Asian Journal of Research in Computer Science

1(4): 1-8, 2018; Article no.AJRCOS.44217

Developing Statistical Machine Translation System for English and Nigerian Languages

Ignatius Ikechukwu Ayogu[1*], Adebayo Olusola Adetunmbi[1] and Bolanle Adefowoke Ojokoh[1]

[1]Department of Computer Science, Federal University of Technology, Akure, Ondo State, Nigeria.

Authors' contributions

This work was carried out in collaboration between all authors. Author IIA designed the study, wrote the protocol, conducted the experiments and wrote the first draft of the manuscript. Authors AOA and BAO validated the design, the protocol and managed the study. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2018/v1i424761
Editor(s):
(1) Dr. Stephen Mugisha Akandwanaho, Department of Information Systems and Technology, University of KwaZulu-Natal, South Africa.
Reviewers:
(1) Xiaodong Shi, Xiamen University, China.
(2) M. Bhanu Sridhar, GVP College of Engineering for Women, India.
(3) Solehah Yaacob, International Islamic University Malaysia, Malaysia.
Complete Peer review History: http://prh.sdiarticle3.com/review-history/26773

Original Research Article

Received 03 July 2018
Accepted 10 October 2018
Published 23 October 2018

ABSTRACT

The global demand for translation and translation tools currently surpasses the capacity of available solutions. Besides, there is no one-solution-fits-all, off-the-shelf solution for all languages. Thus, the need and urgency to increase the scale of research for the development of translation tools and devices continue to grow, especially for languages suffering under the pressure of globalisation. This paper discusses our experiments on translation systems between English and two Nigerian languages: Igbo and Yorùbá. The study is setup to build parallel corpora, train and experiment English-to-Igbo, ($en \rightarrow ig$), English-to-Yorùbá, ($en \rightarrow yo$) and Igbo-to-Yorùbá, ($ig \rightarrow yo$) phrase-based statistical machine translation systems. The systems were trained on parallel corpora that were created for each language pair using text from the religious domain in the course of this research. A BLEU score of 30.04, 29.01 and 18.72 respectively was recorded for the English-to-Igbo, English-to-Yorùbá and Igbo-to-Yorùbá MT systems. An error analysis of the systems' outputs was conducted using a linguistically motivated MT error analysis approach and it showed that errors occurred mostly at the lexical, grammatical and semantic levels. While the study reveals the

**Corresponding author: E-mail: ayoguii@futa.edu.ng;*

potentials of our corpora, it also shows that the size of the corpora is yet an issue that requires further attention. Thus an important target in the immediate future is to increase the quantity and quality of the data.

Keywords: Machine translation; Igbo language; Yoruba language; parallel corpora; SMT.

1. INTRODUCTION

Machine Translation (MT) is an important application of natural, human language processing, the art and science of human language computation. Natural language research develops computational algorithms for the representation, understanding and generation of natural languages by computer systems. The MT process involves the use of computational systems to decode and encode the meanings in a given input sentence from one language, the source language, e.g. Igbo, to the other(s), the target language(s), e.g. Yorùbá, Hausa, English or Chinese.

The process of translation can be described as a two-stage process of decoding the source language input (to obtain the meaning) and encoding the meaning thus obtained into outputs in the target language. Although a number of approaches have been developed to perform these processes, the intuition, principles, requirements and motivations for the use of each approach differ significantly. MT approaches can be rule-based (symbolic), empirical (data-driven) or a blend of rule-based and data-driven techniques in a hybrid system [1]. Each of the approaches is useful in its own right and capable of achieving good results but suitability varies significantly according to the research scenario.

Rule-based methods were the first to be employed in the development of MT systems [2]. The task of building a rule-based MT system is laborious; the requirement for domain knowledge and the cost of assembling the experts is major. Also, the cost of making adjustment to the rules of the working system has been identified as a difficult problem; more still, adaptation to new language pair is nearly impossible as there is no way of avoiding a complete handcrafting of the new rule set that applies to the new language pair. The inability to scale and adapt an existing system to new languages without linguistic expertise is a major concern for rule-based methods. Rule-based methods are not as appealing nowadays, except perhaps when they are used in a hybrid multi-engine [1,3].

