A System for Machine Translation from English to Ebira using the Rule-Based Approach.

ABSTRACT

This research work is aimed at bridging the knowledge gap between the most popular knowledge rich English language and the minority Ebira language spoken by the Ebira people, a minority ethnic group in part of Nigeria. Across the globe and on the internet, English language has become the most widely used language for knowledge dissemination. And presently, the majority of the indigenous people of Ebiraland also known as "Anebira" are still not proficient in their use of English language which as a result prevents them from gaining full knowledge disseminated in English language. Hence, the need to develop an automated Machine Translation System capable of translating English text to Ebira text which will help the people to tap from the abundant knowledge conveyed in English language for effective and fast development in their social, political, scientific, philosophical and economic areas of life. The system was designed to consolidate on human translators' effort and not to replace them. A comprehensive study and analysis of the two languages was carried out with the help of Ebira native speakers in Ebiraland Kogi central and some professional English language tutors at FCE Okene. The knowledge gathered provided the basis for the design and testing of the rule base, inference engine, bilingual dictionary which are important components for the proposed automated system for translation of English text to Ebira text using PHP. Making use of the word in the bilingual dictionary, the system will successfully translate your English text to Ebira. The system was evaluated using one of the popular automatic method of evaluating MT systems BLEU(Bilingual Evaluation Understudy). And an accuracy of 81.5% in translation was achieved. An improved system in the future is recommended to accommodate more complex sentences for the more benefit of the good people of Enebira.

Keyword: *automated, bilingual dictionary, rule base, translators, PHP.*

Introduction

The ever increasing need for cross-regional communication and Information exchange has made translation from one human language to another a matter of absolute necessity in today's highly globalized and connected/networked world. Language is

the medium of communication. Human language is purposively to communicate ideas, emotions, feelings, desires, to co-operate among social groups, to exhibit habits etc. which can be translated along a variety of channels (Bamisaye O, 2000). There are over 6,800 living languages in the world which reflects the scope of linguistic and cultural diversity. Access to information written in another language is of great interest and the means of sharing information across languages is translation, therefore developing technologies for translating from one language to another is very important. Without translation, there can be no cross-regional communication and many voices will not be heard without this critical function.

(Koehn P, 2009) Showed that due to difference in culture and the multilingual environment in India inter-language translation was necessary for the transfer of information and sharing of ideas. The need for translation is also very glaring in the business community. It has been observed that language barriers between companies and their global customers are stifling economic growth and in fact, forty-nine percent of executives say a language barrier has stood in the way of a major international business deal, nearly two-thirds (64 percent) of those same executives said language barriers are making it difficult to gain a foothold in international markets, whether inside or outside your company, your global audiences prefer communicating with you in their native languages. It increases efficiency, receptivity, and allows for easier understanding of concepts (Ayegba F, 2016). Language translation is imperative in the globally united and yet linguistically and culturally separated world in which we live.

Humans were originally responsible for translating from one language to another. At a point the supply of translation services could no long keep pace with the demand for translated content, moreover human translation is costly, time consuming and inadequate for addressing the real-time needs of businesses to serve multilingual prospects, partners and customers (Ayegba F, 2016). The inherent limitations of human translation made the search for an alternative means of translation paramount. The search led to the discovery of what is known today as machine translation or computer assisted translation. Machine Translation (MT) is defined as the use of computers to translate messages in the form of text or speech from one natural language (human language) into another language of nature (Ahmed & Mohd, 2014).

It is the process of using software to translate text from one natural language to another. This need has prompted research organizations and government agencies to develop tools for automatic translation of text in attempt to achieve wider outreach and bridge the gap of language diversity (Koehn P, 2010).

MT has proved to be of social, political, scientific and philosophical importance. Social and political importance emerges from the necessity to understand the other. Binational or multinational countries and organizations need to translate great volumes of texts into many languages in a very limited time. For instance, European Union allocates around €330m a year to translate from and into 23 official languages. In addition, Union allocates nearly %1 of the annual budget for all the language services (DG Translation official website, 2014). European Union uses an internal machine translation engine, which has shifted from rule - based to statistical MT system in the recent years. Commercial importance emerges from the fact that for each step in international markets, from business agreements to instruction manuals, translation is a requirement for people to interact with each other. The delays in translation can be costly, so using MT can help translators and trading parties in the most efficient ways.

