UNIVERSITY OF NAIROBI

School of Computing and Informatics

Use of Bayesian Model for Word alignment in Swahili-English Statistical Machine Translation

By

VINCENT OMWOMA WEKU

Supervisor

Dr. Lawrence Muchemi

A research project submitted in partial fulfillment of the requirement for the award of Masters of Science in Computer Science of the University of Nairobi,Kenya.

AUGUST, 2014

Declaration

This project report is my original work and has not been submitted in support of a degree or any other award in any University

Sign: .... Date: ....

Omwoma Vincent Weku

P58/76972/2012

This project report has been submitted for examination with my approval as University Supervisor.

Sign: .... Date: ....

Dr. Lawrence Muchemi

School of Computing and Informatics

University of Nairobi

Abstract

State of the art word alignment models such as IBM Models (Pietra *et al*.,1993), hidden Markov model(HMM)(Vogel *et al*.,1996), and the jointly-trained symmetric HMM, contain a large number of parameters such; word translation, transition and fertility probabilities, that need to be estimated in addition to desired alignment variables. The common method of inference in such models is expectation –maximization (EM) (Dumpster *et al*., 1977) or an approximation to EM when the exact EM is intractable. The EM algorithm finds the value of parameters that maximizes the likelihood of the observed variables. However, with many parameters to be estimated without prior, EM tends to explain the training data by over fitting the parameters. A well documented example of over fitting in EM-estimated word alignments is the case of rare words, where some of these words act as 'garbage collectors' aligning to excessively many words on the other side of the sentence pair (Pietra *et al*.,1993). Moreover EM is generally prone to getting stuck in a local maximization of the likelihood. Finally EM is based on the assumption that there is one fixed value of parameters that explains the data, that is, EM gives a point estimate. The over fitting problem mentioned among others has been alleviated by the use of word alignment model that uses Bayesian theorem that uses Gibbs sampling for inference as published by (Mermer *et al*., 2013). This approach has been successively applied to English, Arabic, Chinese and a host of other languages. It has however not been investigated for a Bantu language.

This research aimed at exploring the efficacy of a Bayesian based word alignment model for Kiswahili-English statistical machine translation problem. To achieve this, a Kiswahili-English corpus extracted from the Kiswahil-English corpus based on Tanzania constitution (Wagacha, 2014, (unpublished source)) with approximately 23 thousand pairs of sentences was used to train a Bayesian alignment model. The research shows that Bayesian model outperforms EM in the majority of test cases in Kiswahili-English corpus used. Further analysis reveals that the proposed method addresses the rare word problem. It also achieves higher vocabulary coverage rates. For example when using Bayesian, English has 3111 and Kiswahili has 2886 vocabularies compared to EM with 2544 and 2695 vocabularies. This research shows that Bayesian based alignment model can be used to improve alignment in Kiswahili-English statistical Machine Translation.

Acknowledgement

I acknowledge God Almighty for this great gift of life so as to accomplish this far I have come. To my supervisor Dr. Muchemi, who opened my eyes in the research world. His guidance, support, and positive criticism have made this project a success.

My thanks go to Professor P. Waiganjo, who made this project my availing a corpus to me unconditionally.

I can’t forget to thank my wife Irene Ngira, and my daughter Natalia Namatsi for their great support and encouragement throughout my Msc. Course.

To my friends and classmates, who shared ideas and provided assistance during this project. More so to Mr. Samuel Njugu for his seasoned skills in working with Moses Tool kit.

Table of Contents

List of Tables

List of Figures

List of Acronyms and Definitions

AER – Alignment Error Rate: A commonly used metric for assessing sentence alignments. It combines precision and recall metrics together such that a perfect alignment must have all of the sure alignments and may have some possible alignments

BAM – Bayesian Alignment Model: Bayesian based word aligner.

BLEU – Bilingual Evaluation Understudy: A commonly used evaluation metric for determining the performance of alignment models.

E, F – English, Foreign: English contains all English words; Foreign contains words for foreign language, e.g. Kiswahili, Arabic, Chinese and other languages.

EM – Expectation Maximization: A classical algorithm for inference in word alignment form IBM models

GS – Gibbs sampling: An algorithm that is used to infer and samples the posterior alignment distribution.

HMM – Hidden Markov model: State of the art alignment model that uses EM algorithm for inference.

IBM – International Business Machine: A computer manufacturing company, synonymous with the IBM alignment models.

SMT – Statistical Machine Translation: A machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.

VB – Variation Bayes: A model that closely resembles the usual form of EM algorithm which assumes that the parameters and hidden variable are independent and is thus and approximation of Bayesian model.

CHAPTER ONE: INTRODUCTION

1.1 Background of Study

Bantu Language of Swahili is spoken by more than fifty million people in East Africa and central Africa, however it is surprisingly resource scarce from a language technological point of view (De Pauw *et al.,* 2009).An increasing number of publication however are showing that carefully selected procedures can indeed boost language technology for Swahili.

Word alignment can be considered the backbone of Statistical Machine Translation. When Statistical Machine Translation shifted from word based to phrase based paradigm, still remained the base for most phrase-based, hierarchical and syntactic SMT systems (Guzman *et al.*, 2003). Word alignment is a crucial early step in the training pipeline of most statistical machine translation systems (Koehn, 2010).Whether the employed models are phrased-based or tree-based, they use the estimated word alignments for constraining the set of candidates in phrase or grammar rule extraction (Chiang, 2007). As such, the coverage and accuracy of the learned phrase/rule translation models are strongly correlated with those of word alignment. Given a sentence-aligned parallel corpus, the goal of word alignment is to identify the mapping between the source and the target words in parallel sentences. Since word alignment information is usually not available during corpus generation and human annotation is costly, the task of word alignment is considered as an unsupervised learning problem (Mermer *et al*., 2013).

State of the art word alignment models such as IBM Models (Pietra *et al*.,1993), hidden Markov model (HMM) (Vogel *et al*.,1996), and the jointly-trained symmetric HMM, contain a large number of parameters such; word translation, transition and fertility probabilities, that need to be estimated in addition to desired alignment variables.

The common method of inference in such models is expectation maximization (EM) (Dumpster *et al.,* 1977) or an approximation to EM when the exact EM is intractable. The EM algorithm finds the value of parameters that maximizes the likelihood of the observed variables. However, with many parameters to be estimated without prior, EM tends to explain the training data by over fitting the parameters. A well documented example of over fitting in EM-estimated word alignments is the case of rare words, where some of these words act as 'garbage collectors' aligning to excessively many words on the other side of the sentence pair (Pietra *et al.*,1993). Moreover EM is generally prone to getting stuck in a local maximization of the likelihood. Finally EM is based on the assumption that there is one fixed value of parameters that explains the data, that is, EM gives a point estimate.

The research proposed a Bayesian model in which prior distribution on the parameters will be utilized. The alignment is inferred by integrating over all possible parameter values. Word translation probabilities will be treated as multi-nominal- distributed random variables with a sparse Dirichlet prior. Inference is performed via Gibbs sampling which samples the posterior alignment distribution. The research compared EM and Bayesian alignments on the case of IBM models 1 and 2.

The Bayesian model was able to overcome the over fitting problem inherent in maximum likelihood training. The inferred alignment was evaluated in terms of end-to-end translation performance on various language pairs and corpora.

In this project, the research is organized as follows; chapter 2, Literature review i.e. some related work; chapter 3, the methodology, where we dwell deeper in the proposed model; chapter 4, Evaluation of results and finally chapter 5, Conclusion and Recommendation.

1.2 Statement of the Problem

State-of-the-art word alignment models such as IBM, hidden markov and jointly trained-trained symmetric hidden markov models contain a large number of parameters e.g. word translations, transition and fertility probabilities that need to be estimated in addition to the desired alignment variables. This has been used in many languages around the globe. The common method of inference is expectation maximization (EM). However with many parameters to be estimated i.e. word translation, transition and fertility probabilities, which need to be, estimated in addition to desired alignment variables without prior, EM tends to explain the training data by over fitting the parameters, a good example is that of the rare words. Moreover EM is generally prone to getting stuck in local maximum of likelihood. EM is also based on the assumption that there is one fixed value of parameters that explain the data.

This research proposes a Bayesian model that utilizes the prior distribution on the parameters. The Bayesian model overcomes the over fitting problem inherent in maximum likelihood training by placing prior probabilities on the model parameters for the Swahili language, hence address the rare word problem.

1.3 Research Objectives

This outlines the objectives of this project. This was the guiding principle throughout the research period. The research identified three main objectives of the project. The first objective entailed designing of a basic Bayesian word alignment model for English-Swahili statistical Machine Translation of phrases based translations. The second objective entailed development of a prototype based on the above model devoid of over fitting problems, sticking to local variable and assumptions of one fixed local variable. The third objective entailed the evaluation of the model in terms of its performance using an extrinsic evaluation measure with BLEU and intrinsic evaluation measure using Alignment Error Rate.

In addition to the above key objectives, the research identifies others like; comparing of Bayesian model to EM model in terms of performance, test that the model solves the problem of over fitting and sticking to local maxima problem.

These can be summarized as follows;

1. To design a basic Bayesian word alignment model for Kiswahili-English SMT
2. To develop a prototype based on the above model.
3. To evaluate the model in terms of its performance.

1.4 Purpose of the Study

The purpose of the study is to study the use of Bayesian Model for Word alignment in Swahili-English Statistical Machine Translation. The research focuses on Bayesian word alignment model Swahili-English word alignment .Bayesian model will address the major problems of EM which are over fitting problem and high fertility rare word among others mentioned previously in this document.

1.5 Justification and Significance of the Study

From this project, the significance becomes evident in that a better word alignment model will be proposed. This model will be crucial to users as well as developers of Statistical Machine translation (SMT) that performs word alignment for English-Swahili translation. This model is effective compared to the classical EM, since it addresses various problems that are always encountered in EM e.g. over fitting problems, high fertility rare words and intractable optimization problem.

The model will determine the rate at which any word or words will be aligned from the source (English) to target (Swahili) hence efficiency.

1.6 Scope and Limitation of the Study

The study focused mainly on word alignment of sentences or words from English to Swahili (a popular native language commonly used in most parts of East Africa) using Bayesian model. Expectation Maximization (EM) is the only traditional model that the research was being compared to the proposed model in this project.

Due to limited time the research may not analyze Bayesian model’s performance compared to EM’s, in details, but instead restrict itself to designing of a model, development of its prototype and its evaluation.

CHAPTER TWO: LITERATURE REVIEW

2.1 Overview

The literature review was based on identifying and accessing various preliminary or primary sources of relevant information. The research used online books and journals as well as sites and articles to get the immense knowledge word alignment models.

Through the online references as well publications, led to identification of other viable secondary sources of information that has aided much in this research for example the English-Kiswahili corpus.

2.2 Swahili

Swahili is widely spoken in East Africa. It has approximately 80 million speakers spread across several countries such as Tanzania and Kenya, where it has an official status, Uganda where it is a national language, and in regions that border these countries in Malawi, Mozambique, the Democratic Republic of Congo, Rwanda, Ethiopia and Somalia (Ng'ang'a,2005).

It is a highly inflecting language where both prefixes and suffixed morphemes play an important grammatical role. The functions of prefixes are particularly important in both nominal and verbal morphology.

2.3 Statistical machine Translation

2.3.1 Introduction

Statistical machine translation (**SMT**) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The statistical approach contrasts with the rule-based approaches to machine translation as well as with example-based machine translation.

The idea behind statistical machine translation comes from information theory. A document is translated according to the probability distribution $p(e|f)$ that a string $e$in the target language (for example, English) is the translation of a string $f$in the source language (for example, French).

For a rigorous implementation of this one would have to perform an exhaustive search by going through all strings e^*in the native language. Performing the search efficiently is the work of a machine translation decoder that uses the foreign string, heuristics and other methods to limit the search space and at the same time keeping acceptable quality. This trade-off between quality and time usage can also be found in speech recognition. (Och, 2005)

Common examples include Asian Online, Babel Fish, Translator, Google translate, Jollo and SYTRAN among others

2.3.2 Benefits

The most frequently cited benefits of statistical machine translation over rule-based approach are:

- **Better use of resources**
  - There is a great deal of natural language in machine-readable format.
  - Generally, SMT systems are not tailored to any specific pair of languages.
  - Rule-based translation systems require the manual development of linguistic rules, which can be costly, and which often do not generalize to other languages.
- **More natural translations**
  - Rule-based translation systems are likely to result in Literal translation. While it appears that SMT should avoid this problem and result in natural translations, this is counterbalanced by the fact that using statistical matching to translate rather than a dictionary/grammar rules approach can often result in text that include apparently nonsensical and obvious errors.

2.3.3 Types of SMT

i) Word-Based Translation: - An example of a word-based translation system is the freely available GIZA++ package (GPLed), which includes the training program for IBM models and HMM model and Model 6. **(**Shane,2005)

ii) Phrase-based translation

iii) Syntactic-based translations

iv) Hierarchical phrase-based translation

2.3.4 Challenges of SMTs

i) Sentence Alignment

ii) Statistical Anomalies

iii) Data dilution

iv) Idioms

v) Different word order

vi) Out of vocabulary (OOV) words

2.3.5 English-Swahili Machine Translation

The increasing amount of digitally available, vernacular data has prompted researchers to investigate the applicability of corpus-based approaches to African language technology.

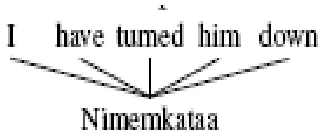
(De Pauw *et al*., 2009), in their research they develop a SAWA corpus project that attempts to collect and deploy a parallel corpus English-Swahili not only for the straightforward purpose of developing machine translation system but also to investigate the possibility of projection of annotation into a resource-scarce African language.

According to (Schimid, 1994), each text in the SAWA corpus is automatically part-of-speech tagged and lemmatized using the TreeTagger for the English part. These extra annotation layers allow (De Pauw *et al*., 2009) to perform more accurate automatic word alignment on the basis of factored data.

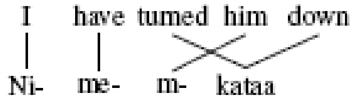
2.4 Automatic Word Alignment

They performed word alignment experiments using GIZA++ on the factored data of the SAWA corpus. However from the results they got, this method is underwhelming on Words.

The main problem in training a GIZA++ model for language pair English-Swahili is the strong agglutinating nature of the latter. There is no parallel corpus exhaustive enough to provide enough linguistic evidence to unearth strongly converging alignment patterns e.g



Morphologically deconstructing the word however can greatly relieve the sparse data problem for this;



The isolated Swahili morphemes can more easily be linked to their English counter parts since there will be more linguistic evidence in the parallel corpus, linking for example *ni* to *I* and *m* to *him.*

(De Pauw *et al.*, 2009a) acknowledges of having no morphologically aligned gold standard data available, so evaluation of the morpheme-based approach was to be done in a roundabout way. First they morphologically decomposed the Swahili data and run GIZA++ again. Next, they recompiled the Swahili words from the morphemes and grouped the word alignment links accordingly. Incompatible linkages were removed and simple majority voting resolves ambiguous alignment patterns.

The SAWA corpus was randomly divided into a 90% training set and a 10% test set. The SMT system was built on the training set and evaluated on the test set using the standard machine translation evaluation measures BLEU and NIST. The results were compared to those of Google translate system for Swahili.

Table 1: BLEU and NIST scores for Bidirectional machine translation Task

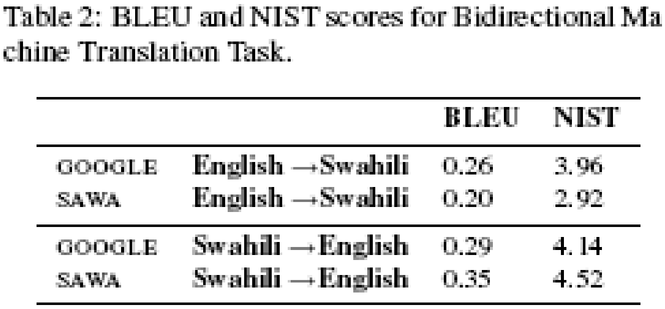


Table 2: BLEU and NIST scores for Bidirectional Machine Translation Task.

		BLEU	NIST
GOOGLE	English →Swahili	0.26	3.96
SAWA	English →Swahili	0.20	2.92
GOOGLE	Swahili →English	0.29	4.14
SAWA	Swahili →English	0.35	4.52

Source: (De Pauw et al., 2009).

Error analysis showed that the SAWA system had significant difficulties generating morphologically correct Swahili words. Swahili-English translation, De Pauw’s et al. (2009) system fared better, not hampered by the morphological generation issues of the target language. In this case, the SAWA system was able to outperform the Google system by a significant margin.

2.5 Standard EM for Word Alignment

Problems with the standard EM estimation of IBM Model 1 were pointed out by (Moore, 2004). A number of heuristic changes to the estimation procedure, such as smoothing the parameter estimates, were shown to reduce the alignment error rate, but the effects on translation performance were not reported. (Zhao and Xing, 2006) address the data sparsity issue using symmetric Dirichlet priors in parameter estimation and they use variational EM to find the maximum *a posteriori* (MAP) solution. (Vaswani *et al,2012.),* encourage sparsity in the translation model by placing a prior on the parameters and then optimize for the MAP objective.

