Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Framework for Many to One Machine Translation

[1]Benson Kituku, [2]Lawrence Muchemi, [3]Wanjiku Nganga
[1] Department of Computer Science, Dedan Kimathi University of Technology, Nyeri, Kenya
[2, 3] School of Computing and Informatics, University of Nairobi, Nairobi, Kenya

Abstract: ***The frequent domestic and international exchanges have created an opportunity for machine translation to flourish since human translators cannot cater for the translation demand. However, there is slow pace in the development of machine translation tools using the one (source language) to one (target language) framework. The paper proposes a many (source languages) to one (target language) conceptual framework that will ensure faster and efficient development of machine translation tools using the Interlingua rule based machine translation approach. The many to one framework make use of shallow structure: lexical similarity and syntactic similarity and deep structure: a unique universal intermediate representation language. A formal computational grammar will be needed and is exemplified by use of two under resourced languages: Swahili and Kikamba. Evaluation model is proposed at the end.***

Key words: ***Lexical similarity, under resourced language, Interlingua, many to one and framework***

I. INTRODUCTION

Machine translation (MT) is a branch of computational linguistics and is defined as an automatic process by computerized system(s) that convert a piece of text (written or spoken) from one natural language referred to as a source language (SL) to another natural language called the target language (TL) with human intervention or not and with the objective of restoring the meaning of the original text in the translated text [1]. Machine translation has been in existence since 1940[2] and over the years a lot of developments and improvements have been witnessed in the approaches and architectures used to build translation systems [3, 4].There are three main approaches to building a machine translation tools summarized in figure 1 namely: knowledge driven approach also known as Rule based Machine translation (RBMT), Data driven Machine translation (DDMT) approach which is also known as corpus based machine translation and hybrid machine translation approach which combines the advantages of the RBMT and DDMT approaches. The approaches are based on underlying theory, for RBMT uses linguistic theory while DDMT uses data theory. All the approaches uses one languages at SL and another one at TL thus one to one framework.

II. MOTIVATION

The frequent domestic and international exchanges within the global village with over 7000 living languages as created an opportunity for machine translation to flourish since human translation cannot cater for the translation demand. However, for under resourced languages (characterized by little or no information technology available, no substantial presence in the internet or digital text and no commercial interest since existing software have not been adapted for use thus technologically marginalized) the pace of development of MT systems is slow and the lexical and syntactic coverage is very far from what has been achieved for better-resourced languages [5, 6, 7, 8].

On the positive side of the under resourced languages, most of the languages are closely related in that they have high lexical similarity (measure of the degree to which the word sets of two or more languages are similar), syntactic similarity of the shallow features of the languages and high similarity of the other grammar features such as: part of speech tags, Morphology etc.[8,9.10]. Since the basic principle and properties of surfaces/shallow structures are universal, expressive capacities equivalent and more so for closely related languages [10, 11]. Thus the richness of the closely related under resourced languages can be exploited by use of a deep structure such as common abstract intermediated universal language [11] which is independent of the any of the source languages and target language. Hence several related source languages are converted to one intermediated universal language from where translation to target language can be done. This approach would ensure one translator can host several languages which result to the many SL to one TL framework as shown in figure 2implying faster and efficient method of modelling machine translation systems for technologically challenged languages thus improving their gross language product.Hence the motivation of this paper is to develop a conceptual framework for many to one framework.