Nigeria is the most populous country in Africa with a population of about 200 million people. It is also the seventh largest country in the world (Ayegba F, 2016). Nigeria is a multilingual country with over 500 ethnic groups. This shows the level of linguistic and cultural diversity in the country. Ebira is one of the ethnic groups in Nigeria.

Ebira is a language spoken by the Ebira people of Nigeria. The language is spoken by over one million natives who live in Nigeria (Adive J, 1989). A figure which is believed to have tripled according to the last census in Nigeria. The word “Ebira” refers to the people themselves, their language and their geographical location. The people call themselves “Anebira”.

The Ebiras who are predominantly of the Niger-Benue confluence area are found in Okene, Okehi, Adavi, Ajaokuta, Lokoja, Koton-Karfe Local Government Areas, Muzzum in the the riverine area of Bassa Local government of Kogi State. Other places where the Ebiras are found are Toto and Ummaisha both in Toto Local Government Area of Nassarawa State. Abaji, Itako and Ukya area of the Federal

Capital Territory and Igara (Ebira-etuno) a town and Local Government headquarters in Akoko Edo Local Government Area of Edo State (Aliyu Y, 1989)

Translation Technology

Translation technology is a general term used to describe the technologies or computerized tools available to translators to help them do their job. Two of the most common technologies are probably Translation Memory (TM) and Machine Translation (MT). It is worth noting that the emergence of translation technology does not mean the superannuation of human translators. Machine translation has not attained and may not attain the interpretive skill of the professional human translator. Machine translation can never be a substitute for human translation because no matter how fast a computer functions, the real problem in translation is not electronic but one of linguistics. It is hard to adequately represent through models the vast array of interconnections between the words of a single language and also the target language and correctly and completely map all the interconnections between them. It is impossible to make correct choice between competing terms in a great majority of cases. Effort to invent a method for fully automatic high quality translation was seen as an exercise in futility and a dream that will not come true in the near future. Although there have been significant improvement in the quality of MT outputs over the years, factors such as infinite diversity of human language, complexity and impreciseness in human language which makes high quality translation difficult for MT has not changed. This means that human translators will remain indispensable in the field of translation. MT should therefore be seen as a translation support tool. Machine translation technology has both direct and indirect benefits (Lambrose & Harris etal, 1994). The direct benefits include: reduced translation cost, improved delivery time, availability, consistency and throughput. The indirect benefits includes: reduced support cost, improved documentation, faster time to market and increased product/brand loyalty.

Machine Translation Approaches

Machine translation approaches can be divided into different categories. Under this classification, two main paradigms can be found: the rule-based approach and the empirical-based or data driven approach. Rule-based translation systems can be divided into three catalogs: literal translation method, interlingua-based method and transfer-based method. Rule-based systems are based on linguistically-informed foundations requiring extensive morphological, syntactic and semantic knowledge. The input is transferred to the target using a large set of sophisticated linguistic translation rules. Translation rules are created manually, demanding significant multilingual and linguistic expertise. Therefore, rule-based systems require large initial investment and maintenance for every language pair (Egbunu, F, 2013). Also within the empirical-based paradigm, two other approaches can be further distinguished: example-based and statistical-based and context based (Ibrahim, S, 2014). Under the empirical-based approach the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. The advantage is that, once the required techniques have been developed for a given language pair, MT systems should – theoretically be quickly developed for new language pairs using provided training data.

Although the rule based system require significant amount of linguistic knowledge, the knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system. (Hieu, H, 2011) posited that rule-based approach is better than its counterpart corpus-based approach for two main reasons: 1: less-resourced languages, for which large corpora, possibly parallel or bilingual, with representative structures and entities are neither available nor easily affordable, and 2: for morphologically rich languages, which even with the availability of corpora suffer from data sparseness. It is clear from this argument that each of these technologies or approaches has their strengths and weaknesses which will be discussed in detail in latter sections.

Choice of technology for building machine translation applications

It is clear from the foregoing discussions on the various technologies or approaches for developing machine translation systems that each approach has its requirements as well as strength and weaknesses. Some languages can be described as resource-poor or low-resource

while others are said to be high-resource depending on the extent of availability of linguistic resources.

