Neural Machine Translator for English to Tamil Translation using K-Means Clustering Algorithm

Dr. (Mrs)V.VIDYAPRIYA. Associate Professor, PG and Research Department of Computer Science, Quaid-E-Millath Govt. College for Women (Autonomous), Chennai, India, vidyapriya2128@gmail.com

E.LAVANYA SHREE. M.Phil Scholar, PG and Research Department of Computer Science, Quaid-E-Millath Govt. College for Women (Autonomous), Chennai, India, venillalavanyashree@gmail.com

Abstract - The research work aims in developing English to Tamil translation using Neural Machine Translation. In an existing word by word translation system there are lot of issues and some of them are ambiguity, Out-of-Vocabulary words, word inflections, and improper sentence structure. To handle these issues, proposed architecture is designed in such a way that, it contains Improved Part-of-Speech tagger, machine learning based morphological analyser, collocation based word sense disambiguation procedure, semantic dictionary, and tense markers with gerund ending rules, and transliteration algorithm. The NMT system allows user to supply search queries in the form of text in one's native language, which are then translated and used to retrieve relevant information in other languages. The objective of this research work is to develop English to Tamil translation system that accepts source text in English language, and retrieves equivalent information in Tamil language. The translation of queries from one language to another is done using Rule Based Approach and Statistical Approach.

Keywords — Neural Machine Translation(NMT), Out-of-Vocabulary words(OOV), Rule Based Approach, Statistical Approach, Word inflections, Word Sense Disambiguation(WSD).

I. INTRODUCTION

Language is not only the means of communication. It could influence our culture and in fact, it influences the thought process of human beings. It is an important element of culture and through the language the culture can be learned and preserved. The native languages all over the world are growing rapidly along with the growth of technology, in general, and information technology, in particular. On the one hand, the world experiences a growth in the native language and on the other hand, precious and nascent information comes through foreign languages.

Knowledge of the mother tongue alone is no longer enough to follow the information supplied by the other languages. Because of this ever-increasing gap and the speed with which information is supplied, there is a possibility of death knell for many native languages. Recent research shows language death is accelerated to the rate of two languages per month. It is necessary to bridge this gap with the help of modern technologies as early as possible, thus enables the information supplied by the other languages available in the native language.

The Global Reach statistics also shows that nearly 80% of the web users prefer to access the Internet in their native languages. According to Global Internet Statistics, over 64% of the global web users are non-English speakers. Non-English speakers are the fastest growing group of new web users and there is a growing interest in non-English sites as the web becomes truly multi-lingual. India is a multilingual country and it ranks second position globally in the usage of internet.

Tamil, a Dravidian language spoken by around 75 million people is the official language of Tamil Nadu state government of India. Tamil in its eagerness to gather information from English resort to build English-Tamil translation systems. Many English-Tamil translation systems are getting built, but none could serve the ambitious need of Tamil.

1.1 INTRODUCTION TO MACHINE TRANSLATION

Machine translation is the task of translating the text in source language to target language, automatically. Machine translation can be considered as an area of applied research that draws ideas and techniques from linguistics, computer science, artificial intelligence, translation theory, and statistics. Even though machine translation was envisioned as a computer application in the 1950's and research has

 DOI : 10.18231/2454-9150.2018.1182

been made for 60 years, machine translation is still considered to be an open problem. The demand for machine translation is growing rapidly.

This notion incites to work on Neural Machine Translation (NMT). The mind blowing applications of NMT and its potentiality and impact in future eventually pursues and led me to develop a NMT system for English to the one of the longest surviving classical, could be the world's oldest surviving, literature rich, culturally significant language Tamil.

1.2 INTRODUCTION TO NEURAL MACHINE TRANSLATION

The NMT system allows user to supply search queries in the form of text in one's native language, which are then translated and used to retrieve relevant information in other languages. NMT models use Deep learning and Representation Learning. Deep learning is a subset of machine learning in Artificial Intelligence (AI) that has networks capable of learning unsupervised from data that is unstructured or unlabeled. It is an artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Representation learning is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data. This replaces manual feature engineering and allows a machine to both learn the features and use them to perform a specific task. The translation of queries from one language to another is done using Rule Based Approach and Statistical Approach.