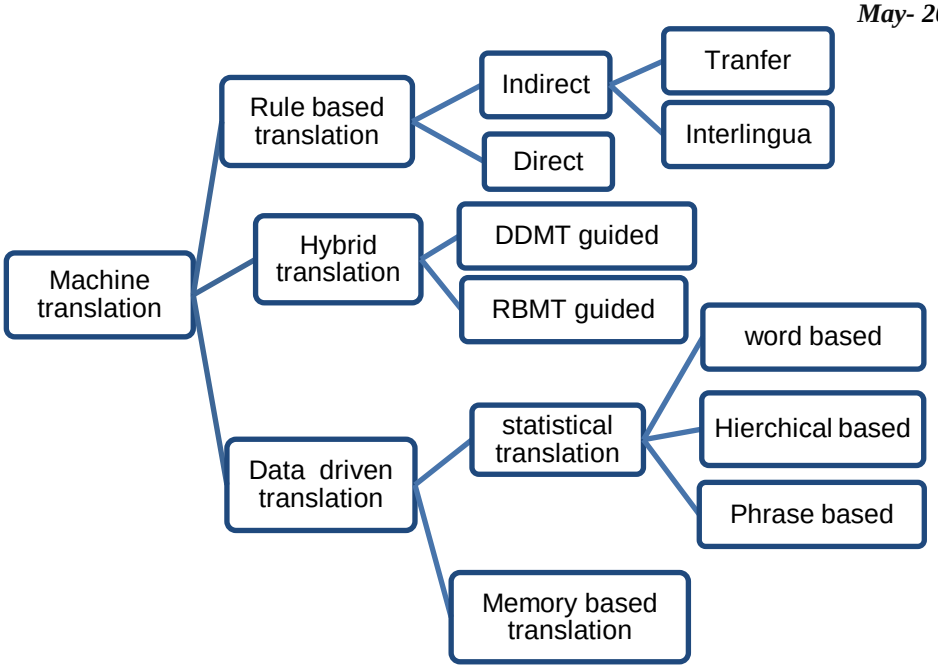

Figure 1 Approaches for machine translation

Figure 2: Many to one framework

III. METHODOLOGY

The methodology involved documents reviews mainly journal and conference papers, university thesis and books on Machine translation and computational grammar plus examination of the various tools or prototypes which has been built using the MT approaches, Triangulation procedure was carried to ensure reliability and viability.

The data collected was inform of documents and interview notes. Establishing categories patterns, features and themes that are outstanding and then Pattern matches them was done at review stage resulting to Qualitative research [12] while categorization was based on the deductive approach [13] Selective and open coding (Coding is the process of organizing the materials into chunks or segments of text before bring meaning to the information [14])was employed in-order to get the patterns from the entire documents on review.

IV. RELATED RESEARCH

Data driven machine translation requires a parallel bilingual corpus and can only work for two languages (TL and SL). These two condition make it impossible to support many to one framework.Thus, the rule based machine translationwhich consist of direct, transfer and Interlingua become the best approach to model the conceptual framework though direct RBMT cannot be used because it is limited to two languages as the case of data driven approach.

The intermediate representation (IR) of SL and TL for transfer is dependent to SL and TL while for Interlingua is independent of the TL and SL thus Interlingua becoming the best option for modelling the framework. In addition , Interlingua approach compared with the other rule based translation method is : the most attractive, better alternative choice, suitable approach for multilingual translation, its performance is better, economical in construction and it has other uses such as question answering, information retrieval and summarization thus making it superior [2,15,16,17]. Therefore, for the purpose of the many to one framework Interlingua rule based approached shall be usedto represent the abstract intermediate universal language.

A. Interlingua

Interlingua is an independent homogenous unambiguous intermediate representation of one or more SLand captures sentence information (words, combination, aspect, mood and sentence style) in a universal way independent of SL and TL and does dependent on generation and analysis of: pragmatic, syntax, semantic and morphology[11,19] of the sentence information.It's the highest level of rule based approaches as illustrate by theVauquois triangle in figure 3.

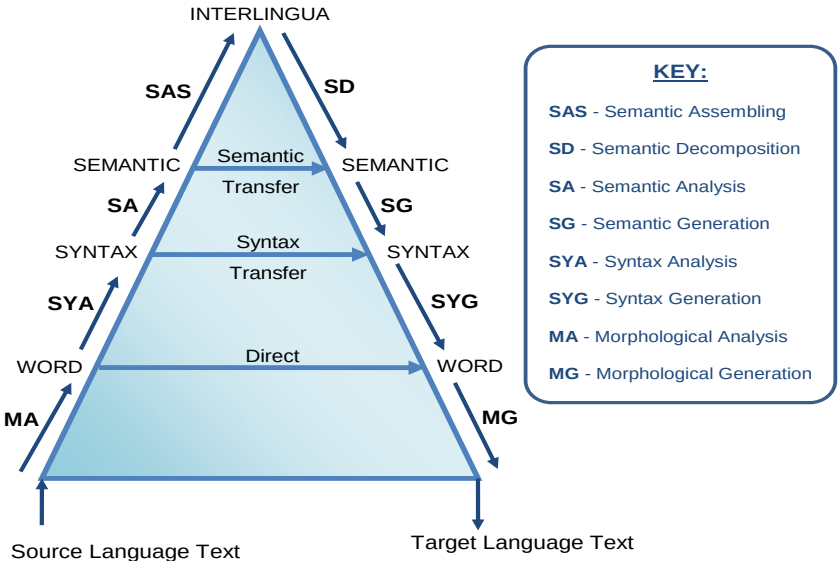

Figure 3 Vauquois triangle [22]