Low-Resource Languages (LRLs) can be understood as less studied, resource scarce, less computerized, less privileged, less commonly taught, or low density, among other denominations (Anil S, 2008, Christopher & Mike etal, 2016, Yulia, T, 2017). From this description of low-resource languages and going by the fact that empirical based methods of machine translation requires large amount of comparable or aligned corpora, empirical based methods cannot be directly applied to LRLs because of data scarcity. The unavailability of massive dataset of translated sentences needed to train a model capable of translating between two languages for LRLs makes NMT unsuitable for building MT systems for LRLs. Some believe that the rule-based approach is the most appropriate for LRLs.

A study of literatures clearly shows that Ebira is a low-resource language and therefore rule based approach is the appropriate technology or method for building MT applications for Ebira language for now. As time progresses and more research is carried out for the language, relevant linguistics resource for use of empirical based methods and NMT will be made available.

Analysis of English and Ebira Language

The two language pairs, English and Ebira were carefully studied in terms of syntax, semantics, morphology, part of speech, word order, pluralization of nouns, structural differences etc. through document study, observation and interaction with Ebira/English professionals and community elders with vast experience and knowledge about Ebira history. The rules that govern the combination of words to form correct sentences in Ebira were identified.

Ebira is a fixed word order SVO (subject, verb, object) like English but the arrangement of words in noun phrase and adjective phrase are not the same. English places modifiers before nouns in noun phrases, Ebira does the reverse, nouns are placed before modifiers.

Significant amount of linguistic knowledge is required for successful deployment of machine translation systems using the rule based machine translation technology; therefore, a comprehensive study and analysis of the two languages was carried out. This knowledge

provided the basis for the design of the rule base, the inference engine and the full-form lexicon which are essential components of the proposed rule based system for automatic translation of English sentences to Ebira sentences.

From the analysis, the transformational rules that govern the translation of English and Ebira phrases presented in the tables below were extracted.

Table 1: Noun Phrases Transformational Rules

Rules	English: NP=	Ebira: NP=	Examples	
R1	DA + N	N + DA	**E**	The car
			Eb	moto ono
R2	IDA + N	N	**E**	A book
			Eb	Uwe
R3	DA + ADJ + N	N + ADJ + DA	**E**	The handsome boy
			Eb	onoru ozoza ono
R4	IDA + ADJ + N	N + ADJ	**E**	A big room
			Eb	Iyara obanyi
R5	DA + ADJ + N	N + ADJ + ADJ + DA	**E**	The intelligent person
			Eb	Oza oniroye ono
R6	PPN + N	N + PPN	**E**	My car
			Eb	moto ami
R7	PPN + N	N + PPN	**E**	Your sister
			Eb	Onyeiza awu
R8	PPN + N	N + PPN	**E**	His mother
			Eb	Onyi aani
R9	DEM + N	N + DEM	**E**	That car
			Eb	moto onoo
R10	DEM + N	N + DEM	**E**	This guy
			Eb	Ozi ononi

Table 2: Verbal Phrases Transformational Rules

Rules	English: VP=	Ebira: VP=	Examples	
R1	AV + V	AV + V	**E**	have written
			Eb	ra chere
R2	AV + V	'a' + V	**E**	is cooking
			Eb	Amisa
R3	AV + V	'ya' + V	**E**	ore cooking
			Eb	Yamisa
R4	AV + V + ADV	AV + V	**E**	has broken up
			Eb	ra hahi
R5	AV + V	V	**E**	am walking
			Eb	Zuse
R6	AV + V + PN	AV + 'nyi' + V + PN	**E**	**might love** me
			Eb	ozù nyi oyisì mi
R7	PN + AV + AV + V	AV + 'oni' + PN + V	**E**	**I have been waiting**
			Eb	osuku onima je
R8	AV + AV + V	AV + 'oniyi' + V	**E**	have been waiting

			Eb	osuku oniyi je
R9	PN + AV + AV + V	AV + ‘ono’ + V	**E**	has been waiting
			Eb	osuku ono je
R10	AV + AV + V	AV + ‘oni' + PN + V	**E**	have been teaching
			Eb	osuku onima zozisa

Part of Speech tag system

Words are grouped into categories called parts of speech. There are eight parts of speech in English language. They are Nouns, Verb, Adjectives, Adverb, Conjunctions, Preposition and Determiners. In Ebira there are some parts of speech such as nouns, adverb, adjectives, verbs etc.