(Zhao and Gildea, 2010) use sampling in their proposed fertility extensions to IBM Model 1 and HMM, but they do not place any prior on the parameters. Their inference method is stochastic EM (also known as Monte Carlo EM), a maximum-likelihood technique in which sampling is used to approximate the expected counts in the E-step. Even though they report substantial reductions in the alignment error rate, the translation performance measured in BLEU does not improve.

Bayesian modeling and inference have recently been applied to several unsupervised learning problems in natural language processing such as part-of-speech tagging (Goldwater et al., 2007), (Gao *et al*., 2007), word segmentation, grammar extraction (Mochihashi et al., 2009) and finite-state transducer training (Chiang et al.,2010) as well as other tasks in SMT such as synchronous grammar induction (Blunsom et al.,2009) and learning phrase alignments directly (DeNero et al.,2008).

Word alignment learning problem was addressed jointly with segmentation learning by (Xu *et al.2008)*, (Nguyen *et al.2010*), and (Chung and Gildea., 2009). As in this paper, they treat word translation probabilities as random variables (with an associated prior distribution).

In (Xu *et al., 2008),* a Dirichlet Process prior is placed on IBM Model 1 word translation probabilities. In (Nguyen *et al., 2010*), a Pitman-Yor Process prior is placed on word translation probabilities in a proposed bag-of-words translation model that is similar to IBM Model 1. Both studies utilize Gibbs sampling for inference. However, alignment distributions are not sampled from the true posteriors but instead are updated either by running GIZA++ (Xu *et al., 2008)* or using a “local-best” maximization search (Nguyen *et al., 2010*).

On the other hand, a sparse Dirichlet prior on the multinomial parameters is used in (Chung and Gildea, 2009) to prevent over fitting.

Bayesian word alignment with Dirichlet priors was also investigated in a recent study using variational Bayes (VB) (Riley and Gildea, 2012).VB is a Bayesian inference method which is sometimes preferred over Gibbs sampling due to its relatively lower computational cost and scalability. However, VB inference approximates the model by assuming independence between the hidden variables and the parameters.

2.6 Bayesian Models and Variational Bayes

2.6.1. Variational Bayes

(Riley & Gildea) in their research to improve the performance of GIZA++, they states that Variational Bayes closely resembles the usual form of EM algorithm which assumes that the parameters and hidden variable are independent and is thus an approximation of Bayesian Model.

Beal (Beal, 2003) gives a detailed derivation of a variational Bayesian algorithm for HMMs. The result is a very slight change to the M step of the original EM algorithm. Beal asserts that during the M step of the original EM algorithm, the expected counts collected in E step are normalized to give the new values of the parameters;

$$\theta_{x_i|y} = \frac{E[c(x_i|y)]}{\sum_j E[c(x_j|y)]}$$

2.6.2. Bayesian models and EM

Thus, there are many benefits to be gained from introducing these Dirichlet priors into the model. Unfortunately, adding a prior over the parameters means that we cannot use EM directly. This is because the EM algorithm makes implicit independence assumptions. First it holds the parameters fixed and optimizes the hidden variables with respect to them; then it holds the hidden variables fixed and optimizes the parameters. EM algorithms typically rely on dynamic programming to compute expected counts for hidden variables.

These dynamic programming algorithms are possible because, for a non-Bayesian model, the parameters are treated as fixed quantities rather than random variables. Thus the probabilities of the hidden variables decompose according to such a model, and there is no independence assumptions required for the E step. For the M step, the assumption is that the parameters can be optimized independently of their effects on the hidden variables, and it is this assumption that makes the EM algorithm an approximate algorithm which converges on the solution rather than finding it directly.

For the Bayesian model, on the other hand, the parameters are random variables, and thus they are interdependent with the hidden variables; both are conditioned on the fixed hyper parameter. This makes it impossible to compute exact expected counts for all the hidden variables with respect to the parameters during the E step, and this is why EM cannot be used with Bayesian models.

Thus, according to (Riley & Gildea, 2012) if Bayesian model was to be used, then one must resort to approximate algorithms. Johnson (2007) tested two such approximate Bayesian algorithms, Gibbs sampling and variational Bayes, and found variational Bayes to be superior.

2.6.3. Bayesian Model

According to (Riley & Gildea, 2012), they took a Bayesian stance and suggested that the parameters themselves have been drawn from the probability distribution. They say that the parameters of this kind were to be referred to as hyper parameters which are denoted by α. Therefore by choosing their values appropriately, the model can be biased towards learning sort of parameters they (Riley & Gildea, 2012) they would like it to learn.

The goal of Bayesian methods is to compute the likelihood of a new data point x^{n+1}. This involves integrating over the model parameters and the hidden states both of which are now encompassed by θ. They are combined into one variable because in a Bayesian model, the parameters act as hidden variables: they, like the hidden variables are random variables that have been drawn from a probability distribution.

2.6.4 Conceptual Model

Ii is a **model** made of the composition of concepts that thus exists only in the mind. Is used to help us know, understand, or simulate the subject matter they represent.

In the below diagram, the model to be, has been conceptualized. The description is as follows;

Step 1: The corpus (parallel text) is fed into the model. The corpus is important since it has Swahili and English correspondences. This facilitates the learning of the model.

Step 2: The corpus is then processed in a learning module of the Bayesian alignment model. In this step, there is a dirichlet prior process that performs the probabilities of both source and target vocabularies.

Step 3: The vocabulary files are then passed to what we can term as the 'database' of the model in order for transition and emission will take place. In this step, issues concerning the sentence rules will be addressed.

Step 4: The vocabulary files are then passed in a construction module that ensure the words/phrases aligned to the target are making sense. This process is iterative, i.e. the files will keep on passing back and forth to component module. This is where the Gibbs sampler algorithm sits. Its work is to keep on checking that the words/sentences have been aligned correctly based on the probability distributions of parameters.

Step 5: The phases/words are then passed through the alignment module. This step, the phrases aligned will be confirmed.

Step 6: Finally the target output (a translated word or phrase). This is the final translated pair of sentences in Swahili and English.

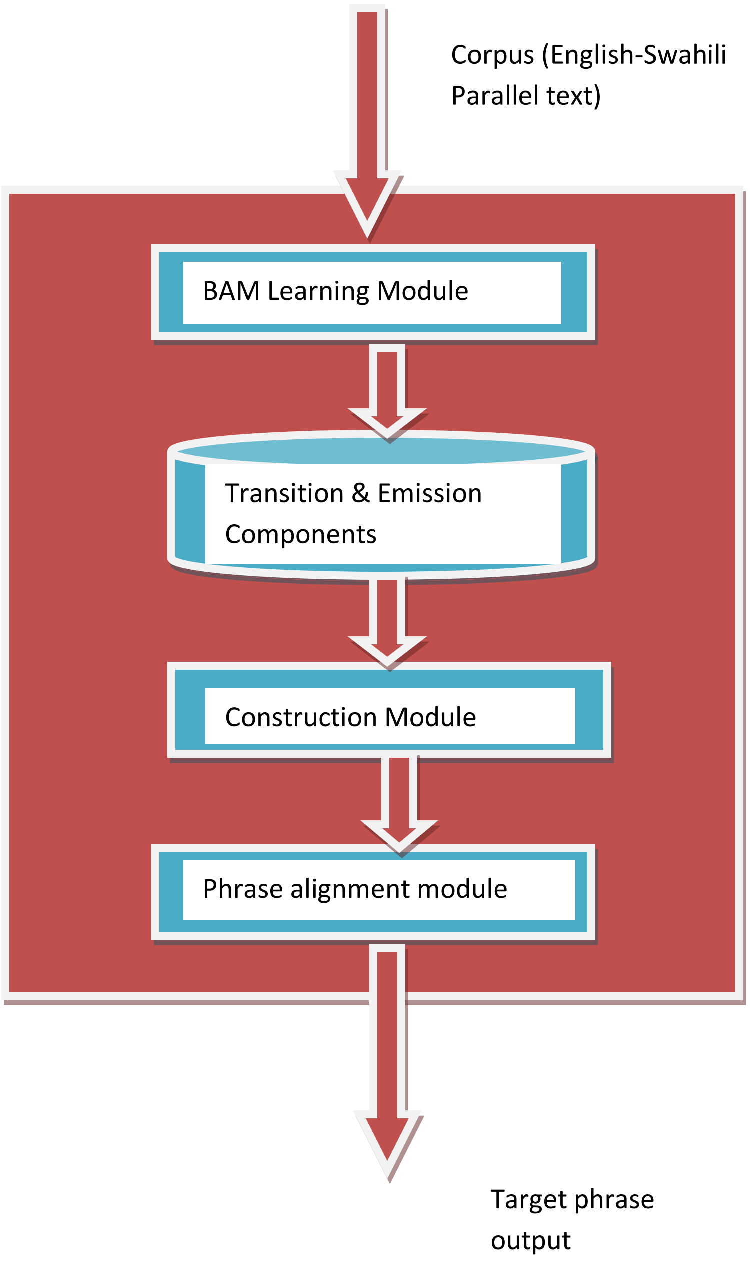


Fig 2.1- Conceptual Model.

CHAPTER THREE: METHODOLOGY

3.1 Introduction

Statistical Machine Translation is crucial when it comes to natural language processing. EM (Expectation Maximization) has been the method of choice for most word alignment, which has its own bottlenecks.

This research was to investigate the use of Bayesian model for Word alignment in Swahili-English Machine Translation. Therefore, in this section; the research established various procedures employed to conduct scientific research for the success of this project. It looked at methodologies e.g. design of the model, data collection, model platform and coding and evaluation. The methodologies were used in order to achieve the objectives being studied.

3.2 Research Design

This section methods and procedures employed to conduct scientific research will be discussed. The design of a study defines the study type and sub-type, research question, hypotheses, independent and dependent variables, experimental design, and, if applicable, data collection methods and analysis.

Design of Bayesian Alignment Model

The research relied heavily on the literature review in order to come up with a model. Literature from Moses tool kit was instrumental since it had a Giza++ model that incorporates EM algorithm which built the foundation of my proposed model.

For better results, research work directly related to Bayesian models or Bayes, was crucial to the design. Probabilistic algorithms also played a very important role. Gibbs algorithm is a probabilistic algorithm, and therefore studying other related algorithm proved to be instrumental to the designing of the proposed model.

Development of a prototype

Data Collection

Data collection was crucial for actualization of the model prototype, based on the information gathered earlier in the literature.

The means of data collection was secondary, whereby a corpus containing Kiswahili-English sentences was used.

Acquiring the corpus was one thing and preparing it another thing. For instance the following tasks would have to be performed to the corpus to make it viable for use;

- **Tokenisation**: This means that spaces have to be inserted between (e.g.) words and punctuation.
- **Truecasing**: The initial words in each sentence are converted to their most probable casing. This helps reduce data sparsity
- **Cleaning**: Long sentences and empty sentences are removed as they can cause problems with the training pipeline, and obviously miss-aligned sentences are removed.

- **Generating Files**: From the corpus, both the English Kiswahili sentences are copied and pasted into two different files after the above three processes have been done. The original corpus is a simple English phrases/words with their Kiswahili (foreign) correspondences.

Model platform and coding

The following tools and languages were needed

- Perl Active state, (for coding the model, which is perl language)
- Linux Preferably Ubuntu 11(operating system) and above,
- Moses Tool kit with Giza++
- Computer with minimum requirements of 2GB RAM and core i3 processor

Evaluation of the model

Pairs of aligned sentence were the output after executing the model. The same corpus was also run through GIZA++. Since the two models used different algorithms which were the centre of the research, outputs (aligned sentences/phrases) from both algorithms were crucial for evaluation of the model.

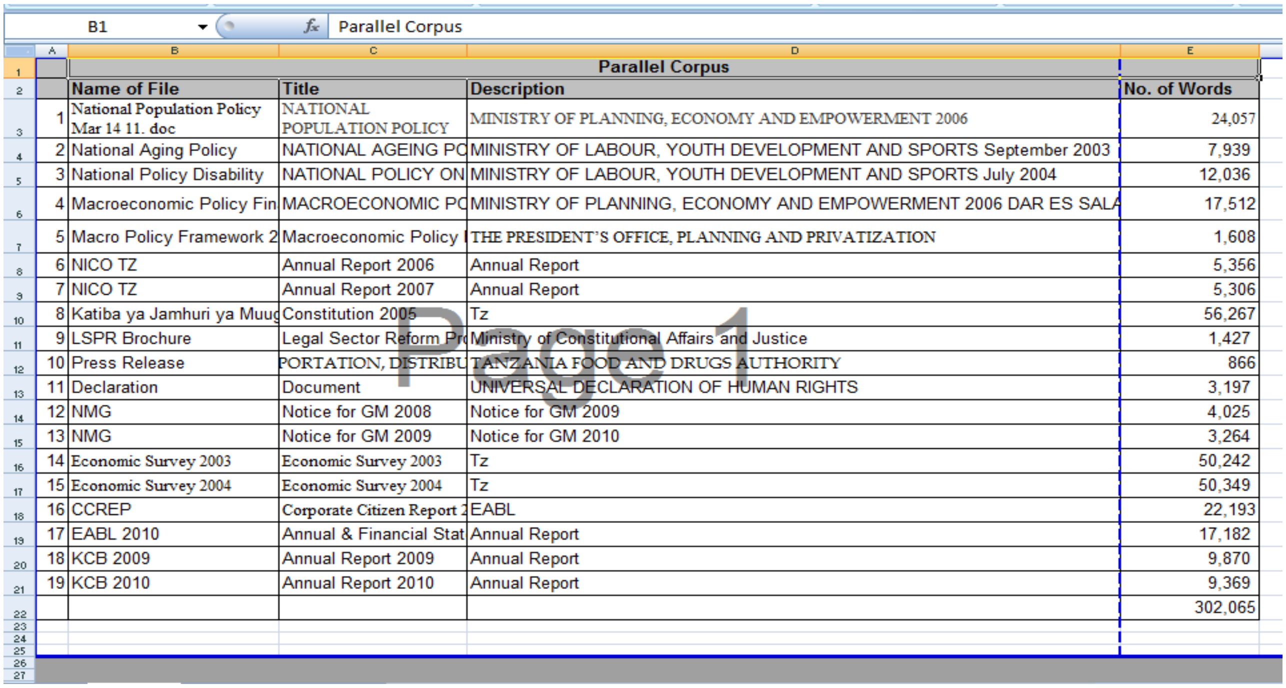
Bilingual Evaluation understudy (BLEU) an evaluation tool, which comes with Moses toolkit was used. The tool sampled all the phrases/words and determined the quality of the alignments in both models.

3.3 Data Collection

3.3.1 Data Source

The data used in this project was secondary data. Unlike traditional data collection methods, a corpus was acquired based on different context. Below is a summary of the corpus. Additional information was available from Moses decoder website on how to carry out the translation.

Table 3.1: Summary of Corpus



Parallel Corpus

	Name of File	Title	Description	No. of Words
1	National Population Policy Mar 14 11. doc	NATIONAL POPULATION POLICY	MINISTRY OF PLANNING, ECONOMY AND EMPOWERMENT 2006	24,057
2	National Aging Policy	NATIONAL AGEING PC	MINISTRY OF LABOUR, YOUTH DEVELOPMENT AND SPORTS September 2003	7,939
3	National Policy Disability	NATIONAL POLICY ON	MINISTRY OF LABOUR, YOUTH DEVELOPMENT AND SPORTS July 2004	12,036
4	Macroeconomic Policy Fin	MACROECONOMIC PC	MINISTRY OF PLANNING, ECONOMY AND EMPOWERMENT 2006 DAR ES SALA	17,512
5	Macro Policy Framework 2	Macroeconomic Policy I	THE PRESIDENT'S OFFICE, PLANNING AND PRIVATIZATION	1,608
6	NICO TZ	Annual Report 2006	Annual Report	5,356
7	NICO TZ	Annual Report 2007	Annual Report	5,306
8	Katiba ya Jamhuri ya Muuç	Constitution 2005	Tz	56,267
9	LSPR Brochure	Legal Sector Reform Pr	Ministry of Constitutional Affairs and Justice	1,427
10	Press Release	PORTATION, DISTRIBU	TANZANIA FOOD AND DRUGS AUTHORITY	866
11	Declaration	Document	UNIVERSAL DECLARATION OF HUMAN RIGHTS	3,197
12	NMG	Notice for GM 2008	Notice for GM 2009	4,025
13	NMG	Notice for GM 2009	Notice for GM 2010	3,264
14	Economic Survey 2003	Economic Survey 2003	Tz	50,242
15	Economic Survey 2004	Economic Survey 2004	Tz	50,349
16	CCREP	Corporate Citizen Report 2	EABL	22,193
17	EABL 2010	Annual & Financial Stat	Annual Report	17,182
18	KCB 2009	Annual Report 2009	Annual Report	9,870
19	KCB 2010	Annual Report 2010	Annual Report	9,369
				302,065

Source: ***Courtesy of Prof. P.Waiganjo (Unpublished source)***

3.3.2 Analyzing Corpus.

From the above summary, it is clear about the various domains of corpora that are being used. The whole point of using various domains is to enhance the performance of the alignment model and better training.