B. Interlingua systems

Interlingua systems have attracted a lot attention despite the cost of the construction of the systems. However, lately data driven system seems to have taken over. Some of the Interlingua systems include: Distributed Language Translation (DLT)[1] which was to develop aMachine translators for 12 Europeans languageswith its IRbeen known as Esperanto and was mainly operated in computer networks. The original version developed was applied between English and French and the performance of the translation was noted to be a bit poor.

Universal translator UNITRANS[2] developed by MIT Artificial Intelligence Laboratory was a bidirectional translator for Spanish, English, and German languages. Its architecture consisted of two parts namely: syntactically which uses linguistic principles and associated parameters while the lexical-semantic part uses lexical conceptual structures [20]. The two stages combine knowledge on languages' specific and language independent [21].

Knowledge-based, Accurate Natural-Language Translation KANT[3] was developed by Carnegie Mellon University for multilingual document production in an industrial setting. KANT is a knowledge based Interlingua system based on coded lexicons and semantics rules. It was experimented using French and English [23, 24]. AnotherInterlingua system is Universal Networking Language (UNL) which was designed for computers to overcome language barriers and therefore itsartificial language that mimic the characteristic of natural languages. It has an engine which converts a natural language into UNLization and vice versa and universal words in the languages are organized using a UNL knowledge bases [11].ATLAS[4] is anInterlinguabidirectional system currently in version 14 and was initially developed to translate from Japanese to English and vice versa butcurrently, can support other languages.

Eurotra[5] was Interlingua system developed by the European community with intention of high quality translation viable for commercial system. The system was developed using prolog programming language andhas three modules for parsing: The first two modules are for syntactic parse that facilitate generation of the intermediate language and third module, backward stage which do the translation. Finally, grammatical framework GF[6] which is a multilingual framework uses parsing and linearization for the translation purposebased on type theory [25]. It has two main modules: abstract syntax which have common features of languages and concrete syntax which have detailed features of languages [26, 27]. The concrete syntax inherits from abstract syntax as the case of inheritance in object programming paradigms. Actually GF is governed by equation 1[26].

$ML_{GF} = < A \{C_i \ldots\ldots C_n)>$……Equation 1.

Whereby

ML is Multi lingual language in GF

A is the abstract syntaxes (common features)

C_i to C_n are the different types of concrete syntaxes

C. Lexical similarity.

Languages have lexical similarity if words set are similar or related in terms of structure for the languages. Let assume there arethree languages L_1, L_2 and L_3then the Venn diagram in figure 4 models lexical similarity in terms of sets

[1] http://www-rohan.sdsu.edu/~ling354/MT-eg.html#DLT-_Distributed_Language_Translation.
[2] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.125.2454&rep=rep1&type=pdf
[3] http://repository.cmu.edu/cgi/viewcontent.cgi?article=1337&context=compsci
[4] https://www.fujitsu.com/global/products/software/packaged-software/translation/
[5] http://en.wikipedia.org/wiki/Eurotra
[6] *www.**grammaticalframework**.org/*

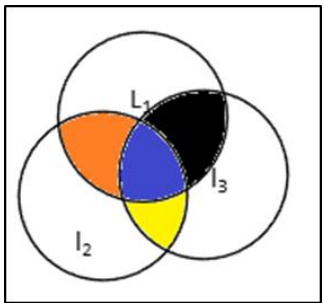

Figure 4 languages Venn diagram

The lexical similarity of language L_1 and L_2 can be model by the intersection $L_1 \cap L_2$ represented by color blue and orange while $L_1 \oplus L_2$ represent the symmetrical lexical difference , likewise the lexical similarity of the three languages can be represented by the intersection of the three languages $L_1 \cap L_2 \cap L_3$ represented by color blue. By extrapolation addition of extra languages up to language L_n, then the lexical similarity will be represented by $L_1 \cap L_2 \cap L_3 \cap L_4 \ldots\ldots\ldots\ldots \cap L_N$. Thepercentages of lexical similarities decrease as the number of languages increase hence indirect proportional. The lexical similarity portion can be used to model a common computational grammar while bootstrapping the dissimilar part of the computational grammar.