Empirical methods can be used to construct MT systems with appreciable results fairly quickly; strictly requiring data, algorithms and the ability to speak the language (in cases where data - parallel corpus- is not already available). Though empirical, data-driven methods have the advantage of being language-independent; their performance is correlated to the size, quality and coverage of the parallel corpus available for the development of the system [4]. Parallel corpora are a scarce resource for Nigerian languages; hence this work is geared toward addressing aspects of the data-related challenges.

The global demand for translation and translation tools currently surpasses the capacity of available solutions. Besides, there is no one-solution-fits-all, off-the-shelf solution for all languages. Thus, the need and urgency to increase the scale of research for the development of translation tools and devices continues to grow, especially for languages suffering under the pressure of globalisation. The translation research scenario across the global community is skewed in favour of the widely spoken languages like English, German, Spanish, French and other members of the league of dominant world languages. Africa accounts for at least two-thirds of the total number of World's languages [5], yet African languages are mostly low-resourced, lacking presence in the electronic media. Language processing tools are required in the quest for the emancipation of African languages; developing tools that promote the use and development of African languages is a research imperative in the globalised world of today since many of the languages still suffer marginalisation from the effects of the over-use of colonial languages. In Nigeria, English, the official language has continued to dominate and endanger indigenous languages – English is fast becoming the official language in many Nigerian homes today. Any keen observer would easily notice that nowadays, young Nigerians who share the same ethnicity are more comfortable discussing in English, effectively showcasing their agnostic tendencies towards their mother tongue.

Nigeria is an ethno-linguistically diverse country and as such developing machine translation systems for the languages represent a potent antidote to some of the perceived ill-effects of the dominance of English language use on our cultural and linguistic diversities and as well provide opportunities for improving the socio-economic well being of the citizenry. MT systems also are a necessary tool for creating culturally sensitive contents on the web and which would in turn create environments for multilingual diversity [6]. Thus the objectives of this research are to create parallel corpora for the development of MT systems for Nigerian languages starting with Igbo and Yorùbá and thereafter investigate the suitability of the parallel corpora for the SMT approach, given that existing indigenous researches have mainly applied rule-based methods. The research has two imports; first, it opens a new frontier for indigenous capacity building and utilisation for the development of corpora and data resources for Nigerian languages and second, it motivates future research on the development of MT systems for Nigerian languages using state-of-the-art approaches. This paper reports on the preliminary results from an English-to-Igbo, English-to-Yorùbá and Igbo-to-Yorùbá phrase-based SMT systems and on the findings of error analyses conducted on the translation outputs from each system. The systems being described are based on Moses SMT toolkit [7].

1.1 On the Syntactic Differences between the Study Languages

The difficulties that arise in MT, regardless of the techniques being employed can be attributed to a number of reasons. First, it is required to know the correspondences between words and phrases between the languages under study. Second, knowing and being able to handle syntactic, semantic and pragmatic differences between the language pairs is crucial to creating a good translation system [8]. A third source of difficulty in MT tasks arises from the relationships between language and culture. Culture and language are intertwined; language is culture and we communicate culture using language, in fact, it can be said that the cultural heritage of a people reflects in the way they speak. Nigerian languages embody the rich culture of the people but are expectedly unable to cater for concepts that are alien to our culture. For instance, certain named entities like mattress and calculator that are 'foreign' to our culture are difficult to be assigned a simple name in the manner it would be given to objects like mat (*ẹní* in Yorùbá; *ute* in Igbo) or stone (*òkúta* in Yorùbá; *okute* in Igbo) which have natural places in our original lives.