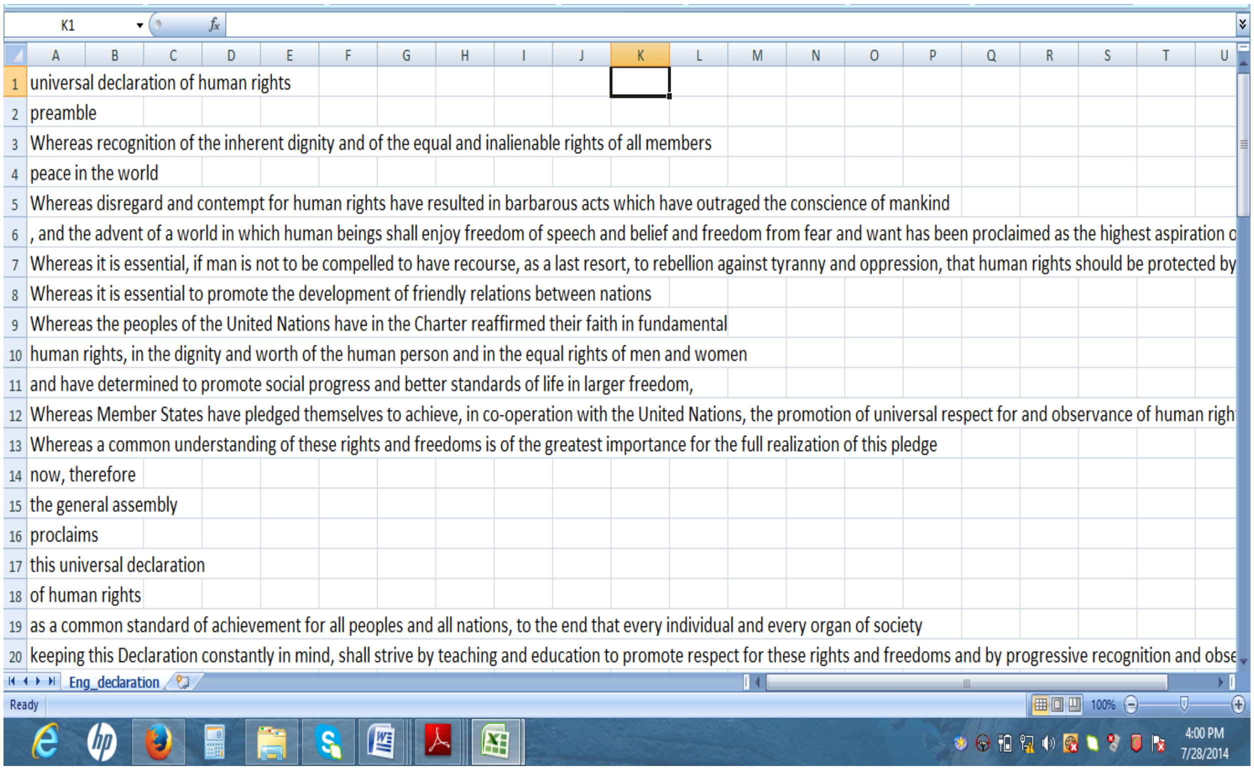
Table 3.2 Sample Corpus

azimio la ulimwengu juu ya haki za binadamu	universal declaration of human rights
Utangulizi	preamble
Kwa kuwa kukiri heshima ya asili na haki sawa kwa binadamu wote ndio msingi wa uhuru,haki na ya watu wote	Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members
amani duniani.	peace in the world
Kwa kuwa kutojali na kudharau haki za binadamu kumesababisha vitendo vya kinyama ambavyo vimeharibu dhamiri ya binadamu	Whereas disregard and contempt for human rights have resulted in barbarous acts which have outraged the conscience of mankind
na kwa sababu Azimio la ulimwengu ambalo litawafanya binadamu wafurahie uhuru wao wa wakusema, kusadiki na wa kutoogopa chochote limekwisha kutangazwa kwamba ndio hamu kuu ya watu wote.	, and the advent of a world in which human beings shall enjoy freedom of speech and belief and freedom from fear and want has been proclaimed as the highest aspiration of thecommon people,

From the above, two columns are seen, the Swahili and English correspondence. From the above summary, the project uses the constitution of Tanzania as the corpus. The project domain is the constitution of Tanzania.

For preparation, each column was extracted and saved in its file with extension dot txt or dot csv. This is vital since the model will interpret the formats comfortably.

Below is an extracted English word from the Declaration corpus. The file is placed on the excel work sheet but saved as a dot csv file for the model to recognize it.

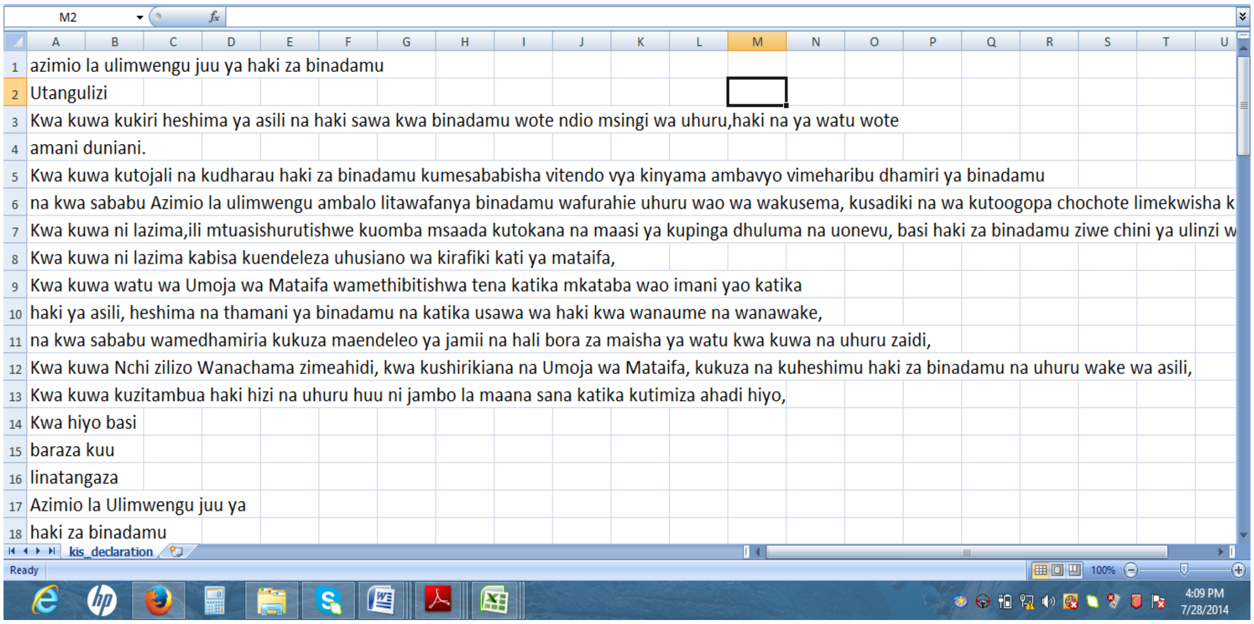


	A
1	universal declaration of human rights
2	preamble
3	Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members
4	peace in the world
5	Whereas disregard and contempt for human rights have resulted in barbarous acts which have outraged the conscience of mankind
6	, and the advent of a world in which human beings shall enjoy freedom of speech and belief and freedom from fear and want has been proclaimed as the highest aspiration o
7	Whereas it is essential, if man is not to be compelled to have recourse, as a last resort, to rebellion against tyranny and oppression, that human rights should be protected by
8	Whereas it is essential to promote the development of friendly relations between nations
9	Whereas the peoples of the United Nations have in the Charter reaffirmed their faith in fundamental
10	human rights, in the dignity and worth of the human person and in the equal rights of men and women
11	and have determined to promote social progress and better standards of life in larger freedom,
12	Whereas Member States have pledged themselves to achieve, in co-operation with the United Nations, the promotion of universal respect for and observance of human righ
13	Whereas a common understanding of these rights and freedoms is of the greatest importance for the full realization of this pledge
14	now, therefore
15	the general assembly
16	proclaims
17	this universal declaration
18	of human rights
19	as a common standard of achievement for all peoples and all nations, to the end that every individual and every organ of society
20	keeping this Declaration constantly in mind, shall strive by teaching and education to promote respect for these rights and freedoms and by progressive recognition and obse

Fig 3.1: English file (Eng Declaration.csv).

Below is an extracted English word from the Declaration corpus. The file is placed on the excel sheet but saved as a dot csv file for the model to recognize

it.



	A
1	azimio la ulimwengu juu ya haki za binadamu
2	Utangulizi
3	Kwa kuwa kukiri heshima ya asili na haki sawa kwa binadamu wote ndio msingi wa uhuru,haki na ya watu wote
4	amani duniani.
5	Kwa kuwa kutojali na kudharau haki za binadamu kumesababisha vitendo vya kinyama ambavyo vimeharibu dhamiri ya binadamu
6	na kwa sababu Azimio la ulimwengu ambalo litawafanya binadamu wafurahie uhuru wao wa wakusema, kusadiki na wa kutoogopa chochote limekwisha k
7	Kwa kuwa ni lazima,ili mtuasishurutishwe kuomba msaada kutokana na maasi ya kupinga dhuluma na uonevu, basi haki za binadamu ziwe chini ya ulinzi w
8	Kwa kuwa ni lazima kabisa kuendeleza uhusiano wa kirafiki kati ya mataifa,
9	Kwa kuwa watu wa Umoja wa Mataifa wamethibitishwa tena katika mkataba wao imani yao katika
10	haki ya asili, heshima na thamani ya binadamu na katika usawa wa haki kwa wanaume na wanawake,
11	na kwa sababu wamedhamiria kukuza maendeleo ya jamii na hali bora za maisha ya watu kwa kuwa na uhuru zaidi,
12	Kwa kuwa Nchi zilizo Wanachama zimeahidi, kwa kushirikiana na Umoja wa Mataifa, kukuza na kuheshimu haki za binadamu na uhuru wake wa asili,
13	Kwa kuwa kuzitambua haki hizi na uhuru huu ni jambo la maana sana katika kutimiza ahadi hiyo,
14	Kwa hiyo basi
15	baraza kuu
16	linatangaza
17	Azimio la Ulimwengu juu ya
18	haki za binadamu

Fig3.2: Kiswahili (Kisw Declaration.cvs)

3.4 Model Design

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements.

3.4.1 Corpus Preparation

The original corpus has lots of impurities which has to be removed .The corpus needs to undergo various processing steps in order for it to be fed into the model.

To prepare the data for training the translation system, the following steps had to be performed:

- **Tokenization**: This means that spaces have to be inserted between (e.g.) words and punctuation.
- **True casing**: The initial words in each sentence are converted to their most probable casing. This helps reduce data sparsity

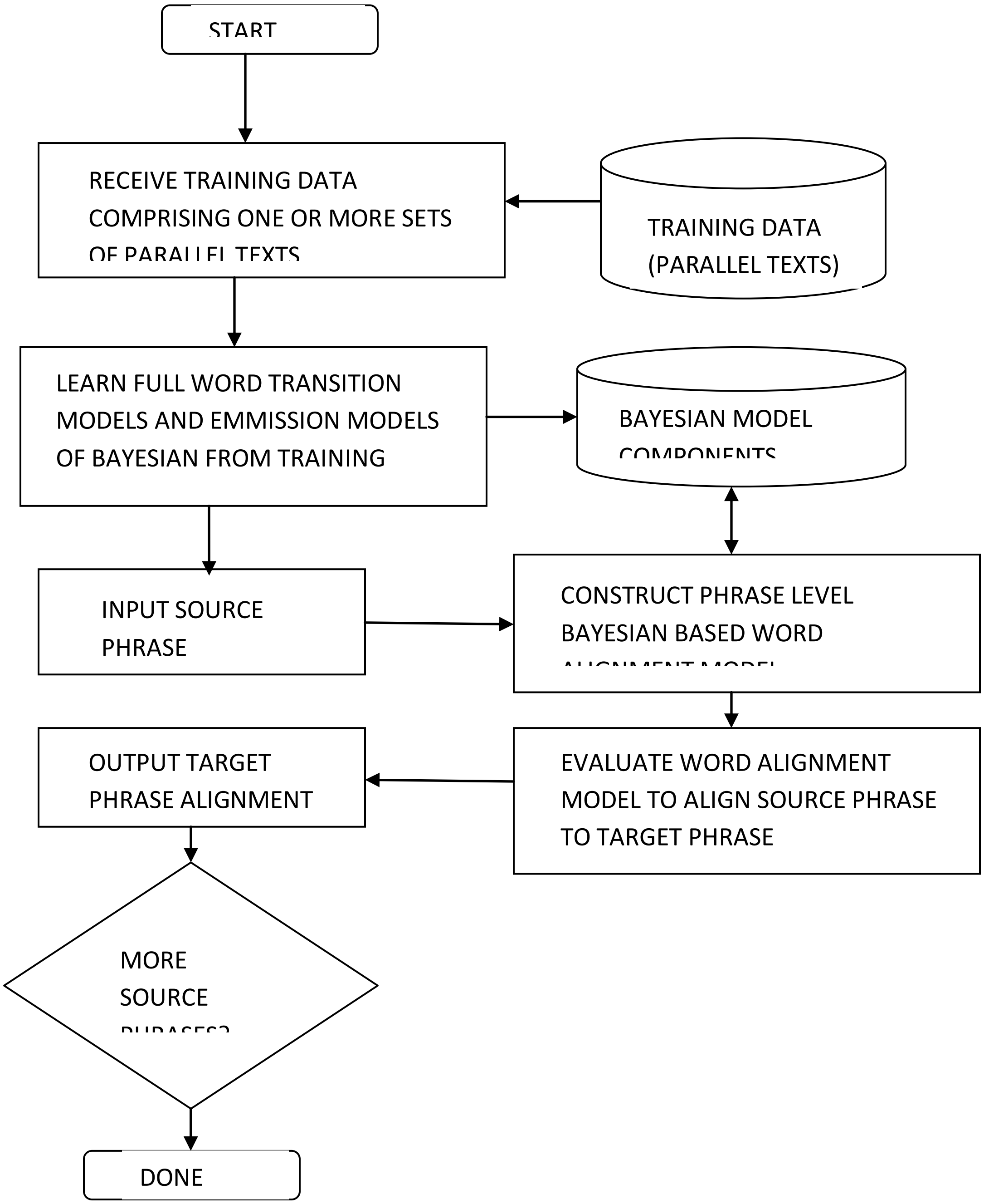
- **Cleaning**: Long sentences and empty sentences are removed as they can cause problems with the training pipeline, and obviously mis-aligned sentences are removed.

Generating Files

From the corpus, both the English Kiswahili sentences are copied and pasted into two different files after the above three processes have been done. The original corpus is a simple word.

Below is a diagram that describes the general process the corpus will go through until our target output is obtained which is the aligned text in Swahili. It is a simple flowchart that is meant to open our eyes on the next diagram, which is our model. From the diagram below, we can deduce where our model sits.

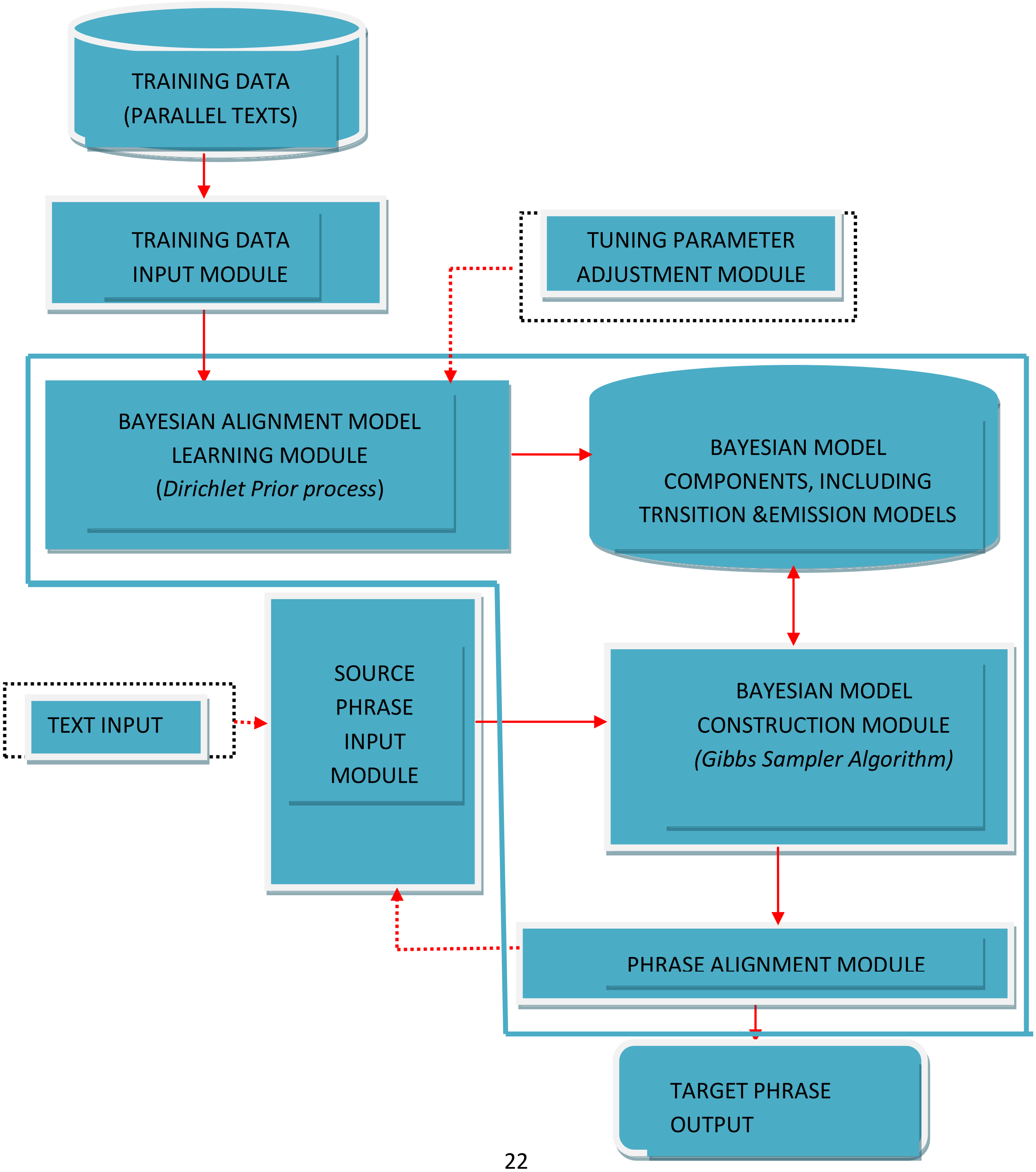
Fig 3.3: Translation Process


START
RECEIVE TRAINING DATA COMPRISING ONE OR MORE SETS OF PARALLEL TEXTS
TRAINING DATA (PARALLEL TEXTS)
LEARN FULL WORD TRANSITION MODELS AND EMMISSION MODELS OF BAYESIAN FROM TRAINING
BAYESIAN MODEL COMPONENTS
INPUT SOURCE PHRASE
CONSTRUCT PHRASE LEVEL BAYESIAN BASED WORD ALIGNMENT MODEL
OUTPUT TARGET PHRASE ALIGNMENT
EVALUATE WORD ALIGNMENT MODEL TO ALIGN SOURCE PHRASE TO TARGET PHRASE
MORE SOURCE PHRASES?
DONE

3.5 Bayesian Model for Word Alignment in Swahili-English SMT

The diagram below has been extracted from the previous one in Fig 3.3; the transition from Fig 3.3 to the one in figure 3.4 is meant to explain our model in detail. Remember the above diagram was general; the research has pinpointed the part that is Bayesian model.