For practical purpose let take a case of Bantu languages in Kenya (example of under resourced languages) which account for 45% of the 39 million Kenya populations [28].The Bantu languagescan be divided into three groups based on their geographical location. The eastern Bantu languages which consist of: Kamba, Gikuyu, Ameru and Aembu languages mostly found in eastern and central provinces of Kenya. Western Bantu languages includes: Luhya, Agusii and kuria languages found in Nyanza and western provinces of Kenya. Finally, the coastal Bantu languages include Mijikenda which has nine dialects and Swahili mostly found in coastal province of Kenya [29,30,31].According to ethnologue report for Kenya languages [7][9] most of the Bantu languages have high degree lexical similarity. Figure 5 below shows analysis of lexical similarity of the eastern Bantu languages. From figure 5 the eastern Bantu languages have a similarity of at least 57%.

V. DISCUSSION

Interlingua systems need a composition of morphological, syntax and semantics generators and analyzers in addition to pragmatics of the sentences or phrase. These concepts are used to create the computation grammar (Stastitical or rule based modelling of natural languages grammar for computational perspective) based on formal language theory of Chomsky [32]. Formal grammar uses the mathematical format which consist of four parts namely: Finite set of vocabularies symbol (terminals) $\sum$, Extra symbols called Non terminals NT and unique non terminal called start symbols S and finally set of rules R. Terminal are the actual words in the language, Non terminal are symbols which can be expanded, start symbol is a non-terminal where the sentence or clause or phrase begin and final R define rules of combing the other three parts. Morphology and syntactic structures rules are the set of rules for natural language.

$$G = (\Sigma, NT, T, R) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots equation\ 2.$$

A. Computational grammar

The grammar will be expressed using Wang and Berwick [33] mathematical model for natural languages which capture all the necessary concepts for Interlingua as stated by the equation 3 and taking into consideration the concepts of equation 2.

$$\mathrm{G} = (\Sigma, \mathrm{W}, \mathrm{Lr}, \mathrm{Sy}, \mathrm{Sm}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots equation3$$

Where G is grammar,
$\Sigma$is alphabet,
W are words,
Lr is lexical relation
Sy is syntax
And Sm is semantics

Alphabets Σ for most of languages consist of upper or lower case of [A—Z] with addition or subtraction of few letter, numbers are based on the decimal system [0—9] and special characters for example punctuation marks. The above various forms of alphabets can be combined using operators of regular expression (union, concatenation and star closure) to strings of words.

Lexical relation include among others: part of speech tagging, Morphology, phrases etc. Morphology involves building words from the smaller meaning bearing units called morphemes (stem and affixes). Stem provide the primary and key meaning of the word and it's the main morpheme, on the other hand, affixes add extra meaning when combined with the stem morpheme. Affixes are further classified into four classes: prefixes suffixes, infixes and circumfixes and can capture tense, action, time, mood etc. There are four methods of generating word form: Inflection, compounding, derivation and cliticization [22].

