International Journal of Conceptions on Computing and Information Technology
Vol. 7, Issue. 1, November' 2019; ISSN: 2345 - 9808

Deep learning approach for Malagasy text summarization

Volatiana Marielle Ratianantitra, Jean Luc Razafindramintsa, Thomas Mahatody and Victor Manantsoa
University of Fianarantsoa
Laboratory of Computer Science and Mathematics Applied to Development (LIMAD)
301 Fianarantsoa, Madagascar
{volatianamarielle, razafindramintsa.jeanluc}@yahoo.fr, tsmahatody@gmail.com and vmanantsoa@moov.mg

Abstract— **In this paper, we deal automatic text summarization in Malagasy language. Malagasy is the national language of Madagascar. Several techniques have been developed to summarize texts in English and other European languages etc. But few research works has been done to summarize the Malagasy text. Therefore, we propose an automatic text summarization approach using deep learning. The main steps of the methodology used include preprocessing, processing and summary generation. These 3 steps work together to assimilate relevant information and generate a coherent and understandable summary. We explore various features and use a Restricted Boltzmann Machine to enhance and summarize these features to improve the resulting accuracy without losing any important information. To validate our strategy, we conducted a case study with the Tontona corpus and applying the proposed approach, the performance judging parameter f-measure has got values, 0.49, 0.55 and 0.50 respectively for the three under corpus sets.**

Keywords- text summarization; deep learning; malagasy language; relevant information

I. INTRODUCTION

The expansion of the internet network around the world has made available documents written in different languages. Text summarization is essential to obtain the important information while dealing with the large collection of documents. Remembering the details of each piece of information is not possible for the human mind. Therefore, text summarization system is needed. The purpose of text summarization system is to produce a condensed representation of an information source, in which the "important" information of the original content is preserved. The sources of information that can be summarized are multiples and heterogeneous: video, audio or text. Our studies will focus only on the text summarization. A summary can be produced from one or more documents.

Several methods are proposed to summarize major languages such as English, European languages, etc.; some problems still remain to be solved for other languages of the world such as the Malagasy language. The Malagasy language can boast of being the product of a miscegenation pushed between the languages of the first occupiers Austronesian, African and Arab sailors [1]. Nowadays, most of the vocabularies used have been enriched over the contacts with foreigners (French, English) but still have their own peculiarities. In the field of Natural Language Processing (NLP), there is some research work on the Malagasy language [2] [3]; but concerning text summarization, to our knowledge, there is only one research work that is considered simple. Therefore, new technique of Malagasy text summarization has been proposed in this paper.

The output of the summary can be of two types: extractive and abstractive. Most researchers have relied on systems based on extraction techniques whose principle is to bring out the relevant information by selecting the phrases that characterize it. The importance of sentences is determined by their statistical and linguistic characteristics. Abstractive summary are produced by rewording sentences of the source text. They use linguistic methods to examine and interpret the text and then search for new concepts and terms to best describe it. The approach provided in this project uses extractive summary.

The rest of the paper is organized as follows: Section 2 describes the reason for doing this work and Section 3 presents some particularities of the Malagasy language. Section 4 presents a brief survey of document text summarization methods that use deep learning. The proposed summary method is presented in Section 5. Section 6 describes the experiments and the last section concludes the paper and gives some future works.

II. MOTIVATION

In this study, we apply the Malagasy language because it is rare to study in the field of NLP and we use deep learning algorithm for the task of text summarization. As we know, the Malagasy language is known as a low-endowed language, deep learning can manipulate that kind of language. Deep learning is the emerging field of machine learning, which is used to solve computer domain number problems such as image processing, robotics, motion. Recently, it is also used in the field of NLP with very encouraging results. An algorithm is deep if its input has gone through many of the non-linearities before being released the most modern learning algorithms, including svm and naive ayes workbook are shallow. Here we use the

Restricted Boltzman Machine (RBM) to extract the highest word text feature.

III. PECULARITIES OF THE MALAGASY LANGUAGE

Malagasy is the national language of Madagascar. Several varieties of Malagasy are spoken in Madagascar. The official Malagasy MO (spoken in the capital) and eighteen regional dialects are generally distinguished [4]. It's important to note that our work focuses on the MO, the most widely used written form in Madagascar. Before starting the study of Malagasy text summarization, it's essential to present the important peculiarities of the Malagasy language. Malagasy is written with the Latin character. That means, there is officially no form of Malagasy-specific letter. On the other hand, compared to the French and English alphabets, those of Malagasy do not have the letters "c", "q", "u", "w" and "x" [2], except the proper noun.