Fig 3.4: Model Architecture



3.5.1 Illustrative Example

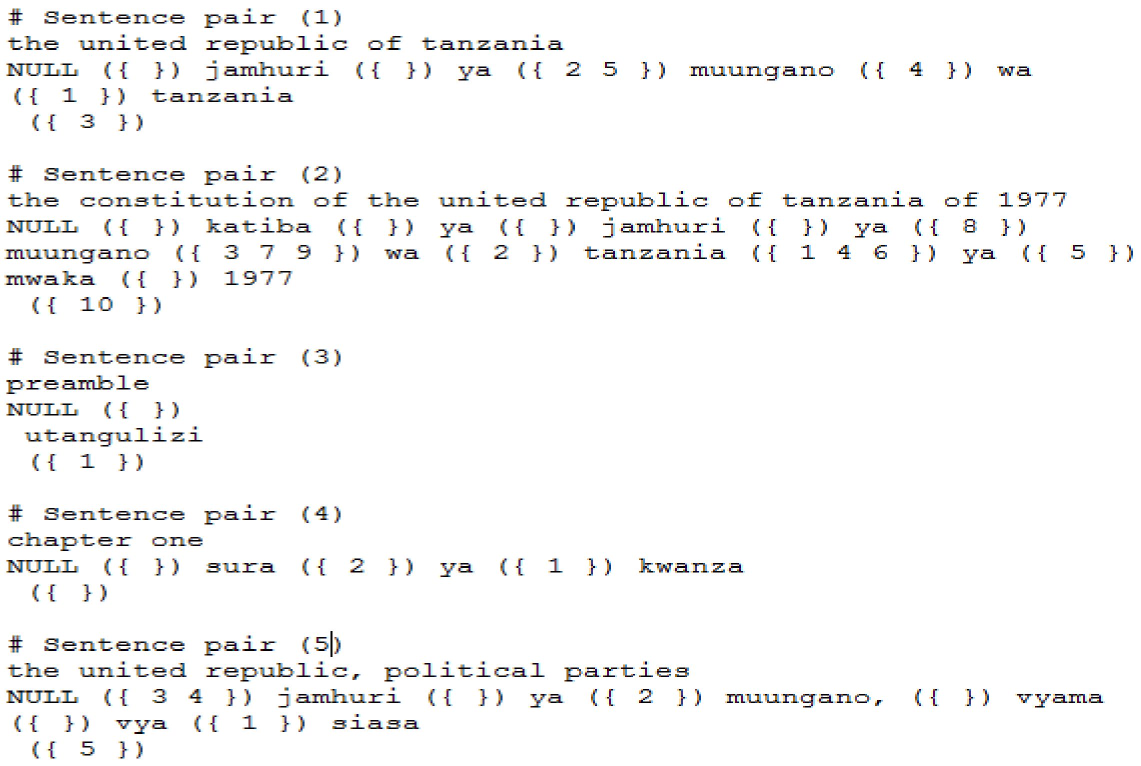
This section explains the working of the model using real examples. It outlines the following areas; Training data (parallel text), training data, which contains vocabulary files generated by the model; learning model that has the dirichlet prior process for probability distributions; Bayesian model components; construction module, which has the Gibbs sampling module for probability sampling distribution; phrase alignment which performs phrase alignment after all the probability distribution, and finally the target output.

Table 3.3: Corpus Example

the united republic of Tanzania	Jamhuri ya muungano wa tanzania
The constitution of the united republic of Tanzania of 1977	Katiba ya jamhuri ya muunguno wa Tanzania ya mwaka 1977
preamble	utangulizi
Chapter one	Sura ya kwanza
The united republic, political parties	Jamhuri ya muungano, vyama vya siasa

The table above is a sample corpus which was used in the illustration. It has both English and corresponding Kiswahili sentences as extracted from the corpus based on the constitution of Tanzania (courtesy of Prof. P. Waiganjo (Unpublished source)).

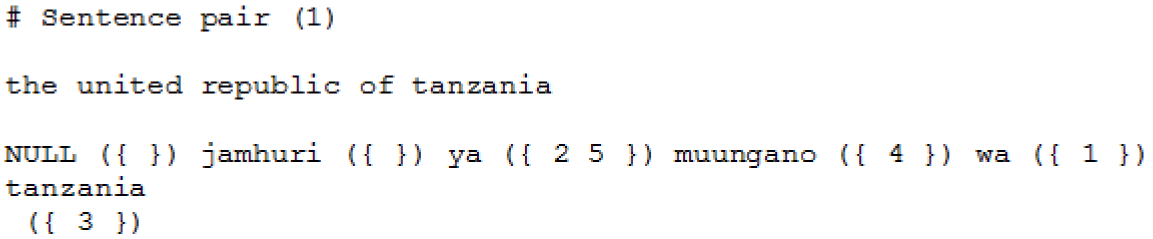
The corpus will be converted to vocabulary files, that is, English and Foreign (Kiswahili) vocabulary files. In summary, the sentences on the left are to be mapped to the sentences on the right in each row. For example, the sentence "*The united republic of Tanzania*" will be mapped to the Swahili sentence "*jamhuri ya muungano wa tanzania*".



```
# Sentence pair (1)
the united republic of tanzania
NULL ({ }) jamhuri ({ }) ya ({ 2 5 }) muungano ({ 4 }) wa
({ 1 }) tanzania
  ({ 3 })

# Sentence pair (2)
the constitution of the united republic of tanzania of 1977
NULL ({ }) katiba ({ }) ya ({ }) jamhuri ({ }) ya ({ 8 })
muungano ({ 3 7 9 }) wa ({ 2 }) tanzania ({ 1 4 6 }) ya ({ 5 })
mwaka ({ }) 1977
  ({ 10 })

# Sentence pair (3)
preamble
NULL ({ })
 utangulizi
  ({ 1 })

# Sentence pair (4)
chapter one
NULL ({ }) sura ({ 2 }) ya ({ 1 }) kwanza
  ({ })

# Sentence pair (5)
the united republic, political parties
NULL ({ 3 4 }) jamhuri ({ }) ya ({ 2 }) muungano, ({ }) vyama
({ }) vya ({ 1 }) siasa
  ({ 5 })
```

Fig 3.5: Output from the model

In the above figure outlines the output and what it entails. This is the final output from the Bayesian model.



```
# Sentence pair (1)

the united republic of tanzania

NULL ({ }) jamhuri ({ }) ya ({ 2 5 }) muungano ({ 4 }) wa ({ 1 })
tanzania
  ({ 3 })
```

In line one, the model extracts and labels the first row to **#Sentence pair (1)** and aligns the contents in that particular pair.

Line two is the English sentence that is to be translated to Kiswahili.

Line three is the translated Kiswahili sentence. This sentence has parts that need to be explained in order to make sense. First, is the "NULL" word is a default word that is added to the foreign vocabulary. This is because the word "the" does not have a matching word in Kiswahili and therefore should be aligned to the word “NULL” so that it does not go unaligned. The “NULL” word is simply a default word in cases where there is no corresponding foreign vocabulary.

The curly braces ({e f}), stores the word counts based on fractional count, what we call viterbi alignment. In the braces, there are up to three parameters. The ({e f}) implies the number of times English word e appears in the same sentence as French word f.

For example ({}) implies that the words “*NULL*”and “*jamhur*i” have zero occurrence in this particular sentence.

In the ({2,5}) implies the number of times the word “*of*” appears in the same sentence of Kiswahili, for example the word “*ya*” can match in most cases with “*of*” or “*for*”

3.5.2 Model process

Training Data

In table 3.3 above, is a corpus of Kiswahili-English words based on the constitution. The column for both Swahili and English were copied each in Eng.txt and Kisw.txt. These two files (parallel text or corpus) were fed into the model.

Training Data

The files entered into the model are converted to English vocabulary and Kiswahili vocabulary files, and stored as *Evcb* and *Fvcb* e.g

‘The’, ‘united’, ‘republic’, ‘of’ and ‘Tanzania’- will be stored in *Evcb*

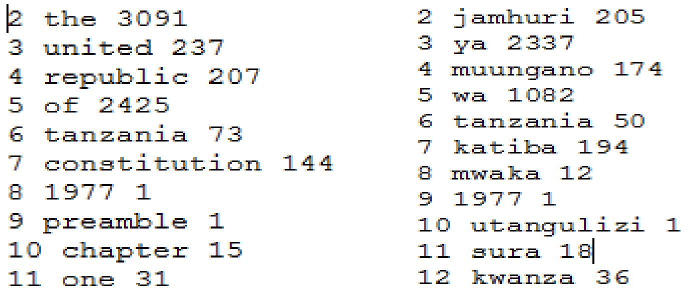
‘Jamhuri’, ‘ya’, ‘muungano’, ‘wa’ and ‘tanzania’- will be stored in *Fvcb*

The vocabulary files are now ready for training.

Learning Module (Dirichlet prior process)

The data is then passed to the Learning module. In this module there is a Dirichlet Process prior. Word translation probabilities will be treated as multi-nominal- distributed random variables with a sparse Dirichlet prior. The Dirichlet prior will prevent over fitting of parameters.

There is a tuning adjustment parameter module that is supposed to tune parameters obtained from the learning module.



```
2 the 3091
3 united 237
4 republic 207
5 of 2425
6 tanzania 73
7 constitution 144
8 1977 1
9 preamble 1
10 chapter 15
11 one 31
```

```
2 jamhuri 205
3 ya 2337
4 muungano 174
5 wa 1082
6 tanzania 50
7 katiba 194
8 mwaka 12
9 1977 1
10 utangulizi 1
11 sura 18
12 kwanza 36
```

Fig 3.6a and 3.6b Distribution of words in both E and F.

$$P(\mathbf{E},\mathbf{F},\mathbf{A}|\mathbf{T}) = \prod_{e=1}^{V_E}\prod_{f=1}^{V_F}(t_{e,f})^{N_{e,f}}\prod_{s}\frac{P(\mathbf{e})}{(I+1)^J} \quad (4)$$

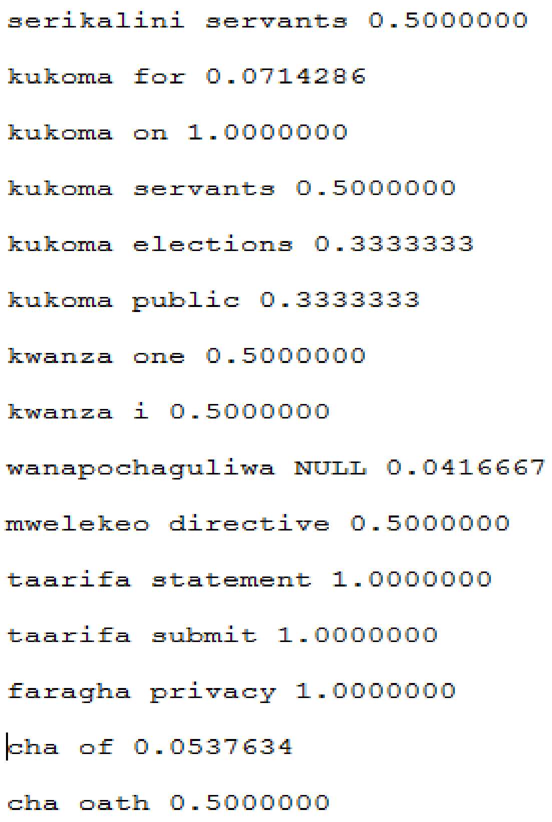
$$\mathbf{t}_e \sim \mathrm{Dirichlet}(\mathbf{t}_e;\mathbf{\Theta}_e)$$
$$f_j|\mathbf{a},\mathbf{e},\mathbf{T} \sim \mathrm{Multinomial}(f_j;\mathbf{t}_{e_{a_j}})$$

In the above formula, *E* is all the English words and *F* are all the Kiswahili words. *V* is the vocabulary words for both E and F.

The probability of the words in English and Kiswahili will be summed up from both Evcb and Fvcb as in **Figs 3.7(a)** and **(b)** and assigned probability distribution with theta as 0.0001, after finding the number of times a given word occurs in English as well as in Kiswahili.

$$\theta_{e,f} = \theta = 0.0001$$

The translation probabilities are the main output in this process e.g.



```
serikalini servants 0.5000000
kukoma for 0.0714286
kukoma on 1.0000000
kukoma servants 0.5000000
kukoma elections 0.3333333
kukoma public 0.3333333
kwanza one 0.5000000
kwanza i 0.5000000
wanapochaguliwa NULL 0.0416667
mwelekeo directive 0.5000000
taarifa statement 1.0000000
taarifa submit 1.0000000
faragha privacy 1.0000000
cha of 0.0537634
cha oath 0.5000000
```

Fig 3.7 Sample translation probabilities

In our example

“The united republic of Tanzania” - “*jamhuri ya muungano wa Tanzania”*

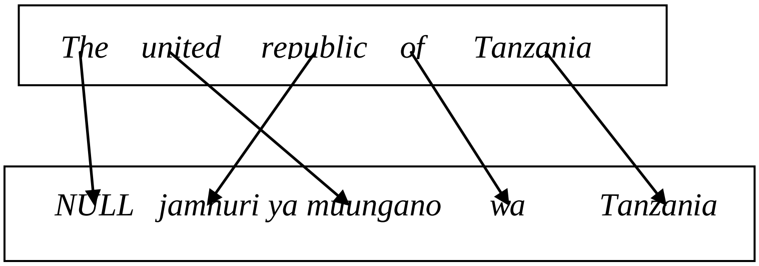
Based on fig 3.7, probable parameters were generated.

```
The null 1.00
The jamhuri- 1.00
United muungano 1.00
Jamuhuri republic 1.00
Of wa 1.00
Tanzania Tanzania 1.00
```

Bayesian Model Components

All the translation probabilities will be sending to the module that has transition, emission and fertility components. In this module, both phrase and word alignment try find a suitable position based the source text. It’s key to note that there are probabilities from which the model continues to learn.

This model components tries to take care of the sentence rules in the below example. Then align the sentence according to the higher probabilities obtained for every sentence pair.