The meaning of some words in Ebira depends on the nouns that follow them, for example the word and when associated with a place or location means ati but when it is associated with something animate means oniri for this reason, we had to develop a part of speech tag system for the machine translation differently from the conventional part of system so that meanings can be appropriately conveyed.

Data dictionary

The data dictionary as the name implies is a catalog, a repository of data items. It stores the description of the elements in the database specifications.

Table 3: Data Dictionary

Field Names	**Full length Field Name**	**Description**
Id	Identification Number	A number that uniquely identifies a record. It is automatically generated.
Englishword	English word	A word in English language.
Pos	Part of speech	The part of speech of an English word.
Ebiraequivalent	Ebira equivalent	The Ebira meaning of an English word.
pair1	Pair1	The first English word in an n-gram.
pair2	Pair2	The second English word in an n-gram
pair3	Pair3	The third English word in an n-gram.
pair4	Pair4	The fourth English word in an n-gram.
Ebirameaning	Ebira meaning	The Ebira meaning of an n-gram

Postag	Part of speech tag	The part of speech tag for an n-gram.

System Architecture

The conceptual architecture of the system is shown in figure 1

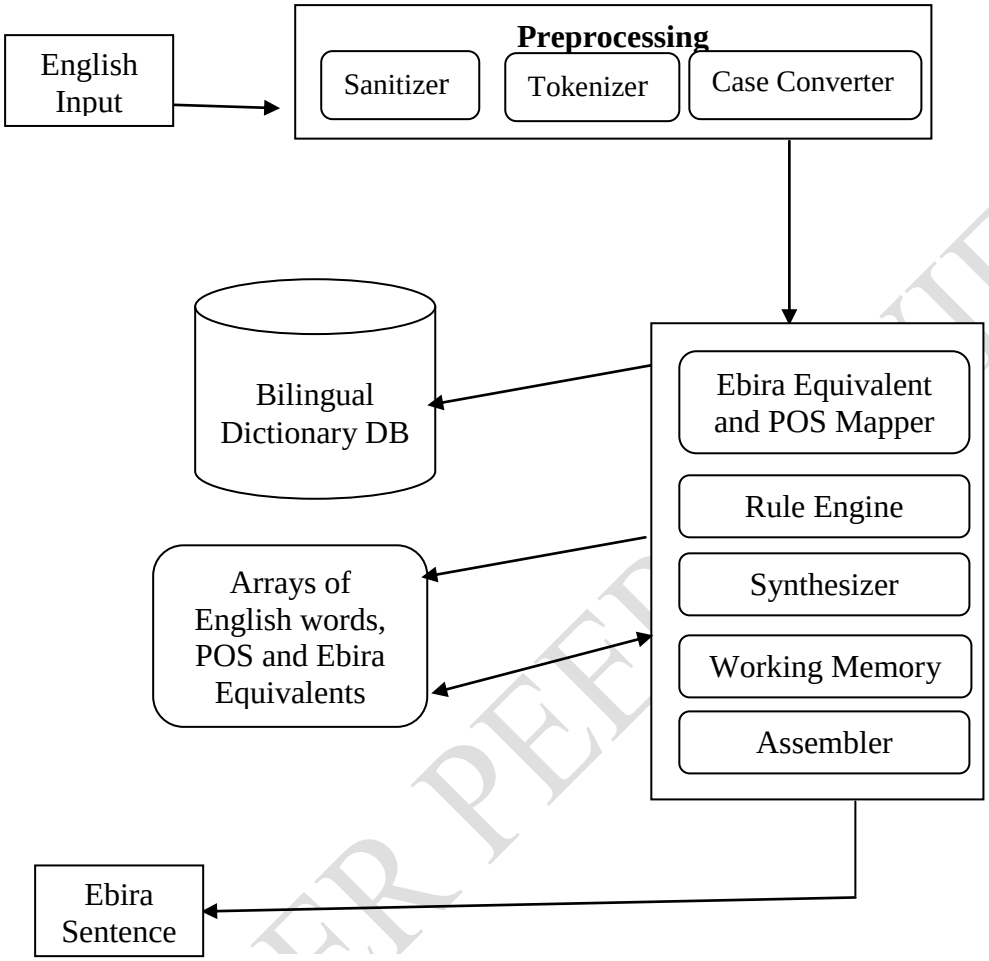

Figure 1: Proposed System Architecture

Figure 2: Sample translation Interface

After entering the text to be translated and then click ***Translate,*** the result is shown above.