Igbo, similar to Yorùbá is considerably a right-branching language, with the SVO word order. English, on the other hand, is a language with right and left-branching structures; though with SVO word order too. The differences between these two languages of Nigeria and English are significant, even though they all share SVO word order. The fact that there are more adjectives in English language than any of Igbo and Yorùbá languages creates some subtle imbalances that these languages attempt to make up for by using descriptive constructs and complementation. While English language is stable in terms of positioning of adjectives and determiners relative to the position of the head noun, Igbo is fluid; much more than Yorùbá. Igbo and Yorùbá rather fluctuate between post-positioning and pre-positioning of adjectives, determiners, demonstratives, genitives, quantifiers depending on the construction [9,10,11].

Morphological inequalities also exist between these languages. Unlike English, Igbo and Yorùbá morphology is centred around the verb. English marks plurality of count nouns by addition of the morph +s for example, Igbo and Yorùbá do not have a way to directly mark plurality of count nouns. Rather they do so by post-positioning cardinal or ordinal numbers (depending on the construction) after the noun. Yorùbá and Igbo plurals are also formed by placing some appropriate plural marker before the noun. The possessive*'s* marker is absent in Igbo and Yorùbá languages. Other subtle differences exist but the discussion is limited to those which are of most important relevance to the presentation in this paper.

1.2 Error Analyses of Translation Outputs

Machine translation error analysis is an exercise aimed at diagnosing the existence, causes and effects of inadequacies (errors) in MT output that significantly impact on the adequacy and fluency of translations. Llitjos et al. [12] described an approach to automatically determine the necessary improvement to transfer rules for reducing errors in translation output in a transfer-based MT scenario. The error types according to these authors are: Missing word, extra word, wrong word order, incorrect word and wrong agreement. Vilar et al. [13], Bojar [14] and Costa et al. [15] are among some excellent research works that have extended the concept of error typology presented in [12]. However, the discussion of errors in our systems' output, is inclined to the classification by Costa et al. [15] because of the linguistic underpinnings of their taxonomy.

Essentially, Costa et al. [15] groups MT error into five types: Orthography, Lexis, Grammar, Semantic, and Discourse error. Extensive discussions of each of the types have been covered in the article. This paper draws inspiration from a number of related works, principally, [15,14,4,16,17,13]. These authors have separately examined and applied similar error analysis framework to MT outputs from various systems. Translation outputs from the systems being described in this paper were examined for orthographic, lexical, grammatical and semantic errors following the style of Costa et al. [15].

2. MATERIALS AND METHODS

2.1 Data Acquisition and Preparation

The development of SMT systems is critically dependent on the availability of large amounts of high quality parallel corpora with good coverage of the language genres. An important challenge for MT research involving Nigerian languages is the unavailability of publicly accessible parallel corpora. Although the Bible and its translated versions available in the majority of world languages can be seen as a single most easily accessible source of multilingual data for NLP research for disadvantaged languages, there is, however, a limit to how much parallel data that can be created from the Bible alone. This is the case because a number of books in the Bible are too figurative or idiomatic to be suitable for the development of open-domain-targeted machine translation systems. In constructing the parallel texts used in this research, data from the English version of the Bible and its translations in Igbo and Yorùbá languages were used as a first source. Further, the authors had a time-limited understanding with the WatchTower Society of Jehova's Witnesses in Nigeria to manually copy and use data from her website for use in the development of our research corpus.

Having collected text data for the three language pairs (English-Igbo, English-Yorùbá and Igbo-Yorùbá), the procedure described in [18] was employed for their transformation into parallel corpora. Following collection, document-level alignments were performed on the files to obtain an input to the sentence alignment algorithm. The Igbo and Yorùbá documents were then cleaned and orthographically normalised to restore tone-marks and other diacritic marks that were damaged during the process of data gathering. Sentence-aligned documents were produced using a re-implementation of the Gale and Church [19] algorithm, a popular algorithm for sentence alignment task. Running each pair of documents through the alignment software produces roughly useable sentence alignments from which well aligned sentences were then selected for inclusion into the final pool of sentence-aligned corpus for each language pair. Fig. 1 shows a sample segment from our parallel corpora for the three languages under study.