The united republic of Tanzania
NULL jamhuri ya muungano wa Tanzania

Construction Module (Gibbs Sampling)

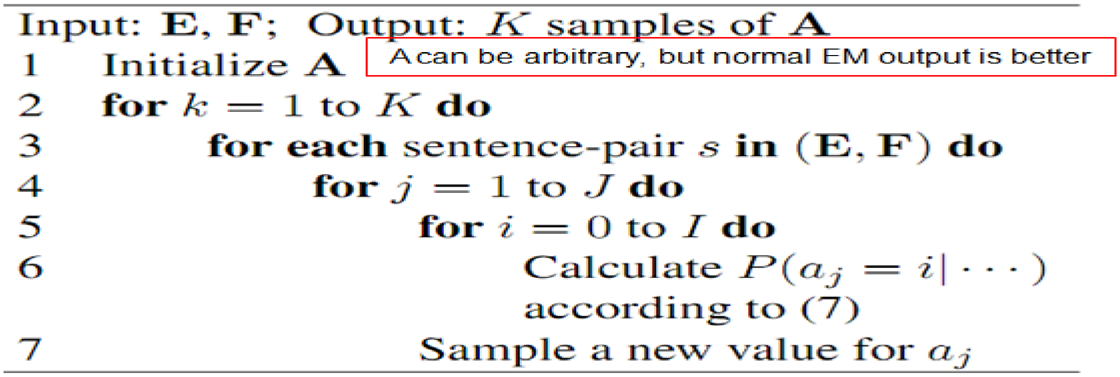
For better results phrase and word probabilities will be sending back and forth to another module called the construction module. This module has the Gibbs Sampling algorithm. Gibbs sampling, a stochastic inference technique produces random samples that converge in distribution to the desired posterior. In general, for a set of random variables z=$\{\mathbf{z}_j\}$, a Gibbs sampler iteratively updates the variables $\mathbf{z}_j$ one at a time by sampling its value from the distribution P($\mathbf{z}_j|\mathbf{z}^{\neg j}$), where **¬j** denotes the exclusion of the variable being sampled in **Fig 3.7.**
The Gibbs sampling formula is found as ;(Mermer &Saraclar, 2013)

$$P(a_j = i|\mathbf{E},\mathbf{F},\mathbf{A}^{\neg j};\mathbf{\Theta}) = \frac{N^{\neg j}_{e_i,f_j}+\theta_{e_i,f_j}}{\sum_{f=1}^{V_F} N^{\neg j}_{e_i,f}+\sum_{f=1}^{V_F}\theta_{e_i,f}}$$

To infer the posterior distribution of the alignments P (A|E,F;Θ) , Gibbs sampling , a stochastic inference technique that produces random samples that converge in distribution to the desired posterior. In general, for a set of random variables z=$\{\mathbf{z}_j\}$, a Gibbs sampler iteratively updates the variables $\mathbf{z}_j$ one at a time by sampling its value from the distribution P($\mathbf{z}_j|\mathbf{z}^{\neg j}$), where **¬j** denotes the exclusion of the variable being sampled.

In the generative model, **N^{-j}** denotes the number of times the source word type **e$_i$** is aligned to the target word type **fj** in **A**, not counting the current alignment link between **f$_j$** and **e$_{aj}$**.

Table 3.4: Gibbs Sampling Algorithm for IBM Model 1(Mermer & Saraclar, 2013)



```
Input: E, F;  Output: $K$ samples of A
    Initialize A        A can be arbitrary, but normal EM output is better
    for $k = 1$ to $K$ do
        for each sentence-pair $s$ in (E, F) do
            for $j = 1$ to $J$ do
                for $i = 0$ to $I$ do
                    Calculate $P(a_j = i|\cdots)$
                    according to (7)
                Sample a new value for $a_j$
```

The Gibbs algorithm, is supposed to group both the E and F sentence in pairs. It matches the E to F sentences and iterate for all other sentence pair. In this example, the output is as shown below;

```
# Sentence pair (1)

the united republic of tanzania

NULL ({ }) jamhuri ({ }) ya ({ 2 5 }) muungano ({ 4 }) wa ({ 1 })
tanzania
 ({ 3 })
```

Phrase Alignment module

After the Gibbs sampler has finished sampling, the possible outcome of alignments is passed to phrase alignment module which generate possible alignments from the probability outcome from the previous module.

Finally, we get the output as the target phrase as in **Fig 3.5**

3.6 Prototype Development.

3.6.1 System requirements

System requirements define how user requirement shall be met by the system.

The system requirements have been divided into two: Functional requirements and non-functional requirements

3.6.1.1 Functional requirements

The system should meet the following requirements:

1) Manage to align sentences from source to target, giving at least 60% accuracy in that domain.
2) The system should give similar accuracy in other domains.
3) Translations should not take more than thirty seconds i.e. time between submitting a text for translation and the resulting translated text.

3.6.1.2 Non-functional requirements

These requirements are not the core functionalities of the system. However, they play a role in ensuring a better presentation of functional requirements. They are:

1) Efficiency: thus ensuring better utilization of the limited memory resources and optimizing the speed of the machine.
2) The system will provide services in optimal speed of translation.
3) The system will have the ability to learn if new data is submitted

3.6.1.3 Hardware and software requirements

Minimum Deployment Environment

PC /Laptop core i3 and above – available with at least the following:

1) Linux operating system Ubuntu—10.04 plus
2) Moses decoder
3) Giza ++ for training the parallel corpora
4) Perl script interpreter
5) Bayesian Model for training the corpora

3.6.1.4 Merits of using Perl and Linux OS

1) Perl is able to interact with the Linux shell thus it can be able to pass commands to the terminal
2) Linux is open source hence most open source products are tailored for Linux in this case Moses decoder is tailored for Linux
3) With Linux it's easy to interact with the system resources and it is also easier to manage them.

3.6.2 Model platform and coding

The following tools and languages will be needed

- Perl Active state, (for coding the model, in Perl programming language)
- Linux Preferably Ubuntu 11 (operating system) and above,
- Moses Tool kit with Giza++
- Computer with minimum requirements of 2GB RAM and core i3 processor.

  All the key components in the model were coded using Perl, especially the Dirichlet prior process and the Gibbs sampling as seen in the Algorithm

3.6.3 Parallel Corpus Used

The data used in the model is a parallel corpus based on the constitution of Tanzania. It has Swahili and corresponding English sentence.

The corpus is in a word document and therefore it had to be papered (refer corpus preparation under data collection).

3.6.4 Actual Transformation

Having the model ready in Perl file, it is now time to test whether it is working.

Step 1: Copy the Swahili and English phrases from the word document to two separate files (One in Swahili and the other in English) and save them as dot txt or csv (.tx or .csv) extensions e.g Eng_Const.txt.

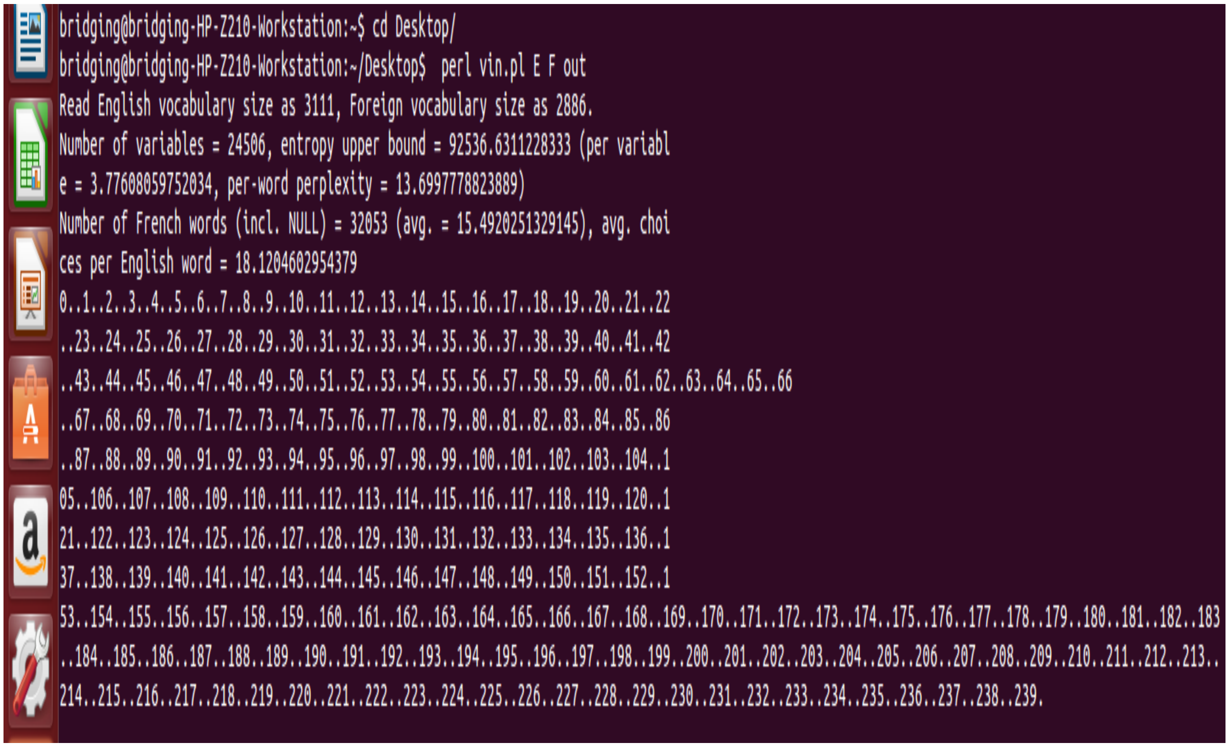
Step 2: Copy the two files on the desktop and rename the files as follows:

Eng_Const.txt as E

Kisw_Const.txt as F

Step 3: On linux terminal we run the following commands from the Desktop, our model is saved as vin.pl(this is the extension given to all the perl codes);

```
cd Desktop/
perl vin.pl E F out
```



```
bridging@bridging-HP-Z210-Workstation:~$ cd Desktop/
bridging@bridging-HP-Z210-Workstation:~/Desktop$  perl vin.pl E F out
Read English vocabulary size as 3111, Foreign vocabulary size as 2886.
Number of variables = 24506, entropy upper bound = 92536.6311228333 (per variabl
e = 3.77608059752034, per-word perplexity = 13.6997778823889)
Number of French words (incl. NULL) = 32053 (avg. = 15.4920251329145), avg. choi
ces per English word = 18.1204602954379
0..1..2..3..4..5..6..7..8..9..10..11..12..13..14..15..16..17..18..19..20..21..22
..23..24..25..26..27..28..29..30..31..32..33..34..35..36..37..38..39..40..41..42
..43..44..45..46..47..48..49..50..51..52..53..54..55..56..57..58..59..60..61..62..63..64..65..66
..67..68..69..70..71..72..73..74..75..76..77..78..79..80..81..82..83..84..85..86
..87..88..89..90..91..92..93..94..95..96..97..98..99..100..101..102..103..104..1
05..106..107..108..109..110..111..112..113..114..115..116..117..118..119..120..1
21..122..123..124..125..126..127..128..129..130..131..132..133..134..135..136..1
37..138..139..140..141..142..143..144..145..146..147..148..149..150..151..152..1
53..154..155..156..157..158..159..160..161..162..163..164..165..166..167..168..169..170..171..172..173..174..175..176..177..178..179..180..181..182..183
..184..185..186..187..188..189..190..191..192..193..194..195..196..197..198..199..200..201..202..203..204..205..206..207..208..209..210..211..212..213..
214..215..216..217..218..219..220..221..222..223..224..225..226..227..228..229..230..231..232..233..234..235..236..237..238..239.
```

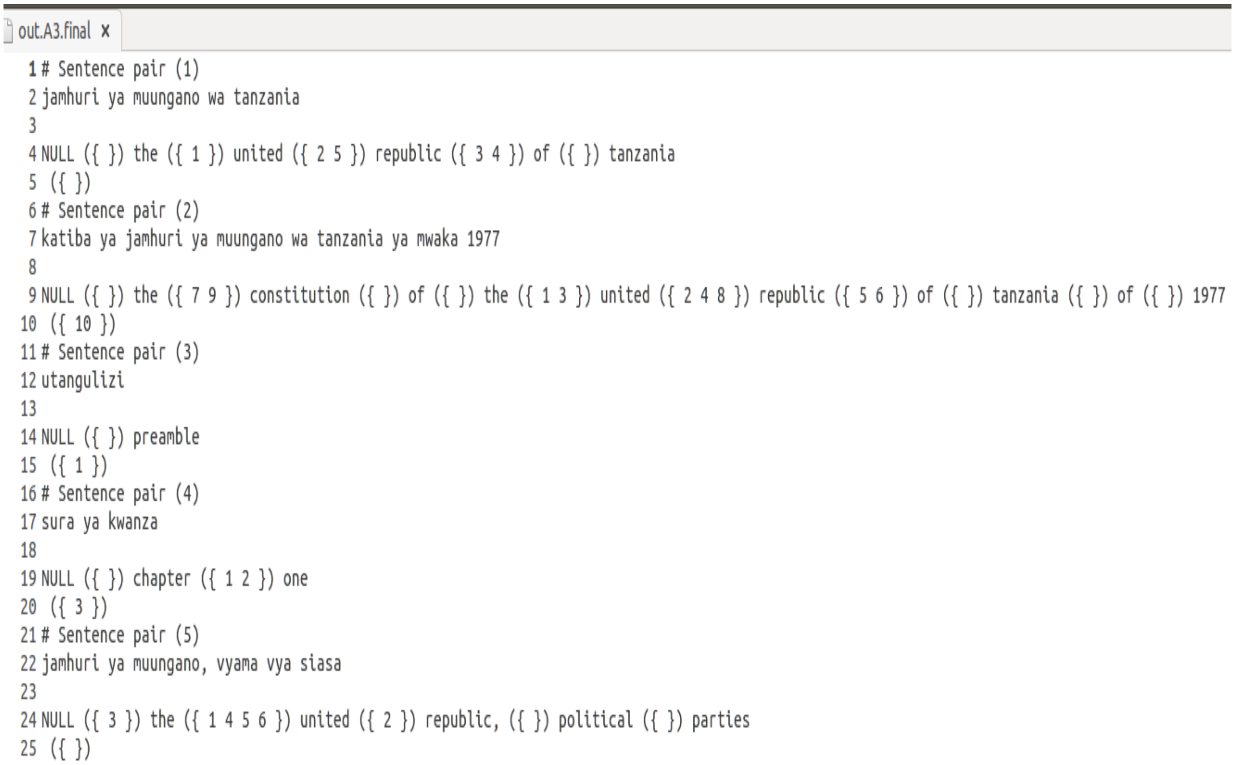
Fig 3.8: Model Execution in Perl on a Linux terminal.

The **E** denotes our source file and ***F*** denotes out target or foreign. ***E*** and ***F*** are the previous Eng_Constitution.txt and Kis_Constitution.txt respectively. They have simply been renamed.

The Bayesian model starts executing by;

- Reading the English vocabulary size.

- Reading foreign vocabulary size.
- Per word perplexity-: Perplexity can be considered to be a measure of on average how many different equally most probable words can follow any given word. Lower perplexities represent better language models, although this simply means that they `model language better', rather than necessarily work better.
- Average choice per English word.



```
# Sentence pair (1)
jamhuri ya muungano wa tanzania

NULL ({ }) the ({ 1 }) united ({ 2 5 }) republic ({ 3 4 }) of ({ }) tanzania
 ({ })
# Sentence pair (2)
katiba ya jamhuri ya muungano wa tanzania ya mwaka 1977

NULL ({ }) the ({ 7 9 }) constitution ({ }) of ({ }) the ({ 1 3 }) united ({ 2 4 8 }) republic ({ 5 6 }) of ({ }) tanzania ({ }) of ({ }) 1977
 ({ 10 })
# Sentence pair (3)
utangulizi

NULL ({ }) preamble
 ({ 1 })
# Sentence pair (4)
sura ya kwanza

NULL ({ }) chapter ({ 1 2 }) one
 ({ 3 })
# Sentence pair (5)
jamhuri ya muungano, vyama vya siasa

NULL ({ 3 }) the ({ 1 4 5 6 }) united ({ 2 }) republic, ({ }) political ({ }) parties
 ({ })
```

The next line that follows is the real execution line by line. The model then reads data into *@ECorpus* and *@FCorpus* and then outputs E and F vocabularies i.e. *Fvcb* and *Evcb* respectively. Once the model completes the execution, three files are generated which will be display on the desktop, namely Entropy, Time taken to Execute and *Out.A3.final*

Fig 3.9: Sentence Pair output

From the above sentence, five pairs of sentence have been extracted from the main output file. The pairs are Swahili and the corresponding English alignments.

On line 9, the model takes into consideration of the number of words (i.e. 10) and their scores as per from the entropy table.

CHAPTER FOUR: EVALUATION

4.1 Evaluation Methodology

In this section, I will be discussing the evaluation methodology .Have large data set to draw my data set from; the methodology of choice will be simple Random Sampling. Each individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process, and each subset of *k* individuals has the same probability of being chosen for the sample as any other subset of *k* individuals

4.2 Fertility and Over fitting

Bayesian model solve key weaknesses of EM which are;

- Explanation of the training data by over fitting the parameters, a good example is that of the rare words.
- Tendency of getting stuck in local maximum of likelihood.
- Assumption that there is one fixed value of parameters that explain the data.
- Inability to take into account fertility.

In this section, we discuss how Bayesian model addresses at least two if not all the weaknesses.

4.2.1Experimental Procedure

The EM algorithm finds the value of parameters that maximizes the likelihood of the observed variables. However, with many parameters to be estimated without any prior, EM tends to explain the training data by over fitting the parameters.

Over fitting occurs when a statistical model describes random error or noise instead of the underlying relationship. Over fitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model which has been over fit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data (Leinweber, 2007)

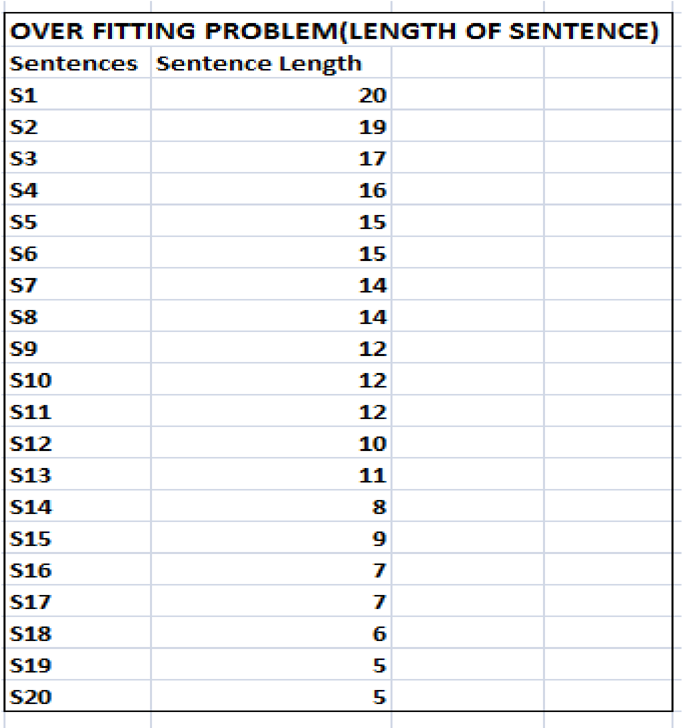
Over fitting arise due to

i. Increase in the length of sentence- determined by the number of words in the sentences.

ii. Number of new words introduced.

