Table VIII in descending order.

TABLE VIII. OVERALL RESULTS OF 5-FOLD CROSS VALIDATION ON THE ENTIRE DATASET BASED ON THE CLASSIFIERS

model	score
w2v + tfidf	0.9956
svc + tfidf	0.9911
svc	0.9875
w2v	0.9875
mult_nb	0.9697
mult_nb + tfidf	0.9692
bern_nb	0.9428
bern_nb + tfidf	0.9416

From the table above, all models performed tremendously on the corpus with a score of 0.9416 and 0.9956. The TF-IDF Word2vec-based Extra Trees classifier is leading, and all versions of SVM as the close second.

The standard version of the Word2vec-based Extra Trees classifier is not far behind, and lastly, Naive Bayes both Bernoulli (bern_nb) and Multinomial (mult_nb). We also check the model's performance depending on the number of labeled training examples of Twi to the handful of language pairs. It helps check the ranking depending on the amount of training data on all corpus for clues into the challenges ahead.

Authorized licensed use limited to: Zhejiang Normal University. Downloaded on November 25,2022 at 08:59:37 UTC from IEEE Xplore. Restrictions apply.

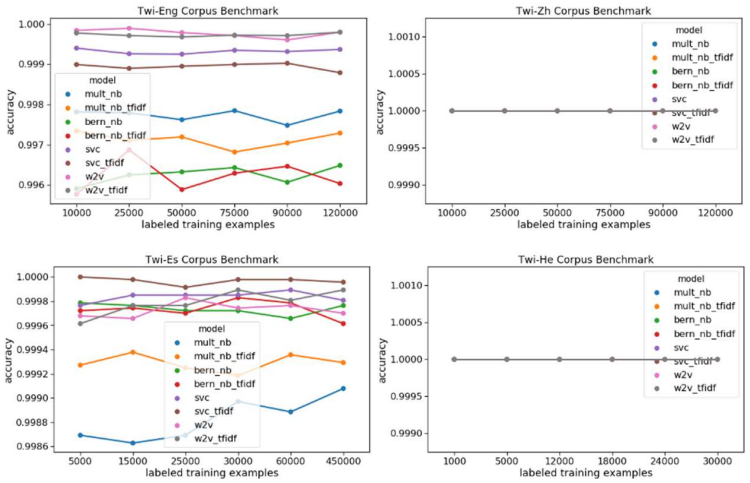

Fig. 3. Model performance on Twi-to-handful language Corpus Benchmark based on labeled training examples

We can see form Figure 3 that versions of word2vec performed well on the "*Tw-En*" corpus with SVM versions as a close second, and vice versa on the "*Tw-Es*" corpus benchmark. All models performed equally on "*Tw-Zh*" and "*Tw-He*" corpus benchmarks with scores of 1.000. It is possible because Twi is an analytic language, and Chinese and Hebrew are morphologically languages, and neither has clear word relations. Perhaps surprisingly, the best results are obtained not by the word embedding method but by the corpus's sparsity. We also measured the inter-annotator agreement on the developed corpus's correctness but has no significant effect on the data. *Cohen's kappa* recorded 0.321 and 0.294 on *tw-en* and *tw-fr*, respectively, which interpretation can be fair agreement. We finally trained the Twi vocabs with 200k training steps for a 2D representation where similar words find their vectors closer by engaging t-SNE [27], [28]. The first 10% vocabs of the corpus are represented in Figure 4. Semantically similar words occur close to each other.

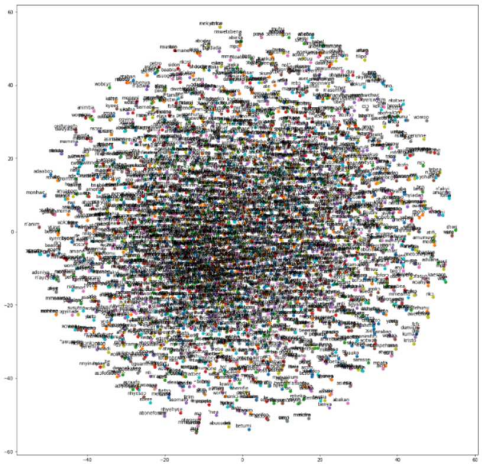

Fig. 4. 2D representation of about 3k words of the Twi Bible corpus.

V. Conclusion

We described the modeling of a massively parallel Bible corpus based on Twi to a handful of languages. We stored the readily processed sentence aligned data for MT work in various files based on the selected language pairs with a tab-delimited separation. It is often perplexing to learn what a good corpus looks like in lower-resource situations, specifically where the target corpus is the only sample of the language's parallel text. We, therefore, performed unsupervised measurements on each sentence pair. We engage the squared Mahalanobis distances that predicted parallelism on the dataset. The experiments show that Mahalanobis offers top performance in both mono and cross-lingual word embedding. The process also shows that two sentences with sufficiently similar topics are considered parallel for the highest-quality corpus pairs in lower-resource situations. To understand the corpus's semantic relationship, we implement an embedding vectorizer model to feed into a text classifier. We also engaged a version that uses the TF-IDF weighting scheme. Finally, we trained the Twi vocabs for a 2D representation. Similar words find their vectors closer by engaging t-SNE, revealing insightful clues into the challenges ahead.

