DOI: 10.5748/18CONTECSI/PSE/DSC/6709

MULTIWORDS EXPRESSIONS IDENTIFICATION THROUGH RECURRENT NEURAL NETWORKS

Edson Marchetti Da Silva - CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS - Orcid: https://orcid.org/0000-0003-4801-0892

Renato Rocha Souza - FUNDAÇÃO GETÚLIO VARGAS-RJ - Orcid: https://orcid.org/0000-0002-1895-3905

Propose an alternative method for to identify Multiwords Expressions extracted from documents through Recurrent Neural Network (RNN) model Multiword Expressions could be used to represent the meaning of the document in the more semantic way, in many activities of the natural language processing. We found few researches that use RNN as a method to obtain MWE. It was the definition of the training corpus created within lines classified as bigrams or not, extracted from documents through traditional statistics methods; It was to train the model based on the corpus created; It was to validate the results obtained aim to generalize the process through RNN method. As results we obtained an accuracy around 80,54% in test train to identify bigram in the new documents. We propose the use of machine learning algorithm to generalized the extraction of bigrams in the documents in the specific domain. The idea is to do the information retrieval process to obtain searched documents through of a document rather than using keywords.

Keywords: Extraction , Multiwords, Recurrent, Neural , Networks, bigrams

MULTIWORDS EXPRESSIONS IDENTIFICATION THROUGH RECURRENT NEURAL NETWORKS

Abstract

This paper aims to propose an alternative method for to identify Multiwords Expressions (MWE) extracted from documents. In this sense, first of all we trained a Recurrent Neural Network (RNN) model, supplying a dataset compound by 186 documents that were previosly preprocessing to identify MWE through of the conventional statistical algorithms that consider the frequency of co-ocorrence of the *n*-grams. Thus, the data set was created with *n*-grams plus frequency information as the independent variables plus the target variables with the two possibilities is or not, a bigram. The idea is to create a learning model that goals generalized the identification MWE using as pattern recognize the behaviour of the frequency for each *n*-gram and of the frequency of when the *n*-grams to be related in a bigram. Then, this model could be used for a tool to identify MWE in documents to presented by the final users. As results we obtained an accuracy around 80,54% in test train.

Keywords: Extraction of Expressions Multiwords, Recurrent Neural Network, bigram.

1. Introduction

One of the biggest chalenges that the cientific community has been research since que 50s of the last century is to reproduce the skills of the human brain in the machines. In other words, making the artificial intelligence a reality. In this context, the one of most complex tasks is develop the ability of machines to interpret the natural language. Thus the Natural Language Processing (NLP) is highlighted which through studies of morphology, syntax and semantic analysis, and statistical processing were designed to predict behavior of a textual content. One way of dealing with understanding the meaning of the text is to understand its parts. In this direction, according Klyueva, Doucet and Straka (2017, p. 60) "Multi-Word Expressions (MWEs) present groups of words in which the meaning of the whole is not derived from the meaning of its parts. The task of processing multiword expressions is crucial in many NLP areas, such as machine translation, terminology extraction etc."

Many works that aim to identify MWE have been successful in the past, using statistical models, such as: Dias, Lopes e Guilloré (1999) which aims identification de MWE independently from language; Silva, Lopes (1999) which aims to obtain *n*-grams from the analysis of the text in a local context called LocalMaxs; Portela, Mamede e Batista (2011) which takes into account the morpho-syntactic characteristics of the text; Silva e Souza (2012) that consider some structure of the text and many others. However, according Klyueva, Doucet and Straka (2017), recently deep learning have been applied to a vast majority of NLP tasks, mainly algorithms based on Recorrent Neural Network (RNN) that address the identification MWE task. In the same direction, this work to use RNN to identify MWE in texts.

To better describe the experiments the work is structured in sections in which are presented in the following contents: Section 2 – theoretical framework about MWE and RNN; Section 3 – Related works; Section 4 – methodology; Section 5 - Results and conclusions; Section 6 - Recommendations for future work.