English	Yorùbá	Igbo
Do you know who a leper is? A leper is a person who has a sickness called leprosy. That sickness can even cause some of the person's flesh to fall off. When Jesus lived on earth, lepers had to live away from other people. And if a leper saw another person coming, he had to call out to warn that person to stay away from him. This was done so that other people would not get too close and maybe get the leper's sickness. Jesus was very kind to lepers.	Njé o mọ ẹni tí a ń pè níi adẹ́tẹ̀? adẹ́tẹ̀ ni eni tó ní àìsàn kan tó ńjẹ́ ẹ̀tẹ̀. Àìsàn yẹn tiẹ̀ lè mú kí apá kan ara èèyàn gé kúrò. Ní ìgbà àtijọ́ tí Jésù wà lórí ilẹ̀ ayé, àwọn adẹ́tẹ̀ kìí gbé pẹ̀lú àwọn èèyàn ní àarin ìlú, wọ́n máa ń gbé lọ́tọ̀ ni. Bí adẹ́tẹ̀ bá sì rí èèyàn kan tóń bọ̀ lọ́dọ̀ rẹ̀, ó ní láti tètè sọ fún eni náà pé kí ó dúró sọ́hùn-ún kí ó má ṣe dé ọ̀dọ̀ òun. Wọ́n máa ńṣe èyí torí kí àwọn èèyàn má ṣe súnmọ́ wọn kí àìsàn ẹ̀tẹ̀ náà má bàa ràn wọ́n. Jésù máa ń sàánú àwọn adẹ́tẹ̀ gan-an.	ị ma ihe bụ́ ekpenta? Ekpenta bụ orịa nke pụrụ obụna ime ka anụ ahụ́ mmadụ na-adapụ adapụ. N'oge Jizọs biri n'ụwa, ndị ekpenta na-ebi ebe dịpuru adịpu site n'ebe ndị ozo bi. ọ bụrụkwa na onye ekpenta ahu ka mmadụ na-abịa, ọ ga na-eti mkpu iji dọọ onye ahu aka ná nti ka ọ ghara ịbịa ya nso. A na-eme nke a ka ndi ọzọ wee ghara ịbịaru ha nso nke ukwuu ma eleghịkwa anya bute ọrịa onye ekpenta ahu. Jizọs nwere obiọma dị ukwuu n'ebe ndị ekpenta nọ.

Fig. 1. A Sample parallel paragraph extract from our corpus

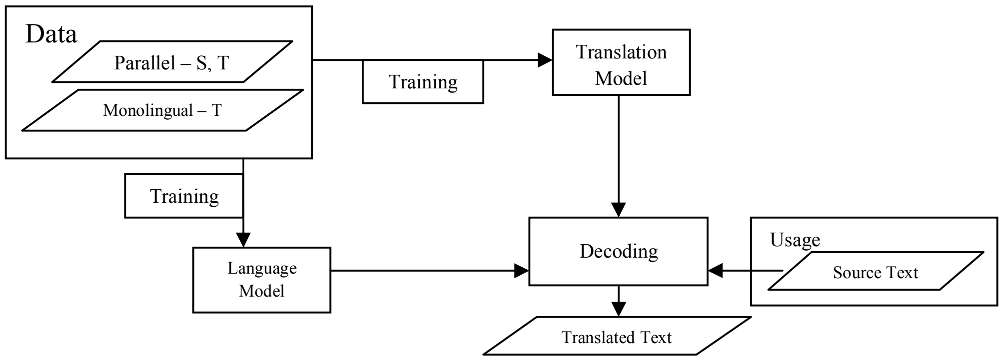

Fig. 2. System development process

Sentences that align to each other across the three language pairs were selected and used; this is to provide an objective basis for comparison of the performances of the resulting MT systems as noted in [18]. A total of 36, 787 parallel sentences was used to build each of the MT systems.