For the case of (i), twenty sentences out of 500 hundred sentences were sampled using the simple random sampling method. The sentences were then subjected to both EM and Bayesian Models and compared the results in a graphical representation.

Table 4.1: Sample Sentences (Sentence Length)



OVER FITTING PROBLEM(LENGTH OF SENTENCE)			
Sentences	Sentence Length		
S1	20		
S2	19		
S3	17		
S4	16		
S5	15		
S6	15		
S7	14		
S8	14		
S9	12		
S10	12		
S11	12		
S12	10		
S13	11		
S14	8		
S15	9		
S16	7		
S17	7		
S18	6		
S19	5		
S20	5		

The sampling was performed using the RAND function in excel. From a total sentence of 500, twenty were sampled out and rearranged from S1….S20 in an excel table from which a graph was plotted as in Fig 4.1. The sentences sampled had different length (determined by the number of words in a sentence). From the above table, it was observed that the longest sentence had 20 words (inclusive all parts of a sentence) while the shortest sentence had had 5 words.

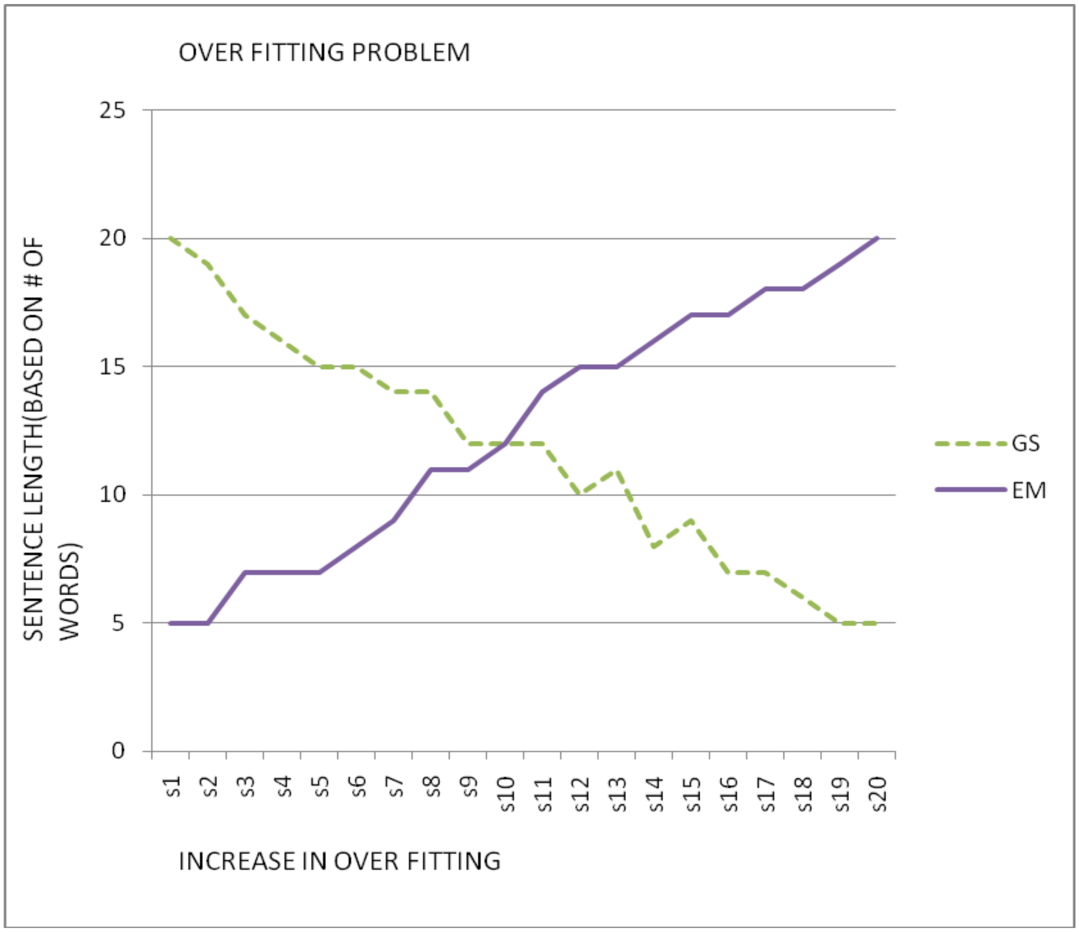


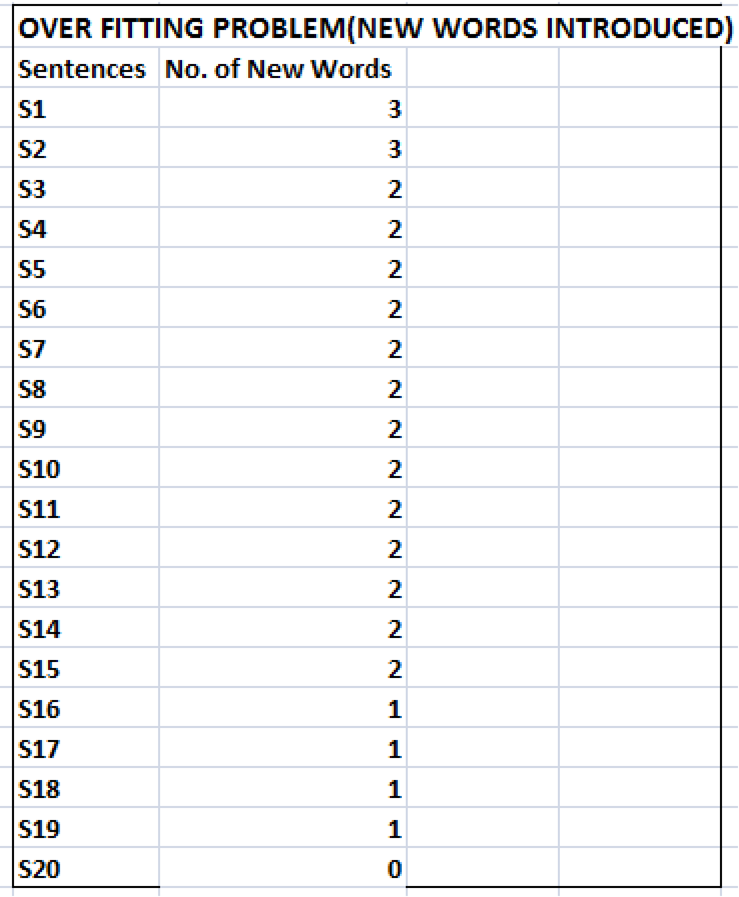
Fig 4.1 Over fitting based on Length of sentence

From the graph above, over fitting increase with increase in the length of the sentences in the case of EM model, while in Bayesian model, there are no traces of over fitting since it has been taken care of.

ii) Number of New words introduced.

The sampling was performed using the RAND function in excel. Again from a total sentence of 500, twenty were sampled out and rearranged from S1….S20 in an excel table from which a graph was plotted as in Fig 4.2. The sentences sampled had different length (determined by the number of new words introduced). From the below table, it was observed that the sentence with most new words introduced had 3, while the sentence with the fewest new words had 0.

Table 4.2: Sample Sentences (New words introduced)



OVER FITTING PROBLEM(NEW WORDS INTRODUCED)			
Sentences	No. of New Words		
S1	3		
S2	3		
S3	2		
S4	2		
S5	2		
S6	2		
S7	2		
S8	2		
S9	2		
S10	2		
S11	2		
S12	2		
S13	2		
S14	2		
S15	2		
S16	1		
S17	1		
S18	1		
S19	1		
S20	0		

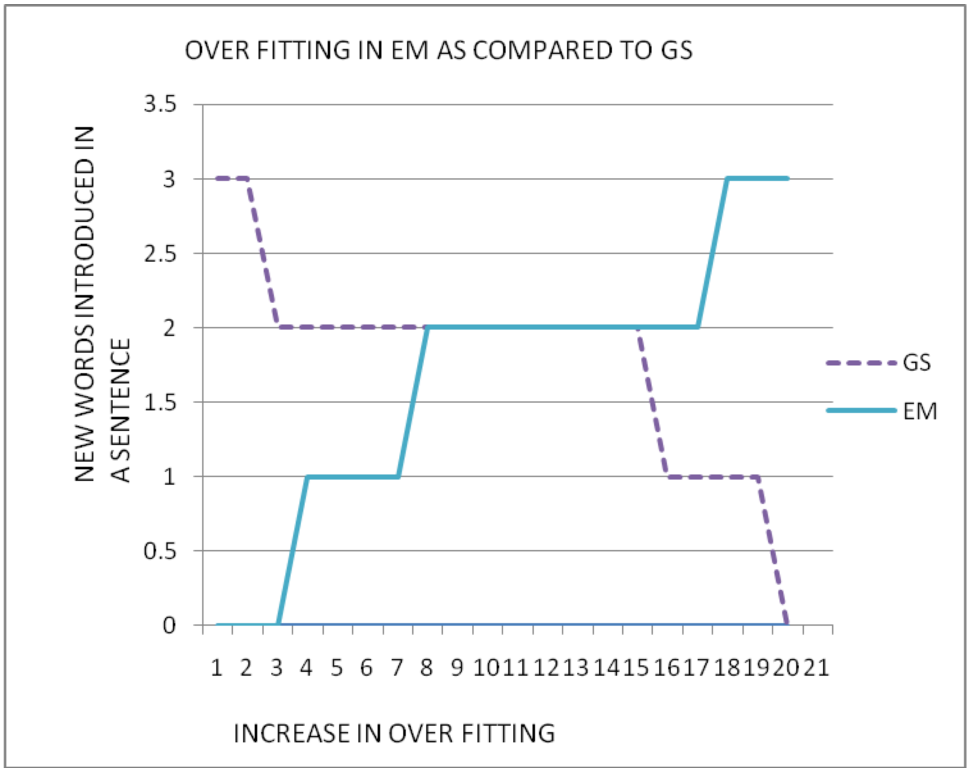


Fig 4.2: Over fitting based on new words introduced.

Let's look at an example of over fitting situation

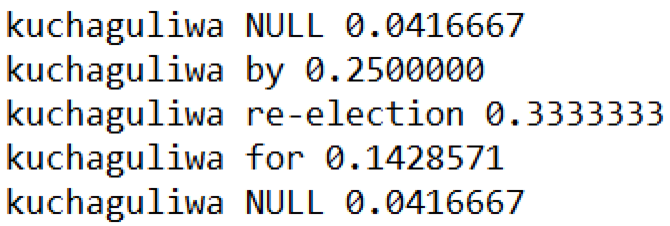
Illustration of how Bayesian Model overcomes over fitting.

Example 1: Length of Sentence (Determined by number of words)

In this illustration, over fitting is investigated based on the sentence length and how this affects EM as a model.

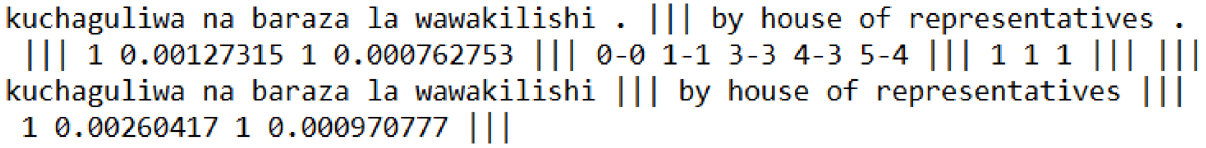
The sentence has 15 words in the English sentence. The significance of this experiment is to show how over fitting is caused in EM due to the length of the sentence. The longer the sentence results to high rates of over fitting. The experiment shows how Bayesian model overcomes this problem.

Below are probabilities for the word "*Kuchaguliwa*" from EM model output



```
kuchaguliwa NULL 0.0416667
kuchaguliwa by 0.2500000
kuchaguliwa re-election 0.3333333
kuchaguliwa for 0.1428571
kuchaguliwa NULL 0.0416667
```

EM Output (From the phrase table)



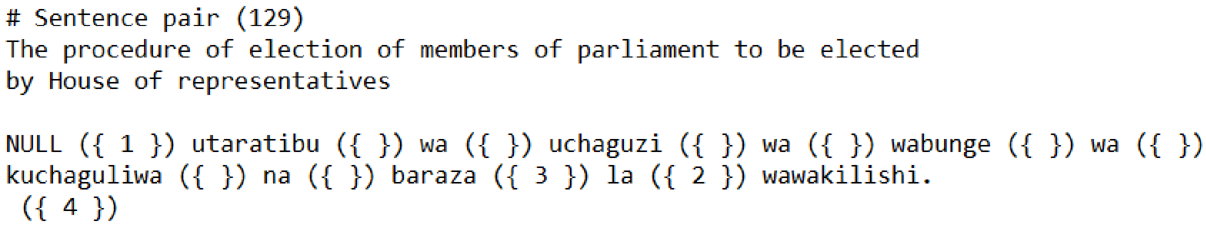
```
kuchaguliwa na baraza la wawakilishi . ||| by house of representatives .
 ||| 1 0.00127315 1 0.000762753 ||| 0-0 1-1 3-3 4-3 5-4 ||| 1 1 1 ||| |||
kuchaguliwa na baraza la wawakilishi ||| by house of representatives |||
 1 0.00260417 1 0.000970777 |||
```

From the results, EM model in Giza++ tried to sum the probabilities of the word '*Kuchaguliwa*' and it did not come up with probable English correspondence.

Therefore we see over fitting of the sentence to '*house of representatives'* only hence ignoring the word '*Kuchaguliwa*' which plays a very important role as a parameter

Bayesian Model Output

On the other hand, the Bayesian model overcomes over fitting by taking into consideration all the parameters and assigns them to possible matches in the foreign word. Therefore during alignment, there is no single (parameter) word that will be over fit to the foreign sentence.



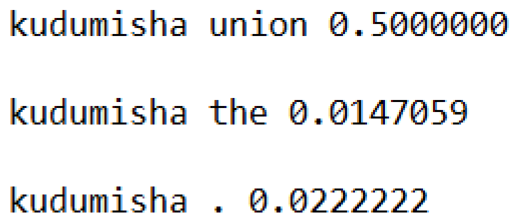
```
# Sentence pair (129)
The procedure of election of members of parliament to be elected
by House of representatives

NULL ({ 1 }) utaratibu ({ }) wa ({ }) uchaguzi ({ }) wa ({ }) wabunge ({ }) wa ({ })
kuchaguliwa ({ }) na ({ }) baraza ({ 3 }) la ({ 2 }) wawakilishi.
 ({ 4 })
```

Example 2: Fertility Problem New words

Over fitting problem is also evident in situation where we have high fertility rate. EM is unable to handle fertility arising towards the foreign sentence. The example below, the sentence “Kudumisha muungano” has lots of words introduced in English, that is, “management of affairs concerning union matters”. From the probabilities below, EM fail to handle fertility and hence miss-aligns the word “*Kudumisha*” to “*Union*”.

The word “*Kudumisha*” can be considered to be new word. Therefore the introduction of this word in the sentence has an effect on over fitting in the EM model. New words tend to be ignored by the EM model hence get miss-aligned.

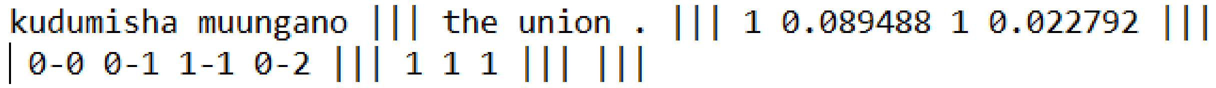
Probability for the word ‘*Kudumisha*’



```
kudumisha union 0.5000000

kudumisha the 0.0147059

kudumisha . 0.0222222
```

EM Output (From phrase table)

It is evident in this EM output that the alignment is wrong.



```
kudumisha muungano ||| the union . ||| 1 0.089488 1 0.022792 |||
| 0-0 0-1 1-1 0-2 ||| 1 1 1 ||| |||
```

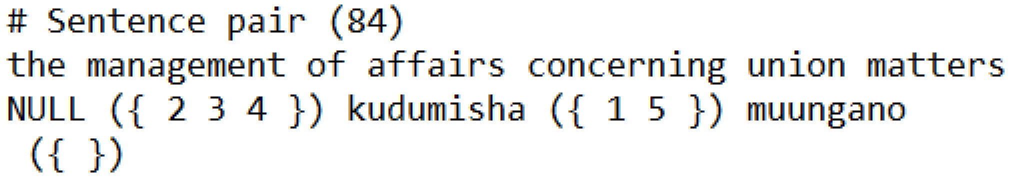
The results indicate that EM matched ‘*kudumisha*’ to Union which is not true. The word was picked without considering the neighboring words/parameters which matters a lot in the alignment of the sentence.