2 Theoretical Framework

This section provides a brief review of the MWE and RNN concepts.

2.1 Multiword Expression

According to Sarmento (2006), the text is not a simple random jumble of words. The order of placement of words in the text is what produces the meaning. Therefore, the study of the co-occurrence of words brings important information. This may indicate that words are related directly by compositionality or affinity or indirectly by similarity.

Along the same lines, Zhang *et al*. (2009), states that the ability to express the meaning of a word depends on the other words that accompany it. When a word appears accompanied by a set of terms, the greater the chances of that set having a relevant meaning. This indicates that not only the word, but also contextual information is useful for processing information. It is based on this simple and direct idea that research on MWE is motivated. In this way, it is expected to capture relevant semantic concepts of the text expressed by MWE. The task of processing multiword expressions is crucial in many NLP areas, such as machine translation, terminology extraction etc.

Although there are many papers on the subject, there is no formal definition of consensus in the literature on MWE. We can consider that MWE are formations composed of two or more adjacents words that occurring in a frequency above a threshold when combined it have a greater semantic expressiveness than when each of its terms are set separately. For Sag *et al*. (2002, p. 2) MWE are "idiosyncratic interpretations that cross the boundaries (or spaces) between words". A further description found in the literature is shown below.

The work Cazolari *et al*. (2002) uses a focused approach in MWE that is productive on the one hand and, on the other shows that regularities that can be generalized to classes of words with similar properties. In particular they seek to find grammatical devices that allow the identification of new MWE motivated by the desire for recognition as possible in the automated acquisition of MWE. In this sense, the research of these authors studied in depth two types of MWE: support verbs and compound nouns (or nominal complex). For according to them these two types of MWE are at the center of the spectrum of compositional variation where the internal cohesion together with a high degree of variability in lexicalization and language-dependent variation can be observed.

The approach used by Evert and Krenn (2005) is based on the calculus of statistical measures of association of the words contained in the text. In empirical tests, these authors used a subset of eight million words extracted from a corpus consisting of a newspaper written in German. The proposed approach was divided into three steps. In the first extracts the tuples from the corpus source contain Lexical pronouns (P), nouns (N) and verbs (V). These data are grouped in pairs (N + P, V) and placed in a contingency table, represented by a three-dimensional structure, where each pair is disposed in a plane P + N V and the third axis is assigned to the frequency information represented by four cells. Thus a comparison is made between all pairs extracted from the lexical corpus with their sentences, accounting for each sentence, one of four possibilities: there are PN and V; there is PS, there is not V; there is not PS and there is V; there are not PS and V. That is, one unit is added whenever one of the possibilities occurs. The second step the association measures are applied to the frequencies collected in the previous step. This process results in a list of pairs of MWE candidates with their association scores calculated and ordered from the most strongly associated to

the less strongly associated. The "n" top candidates on the list are selected for use in the next step. The third step is the evaluation of the list of MWE generated by a human expert. Thus, the approach proposed by these authors is characterized by an extraction of semi-automatic MWE. In order to minimize the intellectual work of an expert, these authors propose the use of a technique of extracting a random sample, representative of the corpus rather than the complete set of documents.

Research conducted by Villavicencio *et al*. (2010) seeks to extract the MWE combining two different approaches: the approach based on associative lexical alignment. At first, the association measures are applied to all bigrams and trigrams generated from the corpus and the result of these measures is used for evaluation. The second approach draws MWE in an automated way based on the alignments of lexical versions of the same content written in Portuguese and English. To combine the results obtained, the authors used two approaches to Bayesian networks. The multilingual perspective, often can not find a direct lexical equivalence; generalization of lexical difficulty (general and terminology) to a specific context.