[7] http://archive.ethnologue.com/16/show_country.asp?name=KE

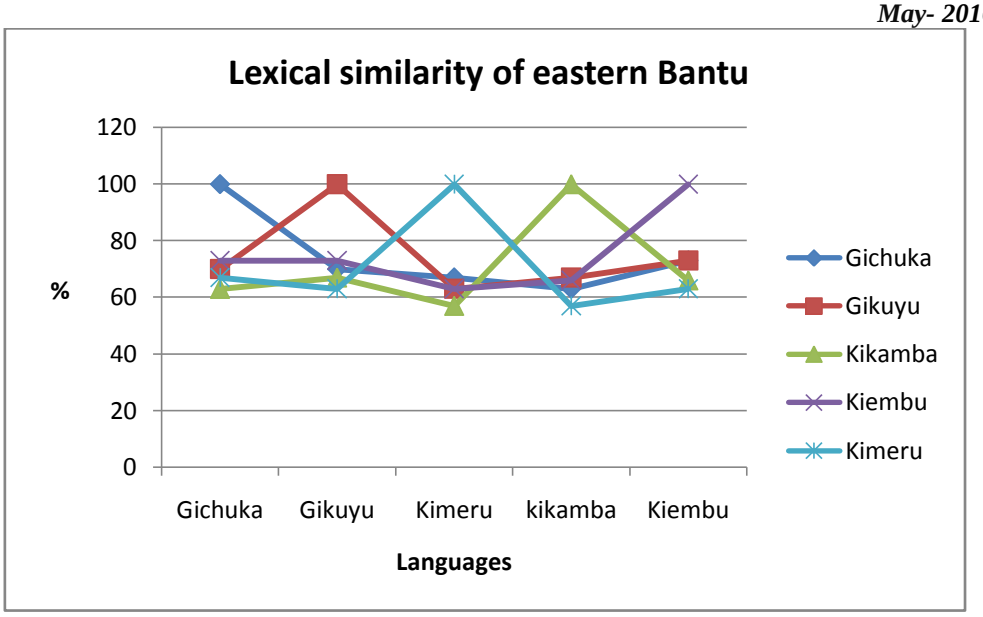

Figure 5 lexical similarity: Source[9]

Part of Speech tagging (PoS) is the process in which syntactic categories are assigned to words or mapping from sentences to strings of tags [19]. They categories include nouns, verbs, conjunction, pronouns etc. in case of English language. Finally, Phrases have to deal with combination of several PoS, for example, noun phrase which is a combination of a noun and adjective while verb phrase is combination of verb and adverb etc.

Syntax deals with sentence formation and structures relation, basically combining words using certain orders (relation) using the rules for putting the words or phrases in the appropriate place in the sentence structure [10,22]. It's a relationship between grammar rules and lexical relation. For example in case of English language the sentences are usually of the form: subject verb object (SVO). Syntactic categories using context free grammar rules are used to model the computation grammar in terms of syntax. Generally computational grammar uses Backus Naur form (BNF) format
Semantics deals with meaning of the sentences or words while pragmatics deals with the knowledge about the relationship of semantics and the intention of the speaker or writer [10,22]. The semantics and pragmatic involves capturing subject of the sentence, behavior or action, the object of the sentence, time when the action is occurring and the space where the action is occurring [10]. The above discussion is summarized in the table below.

Table 1 Formal language frameworks: source [33]

Framework	
Words/Alphabet	Strings of Numbers /letters etc.
Lexical relation	Part of speech tag
	Morphology
Syntactic relation	Phrases making
	Sentence making
Semantic relation	Subject
	Action/behavior
	Object
	Time & space

B. Multilingual source language computational grammar

Let assume language L_1, L_2, L_3……$L_N$have lexical similarity of percentageX and all the languages can form one language L_M then:-

Definition 1: The alphabet Σ of language $L_M$shall be the union of allthe alphabetsΣ of all the languages thatform the languageL_M

$$\Sigma \in L_M = \Sigma \text{ of} L_1 \cup \Sigma \text{ of} L_2 \cup \Sigma \text{ of } L_3 \cup \ldots\ldots\ldots\ldots \Sigma \text{ of } L_N$$

The concatenation of the alphabets symbols results to words that act as the terminals of a particular language. Performing operation on the alphabets symbols of a particular language will result to strings or words (W) and the master language will consist of union of all words/terminals in the languages.

$$W \in L_M = W \text{ of} L_1 \cup W \text{ of} L_2 \cup W \text{ of } L_3 \cup \ldots\ldots W \text{ of } L_N$$

Definition 2: The lexical relation which includes PoS tags,Morphology etc. shall be the union of all the languages forming language L_M. PoS tags are categories such as verb, preposition, pronoun, and quantifiers, demonstrative, personal and possessive pronouns etc. Morphology is based on inflection of the root morpheme in order to form other words e.g plural or singular words, different tenses etc.

$$Lr \in L_M = Lr \in L_1 \cup Lr \in L_2 \cup Lr \in L_3 \cup \ldots\ldots Lr \in L_N$$

Let take case of closely related Kenyan Bantu languages namely Swahili and kikamba languages in particular examine the morphology of verbs and Nouns. Bantu Languages inflection mostly happens at the prefix point. For noun the prefix determine the noun class (gender) and number (plural or singular). Let take example of Noun tree the gender is Mu-Mi and M-Mi for kikamba and Swahili respectively which translate to Mu-ti(singular) ,Mi-ti(plural) and M-ti(singular) and Mi-it(plular) respectively. The lexical relation of the multilingual language would be like

Lr for tree Gender=(Mu-Mi|M-Mi) and Num=singular $\in$ *Muti | Mti*

Lr for trees Gender=(Mu-Mi|M-Mi) and Num=plular $\in$ *Miti | Miti*

The case of verbs, the gender is tense which consist of present tense, past tense and future and can also be illustrated by the verb "fall" which is anguka and valuka in Swahili and kikamba respectively. Note in Bantu languages verbs alone are allowed to form sentences.