1.3 OBJECTIVE

The aim of this research is to translate the English input sentence to Tamil sentence as close as the human translation and to get comprehendible translated Tamil sentence. This research primarily focuses on the development of the NMT system to translate English to Tamil and depends on the success of the prototype model of the NMT system; the approach that employed for the prototype cab be extended to the English to Dravidian languages like Malayalam, Kannada and Telugu.

II. LITERATURE REVIEW

Pasindu Tennage, Prabath Sandaruwan, Malith Thilakarathne, Achini Herath, Surangika Ranathunga, Sanath Jayasena and Gihan Dias [1] in their work they have proposed "Neural Machine Translation for Sinhala and Tamil Languages". In this paper they have addressed the task of developing a NMT system for the most widely used language pair in Sri Lanka- Sinhala and Tamil, focusing on the domain of official government documents. They explore the ways of improving NMT using word phrases in a situation where the size of the parallel corpus is considerably small, and empirically show that the resulting models improve our benchmark domain specific Sinhala to Tamil and Tamil to Sinhala translation models by 0.68 and 5.4 BLEU. In this paper they also presents an analysis on how NMT performance varies with the amount of word phrases, in order to investigate the effects of word phrases in domain specific NMT.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo and Marcello Federico [2] have proposed a research work on "Neural versus Phrase-Based Machine Translation Quality: a Case Study". This article examines to understand in what respects NMT provides better translation quality than PBMT. They perform a detailed analysis of neural vs. phrase-based SMT outputs, leveraging high quality post-edits performed by professional translators on the IWSLT data. Their analysis provides useful insights on what linguistic phenomena are best modelled by neural models such as the reordering of verbs. The outcome of their analysis confirms that NMT has significantly pushed ahead the state of the art, especially in a language pair involving rich morphology prediction and significant word reordering.

Rico Sennrich, Barry Haddow and Alexandra Birch [3] have proposed a "Edinburgh Neural Machine Translation Systems for WMT 16" based on news translation task by building neural translation systems for four language pairs, English ←→Czech, English ←→ German, English ←→ Romanian and English ←→ Russian each trained in both directions. Their systems are based on an attentional encoder-decoder, using BPE subword segmentation for open-vocabulary translation with a fixed vocabulary. They have experimented with using automatic back-translations of the monolingual News corpus as additional training data, pervasive dropout, and target-bidirectional models. All reported methods give substantial improvements. In the human evaluation, this system was the best constrained system for 7 out of 8 translation directions. For all translation directions, they have observed large improvements in translation quality from using synthetic parallel training data which obtained by back-translating in-domain monolingual target-side data.

Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio [4] in their research work they proposed an "Neural Machine Translation By Jointly Learning To Align And Translate" here they have conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this they have achieved a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-

 DOI : 10.18231/2454-9150.2018.1182

French translation. A translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation.

Karthik Revanuru, Kaushik Turlapaty and Shrisha Rao [5] in their work they have proposed "Neural Machine Translation of Indian Languages" In this paper they have studied and applied NMT techniques to create a system with multiple models which we then apply for six Indian language pairs. They compared the performances of their NMT models with system using automatic evaluation metrics such as UNK Count, METEOR, F-Measure, and BLEU. They demonstrated that they can achieve good accuracy even using a shallow network. On comparing the performance of Google Translate on their test dataset, their best model has outperformed Google Translate by a margin of 17 BLEU points on Urdu-Hindi, 29 BLEU points on Punjabi-Hindi, and 30 BLEU points on Gujarati-Hindi translations. They also tried using Convolutional Neural Networks since convolutional approach allows us to discover compositional structure in the sequences more easily since representations are built hierarchically. Optimizing this architecture, this model can be run on embedded devices which opens up new possibilities and gives us the freedom for offline translation.