The statistical approach for the extraction of MWE through the co-occurrence of words in texts has been used in several works, among them: Pearce (2002); Kreen and Evert (2005); Pecina (2006); Ramisch (2009) and Villavicencio et al. (2010). These studies use various statistical techniques that seek to identify MWE as a set of adjacent words that co-occur with a frequency greater than expected in a random sequence of words in a corpus. Thus the associative approach is nothing more than the use of a set of association measures that aim to identify the candidate expressions for MWE. Among the techniques used include: coefficient of Pearson Chi Square; Dice coefficient; Pointwise Mutual Information – PMI; Poisson Stirling among others. For the task of automatic detection of multiword expression researchers use language-independent approaches that combine association measures like mutual information or dice coefficient with machine learning approaches (Tsvetkov and Wintner, 2011), (Pecina, 2008).

On the other hand, neural networks were exploited in a number of papers for the task very related to ours, e.g. Zampieri *et al*. 2018 that presents a technique for identifying a more specific type of MWE, the verbal MWE (VMWE) which is obtained through RNN. VMWEs are classified by combining the VMWE category with Begin Inside-Outside (BIO) tags - a variant of the standard coding scheme. The purpose of the RNN is to predict the correct BIO + category tag for each token.

2.2 Recurrent Neural Network

The ability to interpret the words of a text depends on the interpretation of the context. Therefore, this scenario requires the use of algorithms that treat the independent variables of the learning model as interrelated by sequence. The adoption of a recurrent network aims to map this context. After all, a decision reached in time step t -1 will affect the decision that will be taken later in time step t. Thus, recurrent networks have two sources of input, the present and the recent past, which combine to determine how they respond to new data, just as we do.

According Mikolov *et al*. (2010) sequential data prediction is considered by many as a key problem in machine learning and artificial intelligence. It to deal with sequential data prediction problem is the goal when constructing language models. In dispite of still, many attempts to obtain such statistical models involve approaches that are very specific for language domain, for example, assumption that natural language sentences can be described by parse trees, or that we need to consider morphology of

words, syntax and semantics, in this work we used RNN to for modeling sequential data. The network has an input layer *x*, hidden layer *s* (also called context layer or state) and output layer *y*. Input to the network in time *t* is *x*(*t*), output is denoted as *y*(*t*), and *s*(*t*) is state of the network. Input vector *x*(*t*) is formed by concatenating vector *w* representing words of the each document, and output from neurons in context layer *s* at time $t-1$. Input, hidden and output layers are then computed as follows:

$$x(t) = w(t) + s(t-1) \quad (1)$$

$$S_j(t) = f\left(\sum_i x_i\,(t) u_{ji}\right) \quad (2)$$

$$y_k(t) = g\left(\sum_i s_j\,(t) u_{kj}\right) \quad (3)$$

where $f(z)$ is sigmoid activation function:

$$f(z) = \frac{1}{1+e^{-z}} \quad (4)$$

and $g(z)$ is softmax function:

$$g(z_m) = \frac{e^{zm}}{\sum_k e^{zk}} \quad (5)$$

Networks are trained in several epochs, in which all data from training corpus are sequentially presented. Weights are initialized to small values. The training the network, use the backpropagation algorithm with stochastic gradient descent. Starting learning rate is $\alpha = 0.1$. After each epoch, the network is tested on validation data. If log-likelihood of validation data increases, training continues in new epoch. If no significant improvement is observed, learning rate α is halved at start of each new epoch. After there is again no significant improvement, training is finished. Convergence is usually achieved after 10-20 epochs.

According Zaremba, Sutskever and Vinyals (2015) Recurrent Neural Networks (RNNs) are effective models of natural language that substantially improve over long established state-of-the-art baselines have been obtained using RNNs. as well as in various conditional language modeling tasks such as machine translation.

3 - Related Works

In the paper of Klyueva, Doucet and Straka (2017) the authors describes the MUMULS system that to perform automatic identification of verbal multiword expressions (VMWEs) in 15 languages. The system was implemented using a supervised approach based on RNN using the open source library TensorFlow. The model was trained on a data set containing annotated VMWEs as well as morphological and syntactic information.