Verb past tense= u-li- anguka | nuwa-valuk-ie

Verb present tense= u-na- anguka | niwa-valuka

Verb future tense =u-ta-anguka| u –ka-valuka

The first prefix in Swahili language represent the subject agreement prefix while the second prefix represent tense as well as in Kikambawith exception in past tense of kikamba in addition to the above order the last vowel of the verb changes to "ie"

Defination 3: Syntax relation Sy is modeled based on specific arrangement and agreement of the part of speech tags. The Bantu languages structure of a sentences\phrase follow subject (S), verb (V) and object (O) [34] where basically S represent a noun phrase (NP), V represent VP and O represent NP and can be modelled using Backus Naur form (BNF) format. In Bantu languages the Noun Phrase consist of Noun N, Demonstrative pronoun D, possessive pronoun P, quantifiers Q and adjective A or Personal pronoun PP following in that order. The noun or pronoun must be present what follows after it is optional. The verb phrase consist of a verb followed by optional adverbs and in Bantu languages they have few adverbs. Finally, the object noun phrase behaves the same as subject noun phrase and four Bantu language its optional. Representing the above suing B.N.F for a phrase it would be

PHRASE: = Subject-Noun-Phrase Verb-Phrase Object-noun-phrase;

Subject-Noun-Phrase: = (N|PP) (D|P|Q|A| ε);

Verb-Phrase : = verb (Adverb| ε)

Object-noun-phrase : = Subject-Noun-Phrase | ε

Let consider the sentence "watu wawili weusi waliarifu nyumba" in Swahili and Andu eli aiu nimanaanakie nyumba" in Kikamba, theequivalent in English is "two black people destroyed the house. The subject noun phrase consist of noun (Watu|Andu) followed by quantifier (wawili|eli) then adjective (weusi|aiu). Verb phrase consist of a verb only (waliarifu| nimanaanakie) and finaly the object noun phrase phrase consist of noun only (Nyumba). Therefore the syntax for the two Bantu languages is highly similar.

Defination4: Semantics will be the concatenation of definition two and three. An abstract tree is used to represent the semantic at the highest level of Vauquois triangle. The abstract tree should be complete, coherence and consistence. Complete means all need ingredient for the phrases to make sense are present, while coherence means there are in the right order and consistence tries to avoid conflict

$Sm \in L_M = Sy \in L_M \cap Lr \in L_M$

Defination5: The source language (SL) grammar shall consist of combination of the various components of specific related languages as discussed in definition 1 to 4. Thus words formed out of alphabets $\Sigma \in L_M$, lexical relation$Lr \in L_M$, syntacticcategories, $Sy \in L_M$ and finally semantics$Sm \in L_M$. The equation 4 captures the summary.

$L_M(G) = (W \in L_M, Lr \in L_M, Sy \in L_M, Sm \in L_M)$…..equation 4

C. Proposed architecture of the framework

The architecture is based on equation 4.Firstly, there is need of dictionary for words (the lexicon) that determine the domain (subject) of translation. Secondly, the lexical relation which consists among other morphology, orthographies, and Part of speech based on their rules of formation which will constitute the resource grammar. The similar part of the lexical relation based on lexical similarities can be grouped together without repetitions while the dissimilar each case will be handled on its own.

Syntactic categories and phrases rules to be formed require extraction of the generalized context free grammar (GCFG) based on Chomsky hierarchy. It is worth again to note based on the similarity first, group similar GCFG without repeating and then handle the dissimilar part of GCFGrules. The relation of syntactic relation and lexical relation plus the grammar features create themeaning (semantic relation) which forms the abstract syntax.

The interaction of the all components result to the universal intermediate language as depicted in figure 6. The framework can thenbe used for many to one translation model.

D. Proposed evaluation

Papineni [35] indicate the quality of machine translation is measure of how the output is close to similar human professional translation.we propose use of human evaluators to judge the translation output. Human evaluation involves calling people who have command on the language(s) to judge the output on various dimensions. Jurafsky and Martin [22] identify the dimensions such as fluency which includes intelligent, clear, natural and readable translation output. Fidelity involving how the translation output is adequate and informative and finally, post editing cost which has the idea of how many word need to be added or remove for the translation to make sense.