III. PROPOSED SYSTEM

The research work proposes Neural Machine Translation (NMT) approach, it allows user to supply search queries in the form of text in one's native language, which are then translated and used to retrieve relevant information in other languages. This approach will produce an efficient method to translate English words to Tamil words. The proposed NMT approach combines rule-based machine translation approach and statistical approach to perform English-Tamil Translation. This research work consists of several key tasks that are to be performed in a sequential order to effectively convert the English words to its Tamil equivalent words. The tasks are listed below and the architectural flow is presented in Figure.

- Tokenization
- Pre-processing
- Translation
- Transliteration and error correction
- Query Expansion
- Information Retrieval

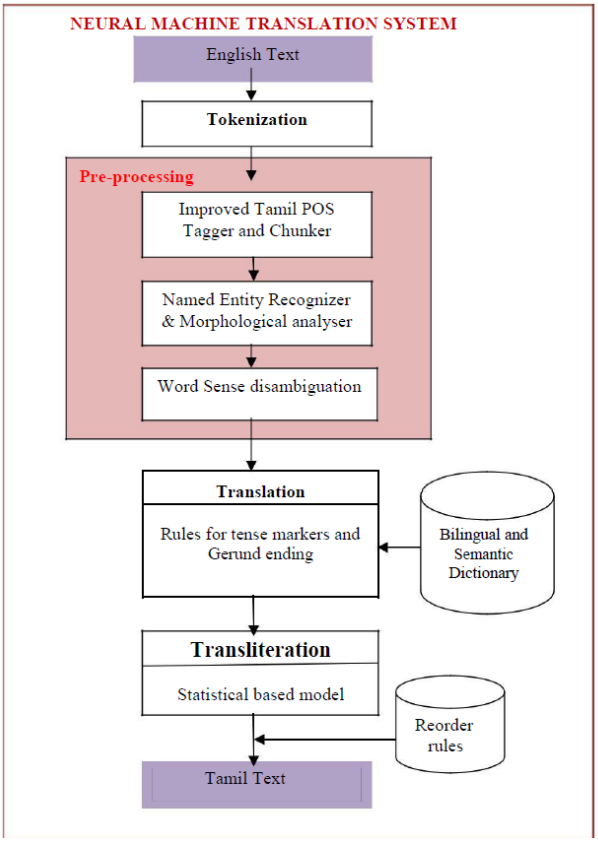

Fig.1 Architecture of NMT

3.1 TOKENIZATION

The first step of NMT approach is tokenization, which is the process of breaking up the input text into units called tokens (words). This process generally use some special symbols like punctuation marks (eg. or -) or spaces as delimiters during word separation. This research work uses blank space as word separator.

3.2 PRE-PROCESSING

The pre-processing step of NMT approach performs five major tasks, namely,

- Part-of-Speech Tagging
- Chunking
- Named Entity Recognition
- Morphological Analysis
- Word Sense Disambiguation.

Part-of-Speech tagging (POS)

The first step in pre-processing of any language sentence is to retrieve Part Of Speech information that helps in processing many language related activities. POS tagging is defined as a task that reads a set of texts and assigns part of speech label to each of them. As English is highly an inflectional language, for tagging each word, one has to depend on the syntactic function or context to decide upon whether the word is a noun, adjective, adverb or postposition. This leads to a complexity in English POS tagging.

Chunking

 DOI : 10.18231/2454-9150.2018.1182

It is a Natural Language Process that separates and segments sentences into their sub constituents such as noun, verb and prepositional phrases. Examples of chunks include noun phrases, prepositional phrases and verb phrases. Chunking works on POS tagged text, so its accuracy depends upon the accuracy of POS tagger.

Named Entity Recognition (NER)

Named entities include the identification of people names, location and companies / organizations, while digits may include time/date stamp and amount. In an English sentence, the NER identifies words that need to be transliterated, and the remaining words are translated using dictionary.