It can also be interpreted that EM ignores the said word or simply miss-aligns it because it is rare in the phrase table.

Bayesian Output

On the other hand, Bayesian model takes into consideration all the word(s) and the neighbors too. It is shown that Bayesian model, takes care of fertility and therefore able to give the correct alignment of the sentence "*the management of affairs concerning union matters*".

From the Bayesian output, the new word is taken care of, hence assigned appropriate probability. This eventually leads to better alignment since the other parameters are considered.



```
# Sentence pair (84)
the management of affairs concerning union matters
NULL ({ 2 3 4 }) kudumisha ({ 1 5 }) muungano
 ({ })
```

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

5.1 Introduction

Bantu Language of Swahili is spoken by more than fifty million people in East Africa and central Africa, however it is surprisingly resource scarce from a language technological point of view (De Pauw et al. 2009).An increasing number of publication however are showing that carefully selected procedures can indeed boost language technology for Swahili.

Word alignment can be considered the backbone of Statistical Machine Translation. Having looked at the most classical alignment model EM, there is need to kick start the Swahili language with better language technology.

The research developed a Bayesian model based on Gibbs sampling algorithm that was used to improve the alignment of the native language Swahili.

Sample data was a corpus of about 300 thousand words, approximately 180 thousand sentences. However due to limited time, the research dwelt on a domain based on the Tanzanian constitution with a total of 56 thousand sentences- end to end, approximately two thousand words.

5.2: Discussion of Findings and Contribution to Society

This section is based on the observations in chapter four. It is meant to discuss what the research meant, what it informs, and conclusions from the observation and drawing of comparisons to other related works/publications.

Performance

i) **BLEU score**

BLEU scores better with for Gibbs compared to EM, given

Table 5.1: BLEU SCORES OF EM ALIGNMENTS AND INFERENCE METHODS ON THE 1M-SENTENCE ARABIC-ENGLISH TRANSLATION

Method	Model 2 EM	Model 2 GS
BLEU	46.97(+-0.15)	47.17(+- 0.14)

In the above results, the Bayesian inference improves the mean BLEU score by 0.2 BLEU regardless of the domain (Mermer *et al.*, 2013).

Table 5.2 BLEU scores of EM and GS models on 56K sentence on English-Kiswahili Translations.

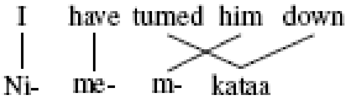
Method	EM Model	Model 2 GS
BLEU	29.3(+-1)	32.6(+-1)

In Table 5.2, the BLEU score results are less compared to those of (Mermer *et al.*, 2013). The score is way below for the English-Kiswahili sentences.

This was attributed to the following;

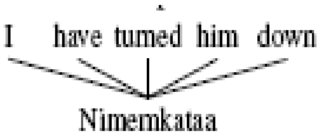
- Morphological differences of Kiswahili compared to other languages. For example, most Kiswahili words have a tendency of yielding more words that are supposed to be aligned.

  e.g



The isolated Swahili morphemes can more easily be linked to their English counter parts since there will be more linguistic evidence in the parallel corpus, linking for example *ni* to *I* and *m* to *him.*

- Agglutinating- This is a process in linguistic morphology derivation in which complex words are formed by stringing together morphemes, each with a single grammatical or semantic meaning (Ng'ang'a, 2005). Languages that use agglutination widely are called agglutinative languages. Kiswahili language has strong agglutination.



ii) Fertility Distribution

Fertility of a source word is defined as the number of target words aligned to it. In general, we expect the fertility values close to the word token ratio between the languages to be the most frequent and high fertility values to be rare. The “garbage collecting” effect was observed more in EM-estimated alignments.

In both alignment tasks, Bayesian method result in fewer high-fertility alignments compared to EM.

In their recent study (Mermer et al., 2013)they deduced that, for example, in English-Arabic Model 1 alignment using EM, 1.2% of the English source tokens are aligned with *nine or more* Arabic target words, corresponding to 22.3k total occurrences or about 0.4 occurrence per sentence.

In our Kiswahili-English alignment Bayesian model has high fertility rate compare to EM. For example, in the same corpus used (Tanzanian constitution), English has 3111 and Kiswahili has 2886 vocabularies compared to EM with 2544 and 2695 vocabularies respectively.

iii) Alignment Error Rate

The table below shows alignment Error Rate using a publicly available 500-sentence manually-aligned reference set. The Bayesian methods achieve better AERs than EM in both alignment directions, i.e. Eng-Swa

Table 5:3 Alignment Error Rate (%) in both models.

Training Model	Correctly Aligned	Wrongly Aligned	Missing phrases
GS	89	8	3
EM	75	20	5

iv) Comparison

Both the models are alignment models that are used to predict probable samples in a given distribution.

In general, EM and Gibbs sampling can handle inputs that are far more complex than the example shown. Many programs that are available by Web can manage hundreds of sequences with thousands of base pairs each. Furthermore, there are several variations to these two algorithms that attempt to improve predictive value based on prior knowledge of the motif's characteristics.

- **Computation**

Kiswahili-English corpus (Tanzanian constitution), Bayesian model is slower compared to EM model, for instance in a 80 sentence pair, the model took approximately 22 minutes, while EM took lesser time on the same data set approximately between 11- 14 minutes. Note that the time varied on every run that was made.

However the above can be sped up by parallel computing, whereby threading can be introduce (This can be future work.)

In their study (Mermer et al., 2013), deduced that in a 100 pair sentence, it cost 18 minutes.

- **Performance**

Bayesian outperforms classical EM in BLEU. Apart from that, it also addresses the rare word problem through the use of Dirichlet prior process that attaches importance to all the surrounding parameters. Bayesian has a smaller phrase table than EM and there better memory utilization than EM.

(Mermer et al., 2013) records that Bayesian out performs Classical EM of up to 2.99 in BLEU score when performing alignments in English-Turkish, English-Arabic and English-Czech languages.

From the results, Bayesian Model showed that it compares favorably to EM estimation in terms of translation BLEU scores as in **Table 5.2**. The largest improvement was observed when data is sparse, e.g., in the cases of smaller corpora and/or more morphological complexity.

The proposed method successfully overcomes the well-known "garbage collection" problem of rare words in EM-estimated current models and learns a compact, sparse word translation distribution with more training vocabulary coverage.

Contribution of this research was to try and apply a better alignment method that addresses weaknesses of a classical EM model especially the over fitting problem which leads to "garbage collection" effect.

This piece of work should be an eye opener to other researchers of likeminded. This kind of research has not been fully studied. Most current research works are basically touching on other international known languages but not native Swahili language.

5.3 Limitation of study

Throughout the research period, it was evident that there was little work done on Swahili language. There are a myriad of limitation right from data collection to implantation;

i) Scarcity of Swahili related corpora, and the few that are available, it is hard to reach them.

ii) The model does not really give results expected with other internationally known languages due to inadequate research and improvement of the Swahili language. For instance there are some words in English that have yet been translated to Swahili.

iii) Morphological, syntactical and synonyms in Swahili have different structures with regard to context.

iv) The Bayesian model is significantly slow especially on large pairs of sentences. This can be sped up by use of parallel computing (which is a whole research on its own).

5.4 Recommendation for Future work

I recommend that more research work should be done in this area, since my work is just an eye opener in this field to other likeminded researchers.

Swahili being our native language, we should have a language bank that has corpora for both international languages and local languages, to enhance as well as facilitate would be researchers.

Bayesian model is slow on large pairs of sentences, this call for research in the field of parallel computing and threading.

REFERENCES

1. Agirre, E. & Martinez, D. (2001), Knowledge sources for word sense disambiguation, *in* R. M. Vaclav Matousek, Pavel Mautner & K. Tauser, eds, 'Proceedings of the Fourth International Conference on Text, Speech and Dialogue', Vol. 2166, Springer Verlag
2. Banerjee, S. and Lavie, A. (2005) "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" in *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005*

   Bayesian Word Alignment and Gibbs Sampling" in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 5, may 2013, pp. 1090-1093.
3. Beal, Matthew J. 2003. Variational Algorithms for Approximate Bayesian Inference. Ph.D. thesis, University College London.
4. Blunsom, P., T. Cohn, C. Dyer, and M. Osborne, "A Gibbs sampler for phrasal synchronous grammar induction," in *Proc. ACL-AJCNLP*, Suntec, Singapore, Aug. 2009, pp. 782–790.
5. Brown, P.,F V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguist.* vol. 19, no. 2, pp. 263–311, 1993.
6. Chiang, Z. J. Graehl, K. Knight, A. Pauls, and S. Ravi, "Bayesian inference for finite-state transducers," in *Proc. NAACL-HLT*, Los Angeles, CA, USA, Jun. 2010, pp. 447–455.
7. Chiang, D. "Hierarchical phrase-based translation," *Comput. Linguist.* vol. 33, no. 2, pp. 201–228, 2007.
8. Chung, T. and D. Gildea, "Unsupervised tokenization for machine translation," in *Proc. EMNLP*, Singapore, Aug. 2009, pp. 718–726.

   Conferences 1999-2001, Hämeenlinna Häme Polytechnic.

9. De Pauw, G., de Schryver, G.-M. & Wagacha, Peter Waignjo. (2009)a. A corpus-based survey of four electronic Swahili–English bilingual dictionaries. Lexikos, 19, p. 340–352.

10. De Pauw, G., Wagacha, P.W. & de Schryver, G.-M. (2009). The SAWA corpus: a parallel corpus English - Swahili. In G. De Pauw, G.-M. de Schryver & L. Levin (Eds.), Proceedings of the First Workshop on Language Technologies for African Languages (AfLaT 2009). Athens, Greece: Association for Computational Linguistics, pp. 9–16.

11. Dempster, A., N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.

12. DeNero, J., A. Bouchard-Côté, and D. Klein, "Sampling alignment structure under a Bayesian translation model," in *Proc. EMNLP*, Honolulu, HI, USA, Oct. 2008, pp. 314–323.

13. Gao, J. and M. Johnson, "A comparison of Bayesian estimators for unsupervised hidden Markov model POS taggers," in *Proc. EMNLP*, Honolulu, HI, USA, Oct. 2008, pp. 344–352.

14. Goldwater, S. and T. Griffiths, "A fully Bayesian approach to unsupervised
in *Proc. ACL*, 2012, pp. 311–319.

15. Johnson, Mark. 2007. Why doesn't EM find good HMM POS-taggers? In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 296–305. Association for Computational Linguistics, Prague, Czech Republic.
Jun. 2007, pp. 744–751.

16. Koehn, *D. Statistical Machine Translation*. Cambridge, U.K.: Cambridge

17. Lavie, A., Sagae, K. and Jayaraman, S. (2004) "The Significance of Recall in Automatic Metrics for MT Evaluation" in *Proceedings of AMTA 2004, Washington DC. September 2004*

Lecture Notes in Computer Science series, Plzen (Pilsen), Czech Republic, pp. 1–10.

18. Leinweber, D. J. (2007). "Stupid Data Miner Tricks". *The Journal of Investing* **16**: 15–22.

19. Mermer, C. and M. Saraclar, "Improving Statistical Machine Translation Using

20. Michie, D., Spiegelhalter, D. & Taylor, C. (1994), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York.
21. Mitchell, T. (1997), *Machine Learning,* McGraw-Hill, New York.
22. Mochihashi, D., T. Yamada, and N. Ueda, “Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling,” in *Proc. ACL-AJCNLP*, Suntec, Singapore, Aug. 2009, pp. 100–108.
23. Moore, R.,C “Improving IBM word alignment model 1,” in *Proc. ACL*, Barcelona, Spain, Jul. 2004, pp. 518–525.
24. Ng’ang’a, W., “Word Sense Disambiguation of Swahili: Extending Swahili Language Technology with Machine Learning” *University of Helsinki, November, 2005,pp. 4-6.*
25. Nguyen, T., S. Vogel, and N. A. Smith, “Nonparametric word segmentation for machine translation,” in *Proc. COLING,* 2010, pp. 815–823.
26. Papineni, K., Roukos, S., Ward, T., Henderson, J and Reeder, F. (2002). “*Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results*” in Proceedings of Human Language Technology 2002, San Diego, pp. 132–137

    Part-of-speech tagging,” in *Proc. ACL,* Prague, Czech Republic,
27. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. (2007) "Moses: Open Source Toolkit for Statistical Machine Translation". *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.*
28. Riley,D. and D. Gildea, “Improving the IBM alignment models using variational Bayes,” in *Proc. ACL: Short Papers,* 2012, pp. 306–310.
29. Ruohotie, R., Nokelainen, P., Tirri, H. & Silander, T. (2001), *Modeling Individual and Organizational Prerequisites of Professional Growth - Papers Presented at International*
30. Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In D. Jones (Ed.), Proceedings of the International Conference on New Methods in Language Processing. Manchester, UK: UMIST, pp. 44–49.

31. Tetko, I. V.; Livingstone, D. J.; Luik, A. I. (1995). "Neural network studies. 1. Comparison of Overfitting and Overtraining". **35** (5): 826–833. Univ. Press, 2010.

32. Vaswani, A., L. Huang, and D. Chiang, “Smaller alignment models for better translations: Unsupervised word alignment with the -norm,”

33. Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. (2000), Self-Organizing Map in Matlab: The SOM toolbox, *in* ‘Proceedings of the Matlab DSP Conference, Espoo, Finland’.

34. Vogel, S., H. Ney, and C. Tillmann, “HMM-based word alignment in statistical translation,” in *Proc. COLING,* 1996, pp. 836–841.

35. Xu, J J. Gao, K. Toutanova, and H. Ney, “Bayesian semi-supervised Chinese word segmentation for statistical machine translation,” in *Proc. COLING,* Manchester, U.K., Aug. 2008, pp. 1017–1024.

36. Zhao, B. and E. P. Xing, “BiTAM: Bilingual topic admixture models for word alignment,” in *Proc. COLING-ACL: Poster Sessions*, Sydney, Australia, Jul. 2006, pp. 969–976.

37. Zhao, Sand D. Gildea, “A fast fertility hidden Markov model for word alignment using MCMC,” in *Proc. EMNLP*, Cambridge, MA, USA, Oct. 2010, pp. 596–605.

APPENDIX ONE: BAYESIAN MODEL CODE

i) Initialization(Data structures and Dirichlet prior)

```
use strict;
$| = 1;

die "Usage: $0 fcorpus_in ecorpus_in out [initial_alignments]\n" if
(@ARGV < 3);

# Dirichlet prior for translation probabilities
my $beta = 0.0001;

# MCMC parameters
my $burninIters = 1000;  # skip this many iterations in the beginning,
first readout will be in iteration N + 1 (provided overlaps with "lag"
-- for now)
my $lagIters = 1;  # Read out every N iterations; set to 1 to sample
at every iteration
my $numReadouts = 1000;
my $randomseed = 0;

# Useful info about how this program was run
open(CALL, ">$ARGV[2]_experiment_info");
print CALL "Started at " . localtime() . "\n";
print CALL $0;
for my $arg (@ARGV) {
     print CALL " $arg";
}
print CALL "\n";
print CALL "beta = $beta\n";
print CALL "(burn-in, lag, numreads) = ($burninIters, $lagIters,
$numReadouts)\n";
print CALL "random-seed = $randomseed\n";
close(CALL);

# Data structures
my @Ecorpus;
my @Fcorpus;
my @A;
my %Count;
my %SumCount;
srand($randomseed);  # for replicable results
```

ii) Reading data and Displaying Evcb and Fvcb

```
 # Read data into @Ecorpus and @Fcorpus, and compile E and F
vocabularies
open(F, $ARGV[0]);
open(E, $ARGV[1]);
my $s = 0;
my %Evcb;
my %Fvcb;
$Fvcb{"NULL"}++;  # Add NULL word to the foreign vocabulary
while (my $fline = <F>) {
      my $eline = <E>;
      $s++;
      chomp $fline;
      chomp $eline;
      my @Fwords = split(/ +/, $fline);
      my @Ewords = split(/ +/, $eline);
      $Fcorpus[$s][0] = "NULL";  # only in the source language, i.e., f
      for my $j (0 .. $#Fwords) {
            $Fcorpus[$s][$j + 1] = $Fwords[$j];
            $Fvcb{ $Fwords[$j] }++;
      }
      for my $i (0 .. $#Ewords) {
            $Ecorpus[$s][$i] = $Ewords[$i];
            $Evcb{ $Ewords[$i] }++;
      }
}
my $Evcbsize = 0;
for my $e (keys %Evcb) {
      $Evcbsize++;
}
my $Fvcbsize = 0;
for my $f (keys %Fvcb) {
      $Fvcbsize++;
      $SumCount{$f} = 0;  # initialize
}
print "Read English vocabulary size as $Evcbsize, Foreign vocabulary
size as $Fvcbsize.\n";

# Initialize
# EITHER from user-provided file
if (defined($ARGV[3])) {
      open (INIT, $ARGV[3]);
      for my $s (1 .. $#Fcorpus) {
            <INIT>;
            my $eline = <INIT>;
            my $fline = <INIT>;
            chomp $eline;
            chomp $fline;
            my @Ewords = split(/ +/, $eline);
            my @Fwords = split(/ +/, $fline);
            my $j = 0;
            my $k = 0;
            while ($k <= $#Fwords) {
                  my $f;
                  while ($Fwords[$k] ne "({") {
                        $f = $Fwords[$k];
                        $k++;
                        last if ($k > $#Fwords);
                  }
                  $k++;
                  last if ($k > $#Fwords);
                  while ($Fwords[$k] ne "})") {
                        # 1. Alignment
                        $A[$s][ $Fwords[$k] - 1 ] = $j;

                        # 2. Alignment counts
                        $Count{$f}{ $Ewords[$Fwords[$k] - 1] }++;

                        # 3. Sum of alignment counts for each fj
                        $SumCount{$f}++;
                        $k++;
                  }
                  $j++;
            }
      }
} else {
# OR from empirical evidence
      # Naively assume uniform probability and compute fractional
counts
      my %FractionalCount;
      for my $s (1 .. $#Fcorpus) {
            for my $ei (@{$Ecorpus[$s]}) {
                  for my $fj (@{$Fcorpus[$s]}) {
                        $FractionalCount{$ei}{$fj}++;
                  }
            }
      }
```