Method Adopted in This Evaluation

This study adopts the BLEU method for the evaluation of the output of the English to Ebira Machine Translation System.

The BLEU Method

BLEU which stands for Bilingual Evaluation Understudy is an algorithm for evaluating the quality of text which has been machine translated from one natural language to another. Quality is considered to be the correspondence between a machine output and that of a human. A number of studies (Doddington, G. 2002, Coughlin, D. 2003, Denoual & Lepage, 2005) showed that BLEU highly correlated with human judgment of translation quality with respect to adequacy and fluency. It is the most popular automatic method for evaluation of machine translation output and a benchmark for the assessment of any new evaluation metric. This is the main thrust for its adoption in evaluating the result of this research.

BLEU Implementation

A VB.net application was developed in Visual Studio 2010 to implement BLEU algorithm and used to evaluate the output of rule based system for automatic translation of English sentences to Ebira language.

A corpus of 200 English sentences was created and given to a professional to translate from English to Ebira. The Professional was in the Department of Linguistics in Federal College of Education, Okene in

Kogi State. The translated English to Ebira sentences was used as the reference translation. The reference translation was stored in a table in the database called Evaluation Table which has the following fields: TranslationID, EnglishSentence, ReferenceTranslation, Candidatetranslation and BleuScore. The same set of English sentences were given to the English to Ebira Automatic Translation System, the translated text was stored in the Candidatetranslation field based on the TranslationID. The BLEU Score Computation module was executed to compute the BLEU score for each of the translated sentence

Results and Discussion

Table 4: shows sample BLEU scores for the sentences.

EvaluationTable Query				
TranslationID	**EnglishSentence**	**ReferenceTranslation**	**CandidateTranslation**	**BLEUScore**
341	the man is good	ọmụha ọnọ ọzọza	ọmụha ọnọ ọzọza	1.00
342	Arise and go to Nineveh that great city and cry.	akwẹtẹ atị na kọ Nineveh ọnọọ ọbọọrọ atị ọrẹrẹ	akwẹtẹ atị na kọ Nineveh ọnọọ ọbọọrọ atị ọrẹrẹ	1.00
343	Bribery and corruption is the main problem in Nigeria.	ọchịrẹrẹnụ atị akụvị ọnọ vị ọdayị ẹngwụkata Nagẹrịya	ọchịrẹrẹnụ atị ọkụvị ịrẹyị ọdayị ọnọ Nagẹrịya	0.69
344	The president will deliver Nigeria from corruption and insecurity.	mịva ọhịrẹsụẹpa sụkụtụ Nagẹrịya hụrị ọkụvị atị ayị-nyayịọnẹtẹ	mịva ọhịrẹsụẹpa sụkụtụ Nagẹrịya hụrị ọkụvị atị ayị-nyayịọnẹtẹ	1.00
345	The tall lecturer wrote the book	Lẹctụrẹr ọgọdọ ọnọ ọchẹrẹ ụwẹ ọnọ	lecturer ọgọdọ ọnọ chẹrẹ ụwẹ ọnọ	0.91
346	The trailer crushed the small car and killed the people inside.	Moto ọgọdọ kwụọrọ moto ọwẹyị atị ẹwụ aza ịnịnẹ ọnọ.	mọtọ-ọbanyị ọnọ kwụọrọ ọnọ ọwẹyị moto atị wụ ọnọ ẹza ịnịnẹ	0.97
347	They carried policemen to intimidate the people.	Ẹ shụọ ọjịsasụnẹ na kụhịrẹ ẹza.	ẹnị shụọ ẹnẹ-ọjịsasụnẹ kọ kụhịrẹ ọnọ ẹza	0.91
348	They came to see the deputy governor.	Ẹ vẹ rị atụra ọhịrẹsụ.	ẹnị ọvẹ kọ rẹ ọnọ atụra ohịrẹsụ	0.81
349	The thieves vandalized properties.	Ọyị ọnọ dụwọha ọgụ	ọnọ ọyị-nịnị dụwọha ogụ	0.93
350	They have been stealing from government treasury.	Ẹ ra tọ hụnị ọgwụ ịjọva	ẹnị ra yọyị hụrị ụgwụ ịjọva	0.89

The result of the 200 test sentences was analyzed. Table 4 shows the analysis.