Morphological Analysis (MA)

The purpose of an MA is to return root word, and their grammatical information of all the possible word classes for a given word. MA also includes extraction of the grammatical information including number, gender and tense information for all the tokens. As Indian languages have a rich inflectional morphology, MA is an essential tool for such languages.

Collocation based Word Sense Disambiguation

Word Sense Disambiguation (WSD) algorithm is used to handle collocations. Collocations are defined as nearby words, that strongly suggests the sense of the ambiguous word, in a given occurrence. WSD is an important and challenging task during translation. In general, a WSD algorithm initially uses a manual process to extract collocations, and it identifies sense-collocation words related to the identified collocation using either a dictionary or a thesaurus. This process is time consuming, and the manual process may introduce errors. To solve this issue, in this research work, the manual collocation extraction process is replaced using an automatic extraction procedure that uses an enhanced K-Means clustering algorithm.

3.3 TRANSLATION

The next step after pre-processing is translation, and it is carried out using knowledge sources and rules set. The outputs of morphological analyser are root word and grammatical morphemes. The root words are directly translated using bi-lingual dictionary. Sometimes several words may not found in the bi-lingual dictionary called OOV words. If a word is not found in the bi-lingual dictionary, then it is searched in semantic dictionary to obtain an equivalent Tamil word. The semantic equivalent word of the OOV word is translated using root word dictionary. A single word may have several translations, and this ambiguity problem is handled using word sense collocation dictionary. WSD procedure helps in choosing the best hypothesis translation from all possible translations. The remaining part of word belongs to grammatical categories which are translated by applying tense marker and gerund ending rules.

3.4 TRANSLITERATION WITH ERROR CORRECTION

Transliteration is task of converting one form of script to another form of script. The words that cannot be translated using dictionary are named entities. The NER identify named entities, and give the entities as input to the transliteration engine. Proper nouns and common nouns are often appears in transliterated forms. Transliteration is first performed to convert named entities and numbers. The first pass retrieval is carried out using a character transformation procedure in which it converts each English character to its Tamil equivalent. For this purpose, an English-Tamil Character Mapping Table is used.

3.5 QUERY EXPANSION

The Query Expansion is defined as the task of reformulating the translated text by selecting or adding terms to the text. The main goal here is to minimize the query-document mismatch and to maximize the retrieval performance. Inclusion of query expansion in CLTR, in general, can improve the retrieval performance by 4-15%.

3.6 INFORMATION RETRIEVAL

Information retrieval is a process of retrieving relevant information related to the user text. It uses Lucene indexer (Lucene is an open source library), which consists of modules for indexing. It is a full-featured text search engine.

K-MEANS CLUSTERING

K-Means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

For example consider the word bass in these statement

Statement 1: I went fishing for a few ocean bass

Statement 2: The bass line of the song is simply too weak.

It is obvious that the primary sentence is victimising the word "bass (fish)", and also the second sentence, the word "bass (instrument)" is getting used within the second sentence. As the word bass has multiple meanings then it will become ambiguous at the time of translation. The good quality of translation can only be achieved by choosing a right sense of an ambiguous word, and this process of identifying a correct sense for a word is done using WSD procedure.

The main concern of K-Means algorithm is an optimal selection of 'K' parameter, which is solved using an ensemble approach. An ensemble of clustering algorithms is built with different K values ranging

 DOI : 10.18231/2454-9150.2018.1182

between 2 to 30. This ensemble generates a set of clusters. Majority voting algorithm is then used to find the optimal clustering set from the different partitions created, thus estimating the optimal K value for clustering. The advantage of this approach is that the estimation of this K value is embedded during the process of clustering and requires no extra optimization procedures. The advantage of using automatic extraction step is that it can save search time while considering large number of ambiguous words in a language and reduces manual errors.

Generally clustering may take number of iterations, eventually to cluster centroids. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. In the proposed K-means clustering, it identifies the equivalent Tamil output text for the given English input text. This algorithm is often generated by assigning input text to the similar cluster output text. Output data are clustered based on input feature similarity. The algorithm is very fast to run it multiple times with different starting inputs. K-means clustering works as follows:

Input: S (instance set), K (number of cluster)

Output: clusters

1: Initialize K cluster centers.