A. Lexicon and morphology

Most linguists and traditional grammarians agree that since the Malagasy name does not know a gender (male, feminine), the adjective and the verb do not know gender flexion. Only the demonstrative determinant and the personal pronoun can receive number marks. The plural can be identified by the -reo- infix.

In the verb, the inflectional marks of time express grammatical meanings of three types: m- indicates the present, h- the future and the n- the past. The three times exist in Malagasy as illustrated in (a), but the verbs are not bent according to the person or the number. Some adjectives conjugate as also the verb. They accept temporal morphemes.

***M**ihinana izy 'he eats'*
***N**ihinana izy 'he ate'* *(a)*
***H**ihinana izy 'he will eat'*

The Malagasy derivational morphology is very rich in the sense that several prefixes, suffixes, infixes and circumfixes can be combined with a given radical. In this case, verbs, nouns and adjectives can be derived. . For example, with the radical *tia* 'love', we can have the following derived forms:

mi- + -tia = mitia 'love'(verb)
mpi- + -tia = mpitia 'lover'(noun)
fi- + -tia = fitia 'love'(noun)
fi- + -tia + v + ana = fitiavana 'love' (noun)
f -+-if-+anka-+tia+v+-ana = fifankatiavana 'mutual love' (noun) …

In addition, the Malagasy spelling uses various signs such as the accent, the apostrophe, the hyphen and the capital letter (initial of the names). And like all languages, the Malagasy language also has stopwords.

B. Syntax

The Commonly used basic sentences correspond to the PCS (Predicate-Complement-Subject) schema. The predicate (core of a sentence) is followed by its possible complements and its subject. It can be a verb, an adjective, a noun, an adverb and a pronoun [5].

Mametraka voninkazo eo ambony latabatra Maria
'Maria puts flowers on the table.' *(b)*

In (b), *mametraka* '*put*' is the main verb, *voninkazo* '*flowers*' is the object, *eo ambony latabatra* '*on the table*' the adverbial and Maria is the subject.

In Malagasy, the subject may be anteposed by two: *dia* and *no* particles. With *di*a, it is the theme that is emphatized. With *no*, it is the rheme that is focused (as new information).

*Maria **dia** mametraka voninkazo eo ambony latabatra*.
'Maria puts flowers on the table'
*Maria **no** mametraka voninkazo eo ambony latabatra*.
' It's Maria who puts flowers on the table.'

According to [5], there are two types of basic sentences according to whether the first complement is a direct object or an indirect object. The sentence is transitive in the first case and intransitive in the second.

IV. RELATED WORK

Automatic text summarization is an axis of research still open. The first initiative to evoke the idea of producing summaries automatically was proposed by Luhn in 1958 [6]. He used frequency of word occurrence to produce summaries of technical documents with the aim of determining the relevance of a document. And over time, new approaches have emerged such as the statistical based approach, graph based approach, topic based approach, discourse based approach and machine learning but mainly for English and European languages. In recent years, more and more searchers have used deep learning to solve problems with text translation and summarization. Deep learning is a branch of machine learning that is also a branch of Artificial Intelligence (AI). It is a machine learning based on neural networks inspired by the human brain.

[7] proposed an artificial intelligence approach. They developed an AI text summarization system architecture with three models: statistical model, machine learning model and deep learning model, as well as evaluating the performance of three models by ROUGE but the paper is based on the fact that they have proposed an automatic text summarization system by applying deep learning to generate short summaries from the titles and summaries of the Web of Science (WOS) database.

[8] proposed a system that uses deep learning to speed up the classification process and recommend relevant documents. The proposed algorithm RCNN, combination of RNN and CNN, guarantees the semantic correlations taken into account during the classification. The proposed system is as follows: preprocessing of data, text summarization and text classification. Thus, the Page-Rank-inspired algorithm used

for the summarization gives better results by its precision and its more abstract results. The classification algorithm offers better semantic information and preservation of the context.

[9] developed multi-document summarization system which incorporates the RBM. They have used four different features for feature extraction phase. The feature score of the sentences is applied to the RMB in which the RBM rules are optimized with the help of Deep Learning Algorithm. The experimentation of the proposed text summarization algorithm is carried out by considering three different document sets. The performance judging parameter f-measure has got values, 0.49, 0.469 and 0.520 respectively for the three document sets.