Table 5: Analysis of Results

BLEU Score range	No. of Sentences	Percentage value
>=0.5and <=1.0	163	81.5
<=0.4	37	18.5

According to (Papineni & Roukos etal, 2001) BLEU scores above 0.50 reflect good and fluent translations. From the table total percentage score above 81.5% Therefore accuracy of 81.5% was achieved.

This result shows clearly that the stated objective of the research was achieved.

Conclusion and Future work

This work has been concentrated on the issues in the design and implementation of Rule based machine translation system which translates English sentences to Ebira. A rule based approach that satisfies the following requirement for translation of sentences: fast, correct, easy to edit was proposed. The system uses full form bilingual dictionary which eliminates the need for morphological analysis as many MT systems do; this increases translation quality and reduce processing time. The part of speech is built into the full form bilingual dictionary. This eliminates the need for an external part of speech tagger used in many MT systems. The result of this approach is reduced processing time.

Future Work

This research work can be continued in the following direction:

1. Develop parallel corpus for English and Ebira to facilitate the development and deployment of translation system for both languages using the corpus based approach.
2. Develop the English to Ebira translation system using other technologies such as Statistical machine translation and Example based machine translation and carry out a comparative analysis of various technologies.

3. Create English/Ebira language Forum targeted at providing Ebira equivalent of English words in areas where there is no linguistic equivalence. This will also lead to better translation between the language pair.

References

Adive J. (1989). *The verbal piece in Ebira. Dallas:* Summer Institute of Linguistics.

Ahmed J, Mohd J., AB Aziz Arabic – Malay Machine Translation Using Rule Based Approach. Journal of Computer Science 2014. pp 1062-1068. ISSN: 1549- 3636

Aliyu Y. O.(1989). In Preservation of An Identity: EPA lecture series. Kaduna: Nagazi printing press.

Anil Kumar Singh. 2008. Natural language processing for less privileged languages: Where do we come from? where are we going? In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages.

Ayegba, F (2016) Ruled Based English to Igala Machine Translation. PhD Thesis, Universidad Azteca Mexico

Bamisaye, O. T (2000) *Essentials of English Syntax*,Balfak Educational Publisher, Ado-Ekiti, Nigeria, ISBN 978-2558-14-04.

Coughlin, D. (2003) "Correlating Automated and Human Assessment of Machine Translation Quality" in MT Summit IX, New Orleans, USA pp. 23–27

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 4543– 4549.

Denoual, E. and Lepage, Y. (2005) "BLEU in characters: towards

automatic MT evaluation in languages without word delimiters in Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing pp. 81–86.

Doddington, G. (2002) "Automatic evaluation of machine translation

quality using n-gram cooccurrence statistics" in Proceedings of the Human Language Technology Conference (HLT), San Diego, CA pp. 128–132.

Egbunu, Fidelis Eleojo, *Education and Re-orientation of Igala Cultural Values,* African Journal of Culture, Religious, Educational and Environmental Sustainability (AJCREES), Vol. 1, No. 2. Pp. 66 – 82. Dec., 2013.

Hieu, H, (2011) *Improving Statistical Machine Translation with Linguistic Information*, PhD Thesis, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh.

Koehn Philipp. 2009. A process study of computer aided translation, *Philipp Koehn*, *Machine Translation Journal*, 2009, volume 23, number 4.

Koehn Philipp, (2010) *Statistical Machine Translation*, Cambridge University Press, ISBN-13 978-0-511-69132-4.

Ibrahim, S (2014) Intelligent hybrid Man-Machine Translation evaluation M.Sc Thesis, Graduate School, Alexandria University.

Lambros Cranias , Harris Papageorgiou, Stelios Piperidis.A Matching Technique in Example Based Machine Translation, 1994. Available online at http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=B2003FC4A79507597C0AEA0CA1281FCA?doi=10.1.1.13.5373&rep=rep1&type=pdf, retrieved on 2nd May, 2021.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2001). IBM RESEARCH REPORT, BLEU: a method for automatic evaluation of machine translation IBM Research Division Thomas J. Watson Research Center, Yorktown Heights, NY 10598 RC22176(W0109-022)

Yulia Tsvetkov. 2017. Opportunities and challenges in working with low-resource
languages. Carnegie Mellon University.