2: while termination condition is not satisfied do

3: Assign instances to the closest cluster center.

4: Update cluster centers based on the assignment.

5: end while

NEURAL NETWORK CLASSIFICATION

An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it. The signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The connections between artificial neurons are called edges. Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times. This type of classification algorithm represents each cluster by a neuron or "prototype". The input data is also represented by neurons, which are connected to the prototype neurons. Each such connection has a weight, which is learned adaptively during learning.

In the Neural Machine Translation, Neural Network classification is used to classify the Tamil output text based on the input text. To translate the English input text to the Tamil output text, one input text may have many suggested output texts with similar meanings. This classification is done to get the perfect translated output text from the many suggested texts.

IV. RESULTS & FINDINGS

The proposed research work is based on Machine Translation that is used to translate text from English to Tamil language. This Neural Machine Translation System supports to user to get the efficient translation. The system can be easily accessible by the users for accurate translation from English to Tamil text. This work will help the user to enhance the knowledge in Tamil language.

Fig.2: Translation of English Text to Tamil Text

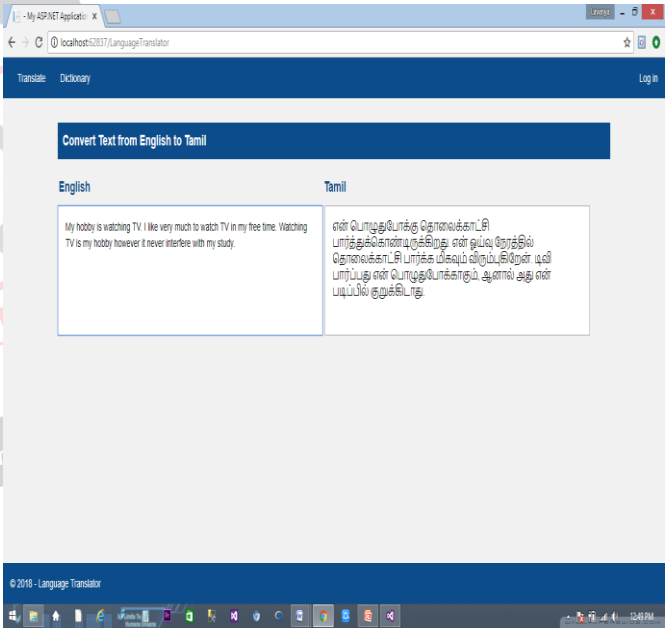

Fig.3: NMT Dictionary Page

 DOI : 10.18231/2454-9150.2018.1182

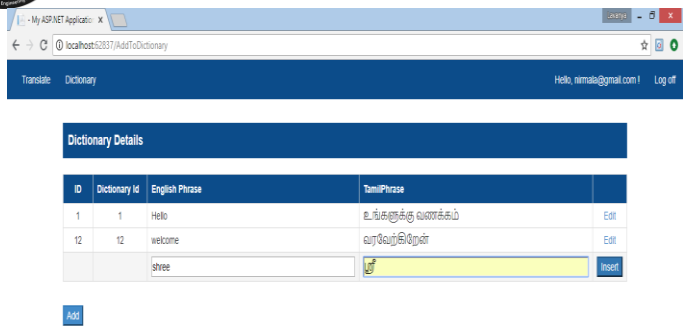

Table 1: Comparison of Google Translator and NMT System

ENGLISH INPUT TEXT	GOOGLE TRANSLATE	NMT SYSTEM
Kalai Vani	காலே வாணி	கலை வாணி
Tamil Arasi	தமிழ் அகாதமி	தமிழ் அரசி
Bat	பேட்	□□□□□
Community	சமூகத்தில்	சமூக□□
Bag	பையில்	பை
Box	பெட்டியில்	பெட்டி
Cover	கவர்	□□□
Bed sheet	படுக்கை விரிப்பால்	படுக்கை விரிப்□□
Home	வீட்டில்	வீ□□
Perverse	விபரீதமான	விபரீத□□
Paint	வரைவதற்கு	□□ய□□
Bunglow	□□□□□□	□□□□கை
Adaptor	அடாப்டர்	இசை□□க்□□
Rocket	ராக்கெட்	□□□க□□
Meaning	அதாவது	அர்□□□□□
Pages	பக்கங்களை	பக்கங்க□□
Locker	லாக்கர்	பெட்டக□□
Calculator	கால்குலேட்டர்	□□□□□□□□□□
Lady's Finger	பெண்ணின் விரல்	□□□□□□க்காய்
Pencil	பென்சில்	□□□□□□□□□

The above Table 1 is the comparison of Google Translator and the NMT System. Google provides an online translator for English to Tamil translation. Google translation from English to Tamil has been checked with many Tamil words and following are some of the erroneous outputs obtained. The Table 1 provides the example of erroneous outputs obtained by google translator, which was corrected by the NMT System. The corresponding correct translations are also provided in the Table 1.

V. CONCLUSION