[10] proposed and developed four (4) new implementation models with a variety of datasets for the Encoder-Decoder architecture, in Keras and TensorFlow. The models proposed are the general model, One-Shot model, the recursive model A and the recursive model B. These models use the LSTM architecture [15] which mainly comprises two main components: an encoder and a decoder. This architecture has mainly been developed for natural language processing problems where it has demonstrated peak performance.

[11] presented a new approach to extract documents based on Deep Q-Network (DQN) which is a reinforcement learning method. Authors explored two hierarchical network architectures (RNN-RNN and RNN-RNN) to deploy DQN. Experiments show that their approach achieves higher or comparable performance than the most recent models, with no access to language annotation.

[12] presented SummaRuNNer, a recursive neuronal network (RNN) network sequence model for extracting summaries of documents and showing that it performs superior or comparable to the state of the art. This model consists of two bidirectional GRU RNN layers. This model has been successful in performance by visualizing its predictions.

[13] proposed a single document summary based on a query using an unsupervised deep neural network and a deep automatic encoder (AE) to compute a function space from the term-frequency input (tf) and thus learn features rather than designing them manually.

Lately, [16] have tried to measure the similarities of sentences for text synthesis using deep learning, which will help summarize the text of any language. In all cases, the proposed sentence similarity measure was found to be effective and satisfactory.

V. PROPOSED APPROACH

The proposed summary method uses various features to find the most important phrases that have three major steps: (*A*) preprocessing, (*B*) processing, (*C*) generating summary.

A. Preprocessing

Preprocessing usually involves the identification of the document language, segmentation, tokenization, and stop words removal. Segmentation consists of transforming the input document into a collection of sentences. In tokenization, sentences are divided into words. For the removal of stop words, we created the txt file containing the Malagasy stop words downloadable via [14]. The most commonly used or frequently used words are called stop words. The stop words are worthless and have no importance in the sentences. So, these types of words must be deleted from the input text document, otherwise the sentence having fewer stop words could have a higher score. For example: *'ny'*, *'dia'*, *'hatrany'*, *'no'*, etc.

B. Processing

The processing step is the heart of the text summarization. Once an input document is preprocessed, the document is structured into a sentence-feature matrix. A feature vector is extracted for each sentence. These feature vectors make up the matrix. The following 8 features were combined:

1) Thematic term: The 5 most frequently occurring words of the text are found. These are thematic termss. For each sentence, the ratio of thematic terms number to total words is calculated as:

$$ThematicTerm = \frac{no.of\ thematic\ term\ in\ Fi}{total\ no.of\ sentence\ Fi} \quad (1)$$

2) Sentence position: In Malagasy, the first sentence is still important. Most journalistic texts in Malagasy contain only one paragraph and if there is more than one paragraph, the first sentence is always important. Hence the positioning score of a sentence is calculated so that the first and last sentence of a document gets the highest score. This feature is calculated as follows:
We get 1 score if it's the first sentence in the text. Otherwise:

$$SentencePos = \cos((pos - min)((1/max) - min)) \quad (2)$$

Where pos = position of sentence in the text,
min = th*N,
max = th*2*N.
N is total number of sentences in document
th is threshold calculated as 0.2*N
By this, we get a high feature value towards the beginning and ending of the document, and a progressively decremented value towards the middle.

3) Named entity: Here, we count the total number of named entities in each sentence. Sentences having references to named entities like a company, a group of people, person etc. are often quite important to make any sense of newpaper.

4) Quotation mark: In Malagasy, we consider the quotation mark as a characteristic because we observe that if the text has quotation marks or double quotation marks around it, this indicates a person's direct speech or important information.

$$QuotedWord = \frac{no.of\ numeric\ data\ in\ Fi}{total\ no.of\ sentence\ Fi} \quad (3)$$

5) Numeric Data: Numeric data in Malagasy phrase represents important information concerning date, age, ariary (monetary unit of the Republic of Madagascar), specific numbers … since figures are always crucial to presenting facts; this feature gives importance to sentences having certain figures. For each sentence we calculate the ratio of numerals to total number of words in the sentence.

$$NumericData = \frac{no.of\ numeric\ data\ in\ Fi}{total\ no.of\ sentence\ Fi} \quad (4)$$

6) Presence of 'hoe' and 'izany hoe': In Malagasy, the presence of the word "hoe" has the same importance as the previous section. It is often possible to see the two characteristics appear simultaneously, that is to say in the same sentence. With "*hoe*", the phrase or word after him is always important. On the other hand, with "*izany hoe*", the sentence or the word before him is always important. "*Izany hoe*" means "that means" in English. All words after izany hoe in a sentence are only a long explanation of what is in front.