2.2 Methodology

The phrase-based SMT approach was employed. In phrase-based SMT modeling, sequences of words (the so-called phrases) are the translation units. Utilising phrase probabilities offer better results than when translation probabilities are conditioned on words [8,6,20]. This approach begins by grouping source language words into sequences of words: $s_1, s_2, s_3, \dots, s_l$, translate each phrase, s_i to the target phrase, t_i and then optionally reorder target phrases according to the target LM, utilising the distortion probabilities before combining them into the target sentence, the translation of the input source sentence.

The core probabilistic model components of a phrase-based SMT are the translation model (TM) probability, $\emptyset(s_i|t_i)$ and the distortion probability $d(a_i - b_{i-1} - 1)$. $a_i =$ the start position of the source phrase generated by the *ith* target phrase; $b_i =$ the end position of the source phrase generated by the *i – 1th* target phrase t_{i-1}. A distortion feature was used to measure and restrict the relative distances by which a phrase can be moved around between two language pairs and with penalty for large distortions. The TM for a phrase-based SMT is thus represented as:

$$p(s|t) = \prod_{i=1}^{l} \emptyset(s_i, t_i) d(a_i - b_{i-1} - 1) \quad (1)$$

Each of the systems uses an n-gram-based LM of order 3 trained on surface words only. Decoding is performed as a search and its model parameters were estimated from parallel corpora. The main set of parameters is the set of phrase translation probabilities, $\emptyset(s_i, t_i)$ which were learned from phrase alignment templates computed with Giza++ [21]. Fig. 2 represents the basic process that was followed in this methodology.

3. EXPERIMENTS, RESULTS AND DISCUSSION

3.1 Training and Tuning

Standard features were used and procedures for training, tuning and testing a typical Pb-SMT system as described in [22] were applied using Moses SMT toolkit. For each language pair, parallel corpora were aligned using Giza++ [21] and symmetrised alignment was generated using the grow-diag-final-and heuristics as described in [20].

Each language model is a trigram model trained on the target side monolingual corpus of 50,000 sentences using IRSTLM [23] with modified Kneser-Ney smoothing. Good-Turing scoring was employed for translation option scoring and performed tuning using minimum error rate training (MERT) on the development data set of 6,000 sentences. 10 iterations of MERT were run on a 100-best list of candidate translations. The respective TM uses a number of features that were combined in a log-linear fashion [24]. The basic features used are log-probabilities for phrase and lexical probability translations, word penalty, phrase penalty and distortion penalty. Phrases of length 5 were extracted from the aligned corpora and with an experimentally determined distortion limit of 6. The BLEU [25] measure to evaluate the performance of the systems. In all the experiments, language models (LM) and translation models (TM) were trained using identical settings as described in above. The data used for tuning and testing were adjusted to close ranges in terms of number of sentences.

3.2 Translation Performance

Table 1 shows the performance of each of the three systems in terms of BLEU score alongside a sample translation for each system. The performances of the English-Igbo and the English-Yorùbá systems are comparable. That of Igbo-Yorùbá however does not compare favourably to these. That said, the performances of these systems are encouraging, given the small size of the data on which they were trained. What is rather of concern is the low performance of the Igbo-Yorùbá system; it is 37.68% lower than the other two. Ordinarily, one would have expected a higher performance of the Igbo-Yorùbá system, given that the two are members of the same language family –The Niger-Congo family.

An examination of the translation by Igbo-Yorùbá system in Table 1 indicates that there exists a 'confusion' of sense error; the word '*igbu*' is wrongly interpreted as '*to kill*' rather than '*to kneel*'. Increasing the size of data should ameliorate this sort of failures since that would improve the contextual information about the word. The implication thus is that more data would be required to improve on the performances of the systems. It is also imperative to further probe into the responses of the Igbo-Yorùbá system for insights into the observed low performance.

Further, the translated outputs were analysed for errors with the aim of studying the failures of the systems in order to understand the causes of failures for the purposes of designing guided improvements to the baseline systems. The BLEU scores attained by the $en \rightarrow ig$ and $en \rightarrow yo$ systems are comparable; $ig \rightarrow yo$ system has a rather low BLEU score compared to the first two systems, despite the fact that the two languages share common characteristics, being members of the same language family. The target of our ongoing work is to improve on the quality and quantity of parallel corpus, enrich these baseline MT systems with extra features from linguistic sources and include advanced features of the Moses system.