iii) Performing Alignment, alignment counts and printing to file.

```
for my $s (1 .. $#Fcorpus) {
            my $i = 0;
            for my $ei (@{$Ecorpus[$s]}) {

                  # 1. Alignments (choose Viterbi alignments based on
fractional counts)
                  my $bestcount = 0;
                  my $bestfj = "";
                  my $bestj = -1;
                  my $j = 0;
                  for my $fj (@{$Fcorpus[$s]}) {
                        if ($fj eq "NULL") {  # don't want to initialize
with NULL alignment
                              $j++;
                              next;
                        }
                        if ($FractionalCount{$ei}{$fj} > $bestcount) {
                              $bestcount = $FractionalCount{$ei}{$fj};
                              $bestfj = $fj;
                              $bestj = $j;
                        }
                        $j++;
                  }
                  $A[$s][$i] = $bestj;

                  # 2. Alignment counts
                  $Count{$bestfj}{$ei}++;

                  # 3. Sum of alignment counts for each fj
                  $SumCount{$bestfj}++;

                  $i++;
            }
      }
} # Initialization
#&printAlignmentsToFile($ARGV[2]);
```

iv) Gibbs sampling(Getting the probability from the sample)

```
# Take successive samples by Gibbs sampling
my @Acollect;
my $t = 0;
while ($numReadouts) {
    print "$t";
    $totalentropy = 0;

    # Keep track of readouts
    my $isReadoutIter = 0;
    $isReadoutIter = 1 if ($t % $lagIters == 0);
    $isReadoutIter = 0 if ($t < $burninIters);
    $numReadouts-- if $isReadoutIter;
    print "*" if $isReadoutIter;

    # Sample each variable
    for my $s (1 .. $#Fcorpus) {
        print "." if ($s % 1000 == 0);
        my $i = 0;
        for my $ei (@{$Ecorpus[$s]}) {

            # Decrement before sampling
            $Count{$Fcorpus[$s][ $A[$s][$i] ]}{$ei}--;
            $SumCount{$Fcorpus[$s][ $A[$s][$i] ]}--;

            # Calculate relative probabilities and draw a sample
from this distribution
            my @probvector;
            my $j = 0;
            for my $fj (@{$Fcorpus[$s]}) {
                my $p_i_j;
                if (defined($Count{$fj}{$ei})) {
                    $p_i_j = ($Count{$fj}{$ei} + $beta) /
($SumCount{$fj} + $Evcbsize * $beta);
                } else {
                    $p_i_j = $beta / ($SumCount{$fj} +
$Evcbsize * $beta);
                }
                $probvector[$j] = $p_i_j;
                $j++;
            }
            my ($outcome, $entropy) = &do_sampling(@probvector);
            $A[$s][$i] = $outcome;
```

```
                  $Acollect[$s][$i][ $outcome ]++ if $isReadoutIter;
                  $totalentropy += $entropy;

                  # Increment according to the new sample
                  $Count{$Fcorpus[$s][ $A[$s][$i] ]}{$ei}++;
                  $SumCount{$Fcorpus[$s][ $A[$s][$i] ]}++;

                  $i++;
            }
      }

      my $log2entropy =  $totalentropy *
$loge_to_log2_conversion_factor;
      printf ENT "%.2f %.6f\n", $log2entropy, $log2entropy /
$numvariables;
      #&printAlignmentsToFileGIZA("$ARGV[2]_iter$t") if $isReadoutIter;
      $t++;
}
close(ENT);
```

v) Print out of a Giza++ like file

```
sub printAlignmentsToFile {
      open (OUT, ">@_");
      for my $s (0 .. $#Fcorpus) {
            my $i = 0;
            for my $ei (@{$Ecorpus[$s]}) {
                  print OUT "A[$s][$i:$ei]=$A[$s][$i]:$Fcorpus[$s][
$A[$s][$i] ]\n";

                  $i++;
            }
      }
      close(OUT);
}


sub printAlignmentsToFileGIZA {
      open (OUT, ">@_");
      for my $s (1 .. $#Fcorpus) {
            print OUT "# Sentence pair ($s)\n";
            for my $ei (@{$Ecorpus[$s]}) {
                  print OUT "$ei ";
            }
            print OUT "\n";
            my $j = 0;
            for my $fj (@{$Fcorpus[$s]}) {
                  print OUT "$fj ({ ";
                  my $i = 1;  # to match the GIZA++ output format
                  for my $ei (@{$Ecorpus[$s]}) {
                        print OUT "$i " if ($A[$s][$i - 1] == $j);
                        $i++;
                  }
                  print OUT "}) ";
                  $j++;
            }
            print OUT "\n";
      }
      close(OUT);
}
```

APPENDIX TWO: BAYESIAN OUTPUT SAMPLE

```
# Sentence pair (1)

the united republic of tanzania

NULL ({ }) jamhuri ({ }) ya ({ 2 5 }) muungano ({ 4 }) wa ({ 1 })
tanzania
 ({ 3 })

# Sentence pair (2)

NULL ({ })
 ({ 1 })

# Sentence pair (3)

the constitution of the united republic of tanzania of 1977

NULL ({ }) katiba ({ }) ya ({ }) jamhuri ({ }) ya ({ 8 }) muungano ({
3 7 9 }) wa ({ 2 }) tanzania ({ 1 4 6 }) ya ({ 5 }) mwaka ({ }) 1977
 ({ 10 })

# Sentence pair (4)

NULL ({ })
 ({ 1 })

# Sentence pair (5)

preamble

NULL ({ })
 utangulizi
 ({ 1 })

# Sentence pair (6)



NULL ({ })
 ({ 1 })

# Sentence pair (7)

chapter one

NULL ({ }) sura ({ 2 }) ya ({ 1 }) kwanza
 ({ })
```

```
# Sentence pair (8)

the united republic, political parties

NULL ({ 3 4 }) jamhuri ({ }) ya ({ 2 }) muungano, ({ }) vyama ({ })
vya ({ 1 }) siasa
 ({ 5 })

# Sentence pair (9)

the people and the policy of socialism and self reliance

NULL ({ 2 9 }) watu ({ 1 4 }) na ({ 6 }) siasa ({ 3 8 }) ya ({ })
ujamaa ({ 10 }) na ({ }) kujitegemea
 ({ 5 7 })

# Sentence pair (10)

NULL ({ })
 ({ 1 })

# Sentence pair (11)

part i

NULL ({ }) sehemu ({ }) ya ({ 1 }) kwanza
 ({ 2 })

# Sentence pair (12)



NULL ({ })
 ({ 1 })

# Sentence pair (13)

the united republic and the people

NULL ({ 3 }) jamhuri ({ 4 }) ya ({ }) muungano ({ }) na ({ 1 2 5 })
watu
 ({ 6 })

# Sentence pair (14)



NULL ({ })
 ({ 1 })
```

```
# Sentence pair (15)

proclamation of the united republic

NULL ({ 5 }) kutangaza ({ }) jamhuri ({ }) ya ({ 4 }) muungano ({ 1 2
})
 ({ 3 })

# Sentence pair (16)

the territory of the united republic

NULL ({ 6 }) eneo ({ 2 }) la ({ 1 4 }) jamhuri ({ }) ya ({ 5 })
muungano
 ({ 3 })

# Sentence pair (17)

declaration of multi-party state

NULL ({ }) tangazo ({ }) la ({ 2 }) nchi ({ 3 }) yenye ({ }) mfumo ({
}) wa ({ }) vyama ({ 1 }) vingi
 ({ 4 })

# Sentence pair (18)

exercise of state authority of the united republic

NULL ({ 4 }) utekelezaji ({ }) wa ({ 8 }) shughuli ({ 2 5 }) za ({ 3 6
}) mamlaka ({ 7 }) ya ({ 1 }) nchi
 ({ })

# Sentence pair (19)

part ii

NULL ({ }) sehemu ({ }) ya ({ 1 }) pili
 ({ 2 })

# Sentence pair (20)

NULL ({ })
 ({ 1 })

# Sentence pair (21)

fundamental objectives and directive

NULL ({ }) malengo ({ }) muhimu ({ }) na ({ }) misingi ({ 3 }) ya ({
}) mwelekeo
```

```
 ({ 1 2 4 })

# Sentence pair (22)

NULL ({ })
 ({ 1 })

# Sentence pair (23)

principles of state policy

NULL ({ 1 }) wa ({ }) shughuli ({ }) za ({ 3 }) serikali
 ({ 2 4 })

# Sentence pair (24)

NULL ({ })
 ({ 1 })

# Sentence pair (25)

interpretation.

NULL ({ }) ufafanuzi.
 ({ 1 })

# Sentence pair (26)

application of the provisions of part ii.

NULL ({ 1 4 })
 ({ 2 3 5 6 7 })
```

APPENDIX THREE: PHRASE TABLE SAMPLE FOR EM MODEL OUTPUT

```
, kupata elimu , na nyinginezo ||| , to educational and other pursuits
. ||| 1 0.00138357 1 0.00401681 ||| 0-0 2-2 4-3 5-4 5-5 1-6 ||| 1 1 1
||| |||
, na nyinginezo ||| and other pursuits ||| 0.5 0.0373564 1 0.209677
||| 1-0 2-1 2-2 ||| 2 1 1 ||| |||
, na ||| and ||| 0.037037 0.0373564 1 0.83871 ||| 1-0 ||| 27 1 1 |||
|||
, uhuru ||| , freedoms ||| 1 0.833333 1 0.185185 ||| 0-0 1-1 ||| 1 1 1
||| |||
, vyama vya siasa ||| , political parties ||| 1 0.104167 1 0.144676
||| 0-0 1-1 2-1 1-2 3-2 ||| 1 1 1 ||| |||
, wilaya za uchaguzi na uchaguziwa wabunge ||| , constituencies and
election of members ||| 1 0.000216182 1 0.0403275 ||| 0-0 1-1 2-1 3-1
4-2 5-3 6-5 ||| 1 1 1 ||| |||
, wilaya za uchaguzi na uchaguziwa ||| , constituencies and election
of ||| 1 0.000216182 0.5 0.0879873 ||| 0-0 1-1 2-1 3-1 4-2 5-3 ||| 1 2
1 ||| |||
, wilaya za uchaguzi na uchaguziwa ||| , constituencies and election
||| 1 0.000216182 0.5 0.318954 ||| 0-0 1-1 2-1 3-1 4-2 5-3 ||| 1 2 1
||| |||
, wilaya za uchaguzi na ||| , constituencies and ||| 1 0.00172946 1
0.318954 ||| 0-0 1-1 2-1 3-1 4-2 ||| 1 1 1 ||| |||
, wilaya za uchaguzi ||| , constituencies ||| 1 0.00192901 1 0.380291
||| 0-0 1-1 2-1 3-1 ||| 1 1 1 ||| |||
, ||| , to ||| 1 0.833333 0.166667 0.0287357 ||| 0-0 ||| 1 6 1 ||| |||
, ||| , ||| 1 0.833333 0.833333 0.833333 ||| 0-0 ||| 5 6 5 ||| |||
. ||| . ||| 0.84 0.488889 1 0.785714 ||| 0-0 ||| 25 21 21 ||| |||
baraza la mawa ziri ||| the cabinet . ||| 1 4.32129e-05 1 0.467532 |||
2-0 0-1 1-1 3-2 ||| 1 1 1 ||| |||
baraza la mawa ||| the cabinet ||| 0.333333 0.00194458 1 0.467532 |||
2-0 0-1 1-1 ||| 3 1 1 ||| |||
baraza la mawaziri na serikali ||| cabinet and the government prime
minister ||| 1 0.0323325 0.25 1.32656e-05 ||| 0-0 1-0 2-0 3-1 4-3 |||
1 4 1 ||| |||
baraza la mawaziri na serikali ||| cabinet and the government prime
||| 1 0.0323325 0.25 0.000384701 ||| 0-0 1-0 2-0 3-1 4-3 ||| 1 4 1 |||
|||
baraza la mawaziri na serikali ||| cabinet and the government ||| 1
0.0323325 0.25 0.0111563 ||| 0-0 1-0 2-0 3-1 4-3 ||| 1 4 1 ||| |||
baraza la mawaziri na serikali ||| the cabinet and the government |||
1 0.00963628 0.25 0.0126876 ||| 0-0 0-1 1-1 2-1 3-2 4-3 4-4 ||| 1 4 1
||| |||
baraza la mawaziri na ||| cabinet and the ||| 1 0.0323325 0.333333
0.0316095 ||| 0-0 1-0 2-0 3-1 ||| 1 3 1 ||| |||
baraza la mawaziri na ||| cabinet and ||| 1 0.0323325 0.333333
0.305559 ||| 0-0 1-0 2-0 3-1 ||| 1 3 1 ||| |||
baraza la mawaziri na ||| the cabinet and ||| 1 0.0174738 0.333333
0.0873025 ||| 0-0 0-1 1-1 2-1 3-2 ||| 1 3 1 ||| |||
baraza la mawaziri ||| cabinet ||| 0.5 0.0360631 0.333333 0.36432 |||
0-0 1-0 2-0 ||| 2 3 1 ||| |||
baraza la mawaziri ||| the cabinet ||| 0.666667 0.01949 0.666667
0.104091 ||| 0-0 0-1 1-1 2-1 ||| 3 3 2 ||| |||
baraza la wawakilishi . ||| of representatives . ||| 0.5 0.0050926 0.5
0.118227 ||| 1-1 2-1 3-2 ||| 2 2 1 ||| |||
baraza la wawakilishi . ||| representatives . ||| 0.5 0.0050926 0.5
0.428571 ||| 1-0 2-0 3-1 ||| 2 2 1 ||| |||
baraza la wawakilishi ||| of representatives ||| 0.5 0.0104167 0.5
0.15047 ||| 1-1 2-1 ||| 2 2 1 ||| |||
baraza la wawakilishi ||| representatives ||| 0.5 0.0104167 0.5
0.545455 ||| 1-0 2-0 ||| 2 2 1 ||| |||
baraza la ||| cabinet ||| 0.5 0.132231 1 0.467532 ||| 0-0 1-0 ||| 2 1
1 ||| |||
binadamu ||| human beings . ||| 1 0.674074 1 0.037037 ||| 0-0 0-1 0-2
||| 1 1 1 ||| |||
bunge . ||| parliament . ||| 0.333333 0.150427 1 0.52381 ||| 0-0 1-1
||| 3 1 1 ||| |||
bunge la jamhuri ya muungano ||| the legislature of the united
republic ||| 1 0.021603 1 0.00124285 ||| 1-0 0-1 1-1 3-2 2-3 3-3 2-4
4-5 ||| 1 1 1 ||| |||
bunge la jamhuri ya ||| the legislature of the united ||| 1 0.0234033
1 0.00134642 ||| 1-0 0-1 1-1 3-2 2-3 3-3 2-4 ||| 1 1 1 ||| |||
bunge la ||| the legislature ||| 1 0.132353 1 0.023416 ||| 1-0 0-1 1-1
||| 1 1 1 ||| |||
bunge laweza kumshtaki rais ||| impeachment by the national assembly
||| 1 0.0220588 1 0.000478928 ||| 0-0 3-2 1-3 2-3 2-4 ||| 1 1 1 |||
|||
bunge ||| impeachment by ||| 1 1 0.166667 0.00574713 ||| 0-0 ||| 1 6 1
||| |||
bunge ||| impeachment ||| 1 1 0.166667 0.166667 ||| 0-0 ||| 1 6 1 |||
|||
bunge ||| parliament ||| 0.666667 0.307692 0.666667 0.666667 ||| 0-0
||| 6 6 4 ||| |||
dhidi ya mashtaka na madai ||| from criminal and civil proceedings .
||| 1 0.0251809 1 0.0310633 ||| 0-0 2-1 3-2 4-3 4-4 4-5 ||| 1 1 1 |||
|||
dhidi ya mashtaka na ||| from criminal and ||| 1 0.0373563 1 0.83871
||| 0-0 2-1 3-2 ||| 1 1 1 ||| |||
dhidi ya mashtaka ||| from criminal ||| 1 0.0416667 1 1 ||| 0-0 2-1
||| 1 1 1 ||| |||
dhidi ya ||| from ||| 0.5 0.0416667 1 1 ||| 0-0 ||| 2 1 1 ||| |||
dhidi ||| from ||| 0.5 0.333333 1 1 ||| 0-0 ||| 2 1 1 ||| |||
```