Varis and Klyeuva (2018), describe ongoing work on improving tagger, MUMULS, using the current state-of-the-art sequence-to-sequence techniques applied in other NLP tasks, including different styles of embedding of the input tokens, creating a context-aware feature representation of the input sequence and generating of the target labels

Erden (2015) in your study examine the impact of data representation format on the VMWE identification task. The results show that data representation format is

important to identify discontinuous VMWEs. So, the author introduce the bigappy-unicrossy tagging scheme in order to recognize overlaps in sequence labelling tasks. Moreover, we enhance our neural VMWE identification model with automatically learned embeddings by neural networks to rise to the variability challenge in VMWE identification. The work compares character-level convolutional neural networks and character-level bidirectional long short-term (BiLSTM) networks, and analyze two different schemes to represent morphological information using BiLSTM networks. The results demonstrate that character embeddings and morphological embeddings improve performance in general. The choice of representation learning method depends on language.

4 Methodology

The purpose of this study is to present a form of the automatic identification of MWE in a documents through of a tool that use a model of RNN.

To perform the experiment, three main tasks were implemented: the first one was the corpus assembly, the second was train the model and the third was to validate and analyze the results obtained. These tasks have been subdivided into several steps, which are described in detail in the next section.

5 Experiment steps and results

This section describes the step by step to achieve the objective of this experiment.

5.1 Corpus assembly

The first step was converting the corpus compound of 186 papers totalizing 47,4 megabytes in PDF format - typically with around 20 pages each one, totaling 644,858 normalized tokens - in a list of 7,970 distincts terms encoded in ASCII.

To perform the conversion into text format, the TET PDF software was used. This software consists of a Dynamic Link Library (DLL) that was coupled in software components developed in C++ by the authors. The process of converting the PDF document page by page was performed in order to identify the header of the pages.

5.2 Preliminary filtering of the contents of the documents

After transforming the document pdf into text, preliminary filtering was performed in order to remove parts of the contents considered as noise. The adopted heuristic evaluates the content that occurs repeatedly in all the pages from the top of each page. This extract, called header, is filtered and therefore eliminated in the converted text. Another filtering process that is performed at this stage the removal of the references, to include terms such as: name of authors and works, which often lie outside the central theme of the document.

After the elimination of the parts considered as noise, the goal in this substep is to perform parsing, ie, to process the string extracted from the document and separate the block into tokens.

The process of separating the text into sentences and words to create the vocabulary words is known as tokenization. Manning, Raghavan & Schütze (2009, p. 22-26) define tokenization as the task of receiving as input a given sequence of characters in a document and split it into parts called tokens, while discarding those characters that indicate the points of separation.

After the text is broken up into sentences, they must be broken into words in order to become or not a term in the vocabulary. The characters usually used to indicate the separation of the words are the comma, the hyphen and blank space. But they can not be considered as separators on an unrestricted basis. For example, the comma can be used to separate whole numbers from decimals in the European model of numerical representation, or the thousands in the Saxon model; the hyphen may be used to divide syllables of a word at the end of a line, or compounds that can be found in different spellings, in the case of the blank space, the problem occurs when it is used to separate the names, in which case the terms should not be separated because they made a semantic sense.