3.3 Probing Failures in the MT Systems

Four hundred (400) sentences each were randomly selected from each of the translation system's output for analysis. The findings of the analyses are as indicated in Table 2. But for orthographic error class, the percentage of each error class is significantly high. Orthographic errors were found to be majorly wrongful insertion or failure to insert punctuation marks, particularly comma and semi colon.

Table 1. The translation scores with sample sentence translations

System	BLEU	Sample output
$en \rightarrow ig$	30.04	**T:** *But then, the Bible explains how we can live forever in God's new world.* **R:** *Ma, Bible na-akọkwa otú anyị pụrụ isi dị ndụ ebighị ebi n'ime ụwa ọhụrụ Chineke.* **O:** *Ma, Bible na-akọwa otú anyị pụrụ isi ndụ ruo mgbe ebighị ebi n'ụwa ọhụrụ Chineke.*
$en \rightarrow yo$	29.01	**T:** *so do you understand what the sign means?* **R:** *Nítorí náà, n`jẹ́ o mọ ohun tí àmì yẹn túmọ̀ sí?* **O:** *Nítorí náà, n`jẹ́ o mọ ohun tí àmì túmọ̀ sí?*
$ig \rightarrow yo$	18.72	**T:** *ì kwesịrị igbu ikpere n'ala?* **R:** *Ṣé ó yẹ kó o kúnlẹ̀ ?* **O:** *Ṣé ó yẹ kí o pạ ?*

T – Test, R – Reference translation, O – Output from the translation system

Table 2. Percentage of error types in a fraction of the output analysed

Category	$en \rightarrow ig$	$en \rightarrow yo$	$ig \rightarrow yo$
Orthography	6.82	3.83	5.92
Lexis	28.41	25.33	29.32
Grammar	39.77	43.24	42.72
Semantics	25.00	27.60	22.04

All spelling errors were traced to the training corpora. All categories of lexical errors were found in the three systems with the 'omission' and 'untranslated' errors being predominant. Error rates at the grammar level are the highest for all the systems. Leading grammatical errors are due to wrong ordering, mis-selection due to tense, person, and number. This is traceable to the divergences in the way that each of the languages handles these syntactic phenomena. At the semantic level, 'confusion of senses' (as seen in the Igbo-Yorùbá system output in Table 1) and 'wrong choice of word' are the most predominant. Collocation errors exist more in the $en \rightarrow ig$ and $en \rightarrow yo$ systems, especially when the concept portrayed in the input sentence does not expressly exist in the culture of the target language.

Specifically, the most frequent errors occur at the lexical, grammatical and semantic levels. The size of training corpus is a strong contributory factor to the kinds and magnitude of errors (failures) that exist in SMT systems [4]. Hence, one of the problems that must be addressed is that of data inadequacies.