7) Term Frequency – Inverse Document Frequency (TF-IDF): Since we are working with a single document, we have always taken TF-IDF feature into account. The TF/IDF value of a term is measured by the product of TF and IDF, where TF (term frequency) is the number of times a word occurs in a document and IDF is a reverse document frequency. The IDF of a word is calculated on a corpus using the following formula:

$$IDF=log\ (N/DF) \quad (5)$$

Where N = number of documents in the corpus and df (document frequency) indicates the number of documents in which a word appears.

8) Sentence to centroid similarity: Sentence having the highest TF-IDF score is considered as the centroid sentence. Then, we calculate cosine similarity of each sentence with that centroid sentence.

$$Sentence\ Similarity = cosine\ sim(sentence; centroid) \quad (6)$$

At the end of this phase, we have a sentence-feature matrix. The sentence-feature matrix has been generated with each sentence having 8 feature vector values. After this, recalculation is done on this matrix to enhance and abstract the feature vectors, so as to build complex features out of simple ones. This step improves the quality of the summary.

To enhance and summarize, the sentence feature matrix is input to a Restricted Boltzmann (RBM) machine with a hidden layer and a visible layer. Depending on the size of our documents, a single hidden layer will be sufficient for the learning process.

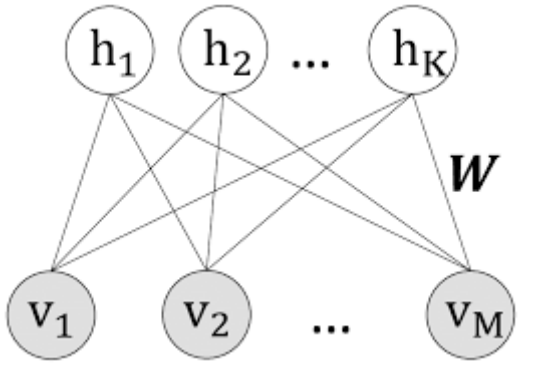

Figure 1. A Restricted Boltzmann Machine

The RBM we use has 8 percepts in each layer with a learning rate of 0.1. We use the persistent contrastive divergence method to perform sampling during the learning process. We have trained the RBM at 5 epochs with a batch size of 4 and 4 parallel Gibbs chains, used for sampling using the persistent CD method. Each sentence feature vector is passed through the hidden layer in which the feature vector values for each sentence are multiplied by the learned weights and a bias value is added to all feature vector values which is also learned by the RBM. At the end, we have an improved matrix. Note that the RBM should be formed for each new document to be summarized. The idea is that no document can be summarized without reviewing it.

C. Generating summary

Enhanced feature vector values are summed to generate a score for each sentence. The sentences are then sorted according to the decreasing value of the score. The most relevant sentence is the first sentence of this sorted list and is chosen from the subset of sentences that will form the summary. Then, the next sentence we select is the one that has the greatest similarity in Jaccard's measure with the first sentence, strictly selected in the upper half of the sorted list. This process is repeated recursively and incrementally to select more phrases until a user specified summary limit is reached. The sentences are then rearranged in the order of appearance in the original text. This produces a coherent summary.

VI. EXPERIMENTATION

In this section, we discuss various experiments conducted to generate a summary of the Malagasy text. For the implementation of the approach described above, Python 3.5 (programming language) and Pycharm IDE 2018.3.4 (platform) were used.

A. Corpus

The corpus, named "TONTONA", was created in 2012. It's a corpus of Malagasy texts in the field of the environment from different authors of the field.

Three criteria summarize the constitution of the corpus: linguistic, extra linguistic and terminological criteria. The corpus was organized according to varieties observed upstream. Other physical criteria related to the electronic format of documents have also been taken into consideration. The source files are classified in a directory while the txt files containing raw texts necessary for the exploration are in another directory. The physical organization of the corpus is summarized in the table below

TABLE I. DESCRIPTION OF TONTONA CORPUS

Corpus	TONTONA			Total
	Scientist	*Judicial*	*Journalistic*	*3 under corpus*
Texts	56	9	15	80
Authors	33	9	11	52
Words	178324	384213	8812	571349

B. Evaluation Metrics

The evaluation of the proposed text summarization method is based three basic evaluation criteria. The different criteria are listed below.