Future work is to engage the state-of-the-art SMT and NMT for translation from Twi to a handful of languages and vice versa. We hope that the corpora's availability would yield an exciting and productive impact in the field of NLP.

Acknowledgment

This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant 17YF1427400, in part by the Fundamental Research Funds for the Central Universities under Grant 17D111206, in part by the Research on Social Sciences Development in Hebei Province under Grant 20200302075, and in part by the Marine Science Research Project of Hebei Normal University of Science & Technology under Grant 2018HY020.

References

[1] V. Webb, "English as a second language in South Africa's tertiary institutions: a case study at the University of Pretoria," World Englishes, vol. 21, no. 1, pp. 49–61, Mar. 2002.

[2] N. M. Kamwangamalu, "Second and Foreign Language Learning in South Africa," in Encyclopedia of Language and Education, Boston, MA: Springer US, 2008, pp. 1280–1292.

[3] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," Comput. Linguist., vol. 29, no. 1, pp. 19–51, Mar. 2003.

Authorized licensed use limited to: Zhejiang Normal University. Downloaded on November 25,2022 at 08:59:37 UTC from IEEE Xplore. Restrictions apply.

[4] H. Hoang and P. Koehn, “Design of the Moses Decoder for Statistical Machine Translation,” in Software Engineering, Testing, and Quality Assurance for Natural Language Processing, Columbus, Ohio: Association for Computational Linguistics, 2008, pp. 58–65.

[5] S. K. Mahata, D. Das, and S. Bandyopadhyay, “MTIL2017: Machine Translation Using Recurrent Neural Network on Statistical Machine Translation,” J. Intell. Syst., vol. 28, no. 3, pp. 447–453, May 2018.

[6] J. Gehring and Y. N. Dauphin, “Conv Seq2Seq,” vol. 1001, no. 1, pp. 160–167, 2016.

[7] J. Gehring, M. Auli, D. Grangier, and Y. N. Dauphin, “A convolutional encoder model for neural machine translation,” in ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2017, vol. 1, pp. 123–135.

[8] A. Vaswani et al., “Attention Is All You Need,” in Advances in Neural Information Processing Systems 30, I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008.

[9] S. B. Cohen, D. Das, and N. A. Smith, “Unsupervised Structure Prediction with Non-Parallel Multilingual Guidance,” in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 50–61.

[10] M. Adjeisah, “English and Chinese↔Twi Parallel-Aligned Corpus for Encoder-Decoder Based Machine Translation,” Doctoral Thesis, Submited to Donghua University, Shanghai, China.

[11] C. Christodouloupoulos and M. Steedman, “A massively parallel corpus: the Bible in 100 languages,” Lang. Resour. Eval., vol. 49, no. 2, pp. 375–395, Jun. 2015.

[12] J. Tiedemann, “News from OPUS — A collection of multilingual parallel corpora with tools and interfaces,” in Recent Advances in Natural Language Processing Vol. V, Amsterdam/Philadelphia: John Benjamins, 2009, pp. 237–248.

[13] J. Tiedemann, “Parallel Data, Tools and Interfaces in OPUS,” in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), 2012, pp. 2214–2218.

[14] P. Koehn and C. Monz, “Shared Task: Statistical Machine Translation between European Languages,” in Proceedings of the ACL Workshop on Building and Using Parallel Texts, 2005, pp. 119–124.

[15] J. Bradbury and R. Socher, “MetaMind Neural Machine Translation System for WMT 2016,” in Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, 2016, pp. 264–267.

[16] Ž. Agić and I. Vulić, “JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3204–3210.

[17] C. Lo, M. Simard, D. Stewart, S. Larkin, C. Goutte, and P. Littell, “Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The NRC supervised submissions to the Parallel Corpus Filtering task,” 2019, pp. 908–916.

[18] P. Littell, S. Larkin, D. Stewart, M. Simard, C. Goutte, and C.-K. Lo, “Measuring sentence parallelism using Mahalanobis distances: The NRC unsupervised submissions to the WMT18 Parallel Corpus Filtering shared task,” vol. 2, pp. 900–907, 2018.

[19] P. C. Mahalanobis, “On the Generalized Distance in Statistics,” pp. 49–55, 1936.

[20] P. Koehn and P. Koehn, “Europarl: A Parallel Corpus for Statistical Machine Translation,” in Conference Proceedings: the 10th Machine Translation Summit, 2005, pp. 79--86.

[21] M. Pagliardini, P. Gupta, and M. Jaggi, “Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 528–540.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings, 2013.

[23] T. Shi and Z. Liu, “Linking GloVe with word2vec,” arXiv Prepr. arXiv 1411.5595, Nov. 2014.

[24] J. Suzuki and M. Nagata, “A unified learning framework of skip-grams and global vectors,” in ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, 2015, vol. 2, pp. 186–191.

[25] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2014, pp. 1532–1543.

[26] X. Rong, “word2vec Parameter Learning Explained,” arXiv Prepr. arXiv 1411.2738, Nov. 2014.

[27] L. Van Der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, pp. 2579–2625, Nov. 2008.

[28] L. Van Der Maaten, “Accelerating t-SNE using tree-based algorithms,” J. Mach. Learn. Res., vol. 15, pp. 3221–3245, Jan. 2015.

Authorized licensed use limited to: Zhejiang Normal University. Downloaded on November 25,2022 at 08:59:37 UTC from IEEE Xplore. Restrictions apply.