4. CONCLUSION

This paper has discussed the creation of a research-size parallel corpus each for the development of phrase-based SMT systems for English-Igbo, English-Yorùbá and Igbo-Yorùbá language pairs, using data from the domain of religion. The performance of the MT systems is encouraging with the English-Igbo and English-Yorùbá systems comparing better in terms of BLEU scores. The Igbo-Yorùbá system exhibited poorer performance relative to the first two despite being of the same language family. An error analysis of the systems' outputs was conducted using a linguistically motivated MT error analysis approach, and it showed that errors occurred mostly at the lexical, grammatical and semantic levels. Evidence exist in research that the incidence rates of these errors can be reduced by increasing the quantity and quality of training corpora and at the same time employing more algorithmic tact.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Costa-Jussa MR, Fonollosa JA. Latest trends in hybrid machine translation and its application. Computer Speech and Language. 2015;32(1):3-10.
2. Hutchins J. Machine translation: A concise history. Computer Aided Translation: Theory and Practice. 2007;13:29-70.
3. Costa-Jussa MR, Rapp R, Lambert P, Eberle K, Banchs RE, Babych B. Hybrid approaches to machine translation. Switzerland: Springer International Publishing; 2016.
4. Popovic M, Ney H, de Gispet A, Marino JB, Gupta D, Federico M, Lambert P, Banchs R. Morpho-syntactic information for automatic error analysis of statistical machine translation output. In: Proceedings of the Workshop on Statistical Machine Translation, New York, USA. 2006;1-6.
5. Simons GF, Fenning CD. Ethnologue: Languages of the world. Twentieth Edition. Dallas: Texas; 2017.
   Available:http://www.ethnologue.com
   (Accessed on April 2nd, 2018)
6. Montviloff V. Meeting the challenges of language diversity in the information society. World Congress on Language Policies, Bacelona; 2002.
   Available:www.linguapax.org/wp-content/uploads/2015/03/CMPL2002_T5_Montviloff.pdf
   (Accessed on April 2nd, 2018)
7. Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic. 2007;177-180.
8. Koehn P, Och FJ, Marcu D. Statistical phrase-based translation. In: Proceedings of the Joint Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics, Edmonton AB, Canada. 2003;127-133.
9. Umeodinka AU. Corpus-based contrastive analysis of Igbo and English adjectives. Journal of Igbo Language and Linguistics. 2011;3:54-62.
10. Uzoigwe BC. A contrastive analysis of Igbo and English determiner phrases. Journal of Igbo Language and Linguistics 2011;3:73-83.

11. Adelabu B. A contrastive analysis of adjectives in English and Yorùbá. International Journal of Educational Research. 2014;2:509-524.
12. Llitjos AF, Carbonell JG, Lavie A. A framework for interactive and automatic refinement of transfer-based machine translation. In: 10th EAMT Conference on Practical Applications of Machine Translation, Budapest, Hungary. 2005;87-96.
13. Vilar D, Xu J, D'Haro LF, Ney H. Error analysis of machine translation output. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy. 2006;697-702.
14. Bojar O. Analyzing error types in English-Czech machine translation. The Prague Bulletin of Mathematical Linguistics. 2011;95:63-76.
15. Costa A, Ling W, Luis T, Rui C, Coheur L. A linguistically motivated taxonomy for machine translation error analysis. Machine Translation. 2015;29:127-161.
16. Popovic M, Ney H. Error analysis of verb inflections in Spanish translation output. In: TC-STAR Workshop on Speech-to-speech Translation, Barcelona, Spain. 2006;99-103.
17. Codon SL, Parvaz D, Abadeen JS, Doran C, Freeman A, Awad M. Evaluation of machine translation errors in English and Iraqi Arabic. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valeta, Malta. 2010;159-168.
18. Koehn P. Europal: A parallel corpus for statistical machine translation. In: Proceedings of the Tenth Machine Translation Summit. 2005;5:79-86.
19. Gale W, Church K. A program for aligning sentences in bilingual corpora. Computational Linguistics. 1993;75-102.
20. Koehn P. Statistical machine translation. Eaglewood Cliffs: Prentice-Hall; 2010.
21. Och FJ, Ney H. Giza: Training of statistical translation models. Available:http://www.fjoch.com/giza-training-of-statistical-translation-models.html (Accessed February 10th, 2018)
22. Huck M, Birch A. The Edinburgh machine translation systems for IWSLT 2015. In: Proceedings of the International Workshop on Spoken Language Translation, Da Nang, Vietnam. 2015;1-38.
23. Federico M, Bertoldi N, Cettolo M. IRSTLM: An open source toolkit for handling large scale language models. In: Interspeech. 2008;1618-1621.
24. Och FJ, Ney H. Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia. 2002;295-302.
25. Papineni K, Roukus S, Ward T, Zhu WJ. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia. 2002;311-318.

Peer-review history:
The peer review history for this paper can be accessed here:
http://prh.sdiarticle3.com/review-history/26773