1) Recall: Recall is the ratio of number of extracted sentence to the number of relevant sentence. The recall is used to measure the reliability of the proposed text summarization method:

$$Recall = \frac{S_{ext} - S_{rel}}{S_{ext}} \quad (7)$$

where, S_{ext} and S_{rel} are the number of extracted and relevant sentences respectively. Relevant sentences are the summaries produced by human.

2) Precision: The ratio of retrieved sentences to relevant sentences based on the extracted sentences is given as the precision measure:

$$Precision = \frac{S_{ext} - S_{rel}}{S_{rel}} \quad (8)$$

3) F-Measure: The precision values and the recall values are considered for finding the F-measure value for the total dataset. Thus the F-measure can be expressed as:

$$F\text{-}Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (9)$$

C. Performance evaluation

Six documents from each under corpus with varying number of sentences were used for experimentation and evaluation. The proposed algorithm was run on each of those. The performance results obtained are shown below in figure 2, 3 and 4.

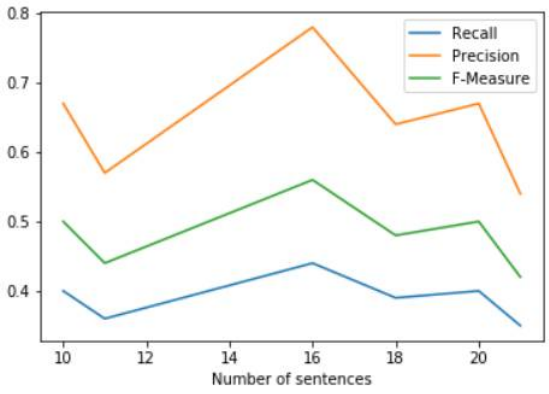

Figure 2. Performance of scientist under corpus

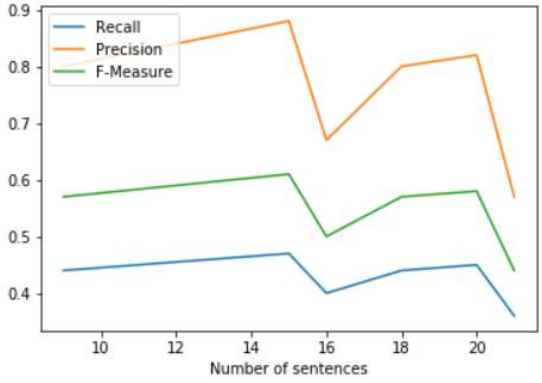

Figure 3. Performance of judicial under corpus

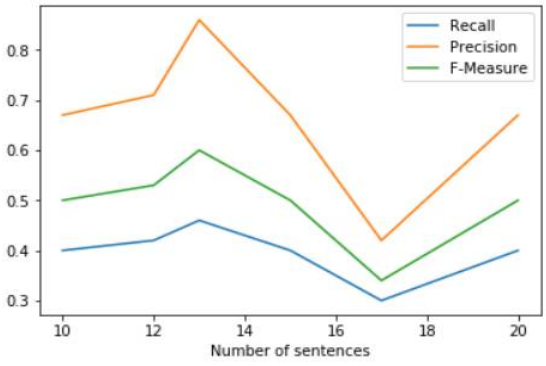

Figure 4. Performance of journalistic under corpus

It can be seen that as the number of sentences in the original document cross a certain threshold, the Restricted Boltzmann Machine has ample data to be trained successfully and summaries with high precision and recall values are generated.

TABLE II. RESULT OF EXPERIMENTATION

Under corpus	**Average result**			
	Domain	***Recall***	***Precision***	***F-Measure***
1	Scientist	0,39	0,64	0,49
2	Judicial	0,43	0,76	0,55
3	Journalistic	0,40	0,66	0,50
Average		0,40	0,69	0,52

The summary is generated with the help of the proposed text summarization algorithm. The maximum recall value obtained for the judicial domain is 0.47. Similarly the maximum precision value obtained is 0.88. The f-measure value is calculated according to the recall and precision value. The maximum value obtained for the f-measure is 0.61. During the experimentation, F-Measure is ranged from 0,35 to 0,61 and the journalistic under corpus had this lowest F-Measure.