Each dimension established is scaled in a questionnaire, then each sentence translate is subjected to the questionnaire, the procedure is repeated for several translation output. This been categorical data [14], it's then subjected to quantitative analysis methods in order to establish the confidence level

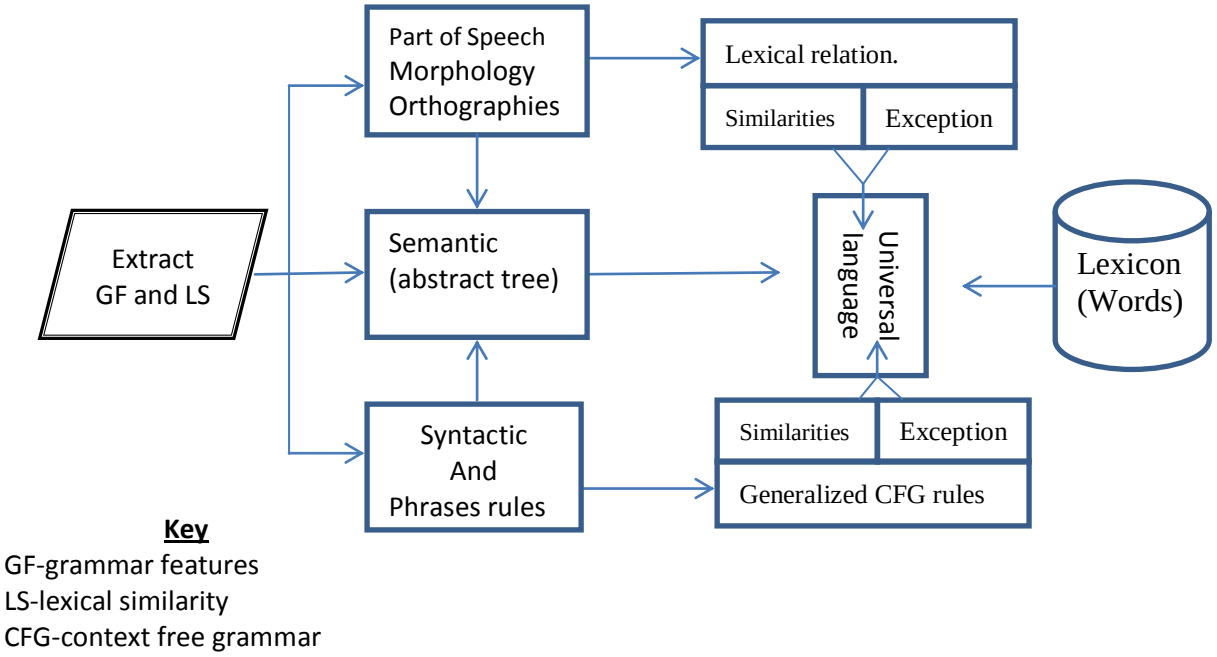

Figure 6 proposed framework architecture

VI. CONCLUSION AND FUTURE WORK

The paper presents a framework for modeling many to one machine translators by exploiting their (languages) shallow structures (lexical similarity, syntactic similarity and grammar features) and deep structures such as intermediate languages. Though under resourced languages have continued to suffer from lack of natural language processing, this work demonstrate that several related languages can be grouped together to form one big language of union of the other languages. Any natural language processing tool developed for the master language can be used by its subset languages. The work will be of significance to the under resourced languages (African languages) users, developers and researchers. Recommend is made for implementation and evaluation of the framework as a future work

REFERENCE

[1] Kituku, Benson, Lawrence Muchemi, and Wanjiku Nganga. "Machine Translation Approaches: A Review." *TELKOMNIKA Indonesian Journal of Electrical Engineering* 17.1 (2016).

[2] Chéragui, Mohamed Amine. "Theoretical Overview of Machine Translation." *Proceedings ICWIT* (2012): 160.

[3] Gupta S. *A survey of Data Driven Machine Translation*. Doctoral dissertation, Indian Institute of Technology, 45] Bombay, 2010

[4] Tripath s ,Sarkhel. K. Approaches to machine translations. *Annals of Library and information studies.*2010; 57: 388-393

[5] Babych, Bogdan, Anthony Hartley, and Serge Sharoff. "Translating from under-resourced languages: comparing direct transfer against pivot translation." *Proceedings of MT Summit XI, Copenhagen, Denmark* (2007).