The present research entitled "English to Tamil translation using Neural Machine Translation" in a novel attempt in the area of machine translation from English to Tamil. This Neural Machine Translation accepts the input text in the English language and translates the output text into Tamil language. The proposed NMT system is a combination of both rule based approach and statistical approach. Rule based MT system involves several tasks such as tokenization, pre-processing, and translation. The transliteration is carried out using statistical MT system. The proposed English to Tamil translation system showed effective results when compared to existing system.

VI. FUTURE WORK

This work translates only the English input text to Tamil output text, where as in future it can be extended to many languages. It can be extended to analyses the performance of English document to Tamil document translation process and vice versa. Semantic or ontology based text retrieval can also be probed and combined with the proposed classification algorithm in the future. It can also be extended to upload PDF files and have the translated version displayed using Machine Translation.

REFERENCES

[1] Pasindu Tennage, Prabath Sandaruwan, Malith Thilakarathne, Achini Herath, Surangika Ranathunga, Sanath Jayasena and Gihan Dias - Neural Machine Translation for Sinhala and Tamil Languages, ISBN: 978-1-5386-1981, IEEE 2017.

[2] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo and Marcello Federico - Neural versus Phrase-Based Machine Translation Quality: a Case Study, Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016.

 DOI : 10.18231/2454-9150.2018.1182

[3] Rico Sennrich and Barry Haddow and Alexandra - Edinburgh Neural Machine Translation Systems for WMT 16, Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, 2016.

[4] Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio - Neural Machine Translation by Jointly Learning to Align and Translate, Proceedings at ICLR 2015.

[5] Karthik Revanuru, Kaushik Turlapaty and Shrisha Rao - Neural Machine Translation of Indian Languages, Association for Computing Machinery, ISBN 978-1-4503-5323, 2017.

[6] Dr.R.Padmamala - Word Level Translation (Tamil-English) with Word Sense Disambiguation in Tamil Using OntNet, ISBN: 978-1-4799-7623, IEEE 2015.

[7] Parth J. Vasoya, Tarjni Vyas - A Survey on Word Sense Disambiguation Approaches, International Journal of Trend in Research and Development, Volume -1(1), 2014.

[8] S. M. Fakhrahmad, A.R. Rezapour, M.H. Sadreddini and M. Zolghadri Jahromi - Machine Translation Based on Data Mining and Deductive Schemes, Proceedings of the World Congress on Engineering, Volume II, 2012.

[9] Xing Wang, Zhaopeng Tu ,Deyi Xiong and Min Zhang - Translating Phrases in Neural Machine Translation, Proceedings Conference on Empirical Methods in Natural Language Processing, 2017.

[10] Mary Priya Sebastian, Sheena Kurian K, G. Santhosh Kumar - English to Malayalam Translation: A Statistical Approach, ISBN: 978-1-4503-0194, 2010.

 DOI : 10.18231/2454-9150.2018.1182