D. Comparative analysis

As we mentioned in the beginning of the paper, recently, there is one research work of summarizing Malagasy text that we performed too. It used statistical based method, our first system that is judged simple. The existing approach was run for the same document set in the same corpus. It has been evaluated by Accuracy in percentage. The average result of the experiments indicates a precision of 77% for a compression ratio of 25% and 80% for a compression ratio of 50% with a human evaluation by judgment of the obtained summary of 57.6%. The approach proposed in this paper, which compares the summaries obtained by the system and the reference summaries, obtained a human evaluation by judgment of 65%.

VII. CONCLUSION

An approach using deep learning for the Malagasy text summarization has been proposed in this paper. It incorporates the Restricted Boltzmann Machine algorithm and can extract one or more documents. The algorithm is executed separately for each input document, instead of learning the rules of a corpus, because each document is unique in itself. We extract 8 features of the given document and improve them to mark each sentence. Our approach uses one RBM and works effectively. This has been demonstrated by selecting documents from 3 different domains and comparing the summaries generated by the system to those written by humans. The proposed system performance is also a start for the Malagasy text summary and can be improved by adapting features according to the needs of the user and further adjusting the RBM parameters to minimize processing and errors in the coded values. . One can also use 2 RBMs stacked on top of each other for feature enhancement.

REFERENCES

[1] M. Jaozandry, "Les prédicats nominaux du Malgache: étude comparative avec le français, " Doctoral dissertation, Paris 13, 2014.

[2] T. H. Ravelonjatovo, "Traitement informatique de variations terminologiques malgaches. Cas du domaine de l'environnement," Western Papers in Linguistics/Cahiers linguistiques de Western, 2015, vol. 1, no 1, pp. 19.

[3] A. J. Ratovondrahona, J. L. Razafindramintsa, H. F. Rafalimanana, T. Mahatody and V. Manantsoa, "Word Embedding: Machine translation according to the context," In 8th International Conference Sciences of Electronics, Technologies of Information and Telecommunications, 2018.

[4] R. Hanitramalala, "Vers une typologie des collocations à verbe support en malgache, " 2018.

[5] R. B. Rabenilaina and J. Morin, "Vitasoa,"Ambozontany Editions, pp. 11-28, 2015.

[6] H. P. Luhn, "The automatic creation of literature abstracts," IBM Journal of research and development, vol. 2, no. 2, pp. 159–165, 1958.

[7] M. Y. Day and C. Y. Chen, C. Y, "Artificial intelligence for automatic text summarization," In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 478-484. July 2018.IEEE.

[8] K. Sharma, A. Gaikwad, S. Patil, P. Kumar, D. P. Salapurkar, "Automated Document Summarization and Classification Using Deep Learning," In: International Research Journal of Engineering and Technology, 5(6), 2018.

[9] G. PadmaPriya and K. Duraiswamy, "An approach for text summarization using deep learning algorithm," Journal of Computer Science 10 (1): pp. 1-9, 2014

[10] M. Patel, A. Chokshi, S. Vyas, K. Maurya, "Machine Learning Approach for Automatic Text Summarization using Neural Networks," In: International Journal of Advanced Research in Computer and Communication Engineering, 7(1), 2018.

[11] K. Yao, L. Zhang, T. Luo, Y. Wu, "Deep reinforcement learning for extractive document summarization," *Neurocomputing*, *284*, pp. 52-62 2018.

[12] R. Nallapati, F. Zhai, B. Zhou, "Summarunner : A recurrent neural network based sequence model for extractive summarization of documents," In *Proceedings of the Thirty-FirstAAAIConference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pp. 3075–3081, 2017.

[13] M. Yousefi-Azar and L. Hamey, "Text summarization using unsupervised deep learning," *Expert Systems with Applications*, *68*, pp. 93-105, 2017.

[14] https://uptobox.com/xmv8x2ro1zon

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, *9*(8), pp. 1735-1780, 1997.

[16] S. Abujar, M. Hasan, S. A. Hossain, "Sentence Similarity Estimation for Text Summarization using Deep Learning," In *Proceedings of the 2nd International Conference on Data Engineering and Communication Technology* , pp. 155-164, Springer, Singapore, 2019