[6] Berment V. "Méthodes pour informatiser des langues et des groupes de langues peu dotées" PhD Thesis, J. Fourier University – Grenoble I, May 2004.

[7] De Pauw, Guy, Gilles-Maurice De Schryver, and Peter W. Wagacha. "Data-driven part-of-speech tagging of Kiswahili." *Text, speech and dialogue*. Springer Berlin Heidelberg, 2006.

[8] Muhirwe, Jackson. "Towards human language technologies for under-resourced languages." *Strengthening the Role of ICT in Development* (2007): 38.

[9] Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig. *Ethnologue: Languages of the world*. Vol. 16. Dallas, TX: SIL international, 2009.

[10] Wang, Yingxu. "A formal syntax of natural languages and the deductive grammar." *Fundamenta Informaticae* 90.4 (2009): 353-368.

[11] AlAnsary S. *Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas*. In 11th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt. 2011:

[12] Myers M D. Qualitative research in information systems. *Management Information Systems Quarterly*. 1997; *21*: 241-242.

[13] Strauss A, Corbin J. *Basics of Qualitative Research Techniques*, London.Sage Publications: 1998

[14] Creswell, John W. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2013.

[15] Peng L. A Survey of Machine Translation Methods. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(12): 7125-7130
[16] Hutchins W.J, Somers H L. An introduction to machine translation.*London: Academic Press*.1992:
[17] Hutchins John. A new era in machine translation research. In *Aslib proceedings*. 1995; 47(10) 211-219
[18] Hiroshi U,Meiying Z. *Interlingua for multilingual machine translation*. Proceedings of MT Summit IV, Kobe, Japan. 1993:157-169.
[19] Daelemans, W. Zavrel, J.van der Sloot, V and van den Bosch, (2002). "MBT: Memory- Based Tagger, version 1.0, Reference Guide," ILK Technical Report ILK-0209, University of Tilburg, The Netherlands. 2002
[20] Dorr, Bonnie J. "A cross-linguistic approach to translation." *Austin, TX: Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*. 1990.
[21] Hutchins, John. "Latest developments in machine translation technology: beginning a new era in MT research." *MT Summit IV, Kobe Japan* (1993).
[22] Martin, James H., and Daniel Jurafsky. "Speech and language processing." *International Edition* (2000).
[23] Nyberg, Eric, Teruko Mitamura, and Jaime G. Carbonell. "The KANT Machine Translation System: From R&D to Initial Deployment." (1997).
[24] Nyberg, Eric H., and Teruko Mitamura. "The KANT system: Fast, accurate, high-quality translation in practical domains." *Proceedings of the 14th conference on Computational linguistics-Volume 3*. Association for Computational Linguistics, 1992.
[25] Ranta, Aarne. "Grammatical framework." *Journal of Functional Programming* 14.02 (2004): 145-189.
[26] Ranta, Aarne, Ali El Dada, and Janna Khegai. "The GF resource grammar library." *Linguistic Issues in Language Technology* 2.2 (2009): 1-63.
[27] Ranta, Aarne. *Grammatical framework: Programming with multilingual grammars*. CSLI Publications, Center for the Study of Language and Information, 2011.
[28] Oparanya.,A 2009 Population & Housing Census Results. Available from [http://www.knbs.or.ke/Census Results/Presentation by Minister for planning revised.pdf], Nairobi, Kenya,2010
[29] Hinnebusch, Thomas J. "Prefixes, sound change, and subgrouping in the coastal Kenyan Bantu languages." (1973).
[30] Wagner, Gunter. "The Bantu of Western Kenya, Vol. 1." *London: Oxford University* (1970).
[31] McIntosh, B. G. The Eastern Bantu Peoples. *Zamani: A Survey of East African History*, 198-215, 1968.
[32] Jäger, Gerhard, and James Rogers. "Formal language theory: refining the Chomsky hierarchy." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367.1598 (2012): 1956-1970.
[33] Wang, Yingxu, and Robert C. Berwick. "Towards a formal framework of cognitive linguistics." *Journal of Advanced Mathematics and Applications* 1.2 (2012): 250-263.
[34] Kioko, Angelina. *Theoretical issues in the grammar of Kikamba: a Bantu language*. Lincom GmbH, 2005.
[35] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.