To mitigate these problems we used some strategies described below. In the case of the comma it is discarded, so the numerical representations are expressed only by numbers without separators. In the case of the hyphen in the Portuguese language such as: “infraestrutura[1]”, “infra-estrutura[2]” or “infra estrutura[3]”; by making a Google search for three terms two different results are found. When searching for “infra-estrutura” or “infra estrutura”, approximately 6,880,000, links were found while “infraestrutura” found approximately 39,100,000 responses. Therefore this is still an open question. In this study we will ignore the hyphen, thus, words spelled with a hyphen will be treated as a single word, and syllable breaks of the dash, when removed, will regroup the word. In the case of the blank space, the problem is found in the contents with proper nouns, such as “New York”, because the semantic meaning in this case must be made by the two words together, not as two separate entries in the vocabulary. In this work, this problem becomes irrelevant, because if these words are relevant in the context of the document, they will become a bigram and will be found only if the sequence in the document collection. Therefore, during the process of converting a text, a treatment was carried out byte by byte characters where the following tasks were performed: (1) To identify and convert all accented characters, which are represented by multibyte characters, the usual Portuguese, transforming them into non-accented characters while preserving the original spelling of the text of uppercase and lowercase letters, (2) to remove the hyphens, (3) to remove the dot (.) that is used to abbreviate words, (4) to remove the periods (.) and commas (,) used as separators of numbers, (5) to remove expressions such as "[...]" "(...)"; (6) Delete all ASCII bytes whose value is less than 1 or greater than 126. All these steps were performed in order to minimize the error parser separation of sentences.

Thus, the rules adopted to consider the existence of a delimiter sentence were: (1) If you find any of the following characters: question mark, exclamation point, (2) If after the (.) period there is a line breaking character, a new paragraph, end of a text or a capital letter. All characters used as separators are eliminated from sentences.

[1] Under the new Portuguese orthographic agreement in effect as of 2009.
[2] Spelled before the agreement.
[3] Spelled incorrectly, but that could be found.

5.3 Decoding Acronyms

A very common practice of writing, especially in science, is the use of abbreviations. Typically, the first appearance terms are shown in full with the letters that make up the acronym in each term presented in uppercase followed by the acronym itself with capital letters separated or not by a period between brackets. From this premise, in this sub step the goal is to identify acronyms in order to build a table of acronyms used in each document, and add to the part in full text whenever when the acronym occurs. This strategy is important to be adopted, since the content expressed in the text only as an acronym would not be interpreted as MWE. While in fact this kind of content is usually high in semantic content to express the meaning of the document, and when it is placed in full, depending on its frequency of occurrence, it makes this set of terms become MWE.

5.4 Segmentation sentences into terms

In this sub step, the goal is to separate the sentences into terms in order to create the vocabulary of terms. Tokens, ie, the pieces that were targeted, normally go through a standardization process before they become a term of the vocabulary. Normalization aims to reduce the number of dictionary entries. In this sense all words are transformed into lowercase.

5.5 Secondary Filtering – Stop Words

In this substep, after breaking the documents into a word sequence a new filter is executed. The goal is to remove the vocabulary words that appear very frequently in all documents and, therefore, have little power of discrimination. Manning, Raghavan & Schütze (2009, p. 27) defined stop words as common words that seem to have little value to select the corresponding documents. These words usually belong to the class of articles, prepositions and some conjunctions.

5.6 Train the RNN model

After pre-processing all the documents, and putting each one in a text file, we started using the Python language to handle this data. Then the first step was to prepare the dataset for trainning the deep learning model. After each text file were preproced the dataset had 644,858 tokens, then the next step were identified the MWE though statistical algorithm of the Nltk library. It was seleted only the bigrams that have the frequency greather than two. So, the resultant dataset was compound by 7,306 distinct bigrams, each one with the frequency obtained in all documents in the corpus. The resultant dataset was created with three independent variables, being each one of the *n*-grams plus the frequency that the bigram have in the document. Furthermore one target variable with two possible categories: “is a bigram” or “isn´t a bigram” as dataset of supervisoned model.

After to insert all bigrams in the dataset, for to have the corresponding quantity of instances at oposite classe “isn´t a bigram”, it was necessary to insert in the dataset the same quantity of instances with distinct words forming the new "isn't bigrams" to balanced the dataset with 50% of the instances of each classe. So, the dataset to became with 14,612 instances with both target classes. However, for put the dataset in the correct format was necessary to convert the two first categorical attributes in a embedding vector. Then it was created a dictionary with 3,580 distinct *n*-grams.

With the dataset ready, it was necessary to divide the dataset in train and test. We used 25% of the size to compound the test data. As the algorithm for to trainning the model, we use the Pytorch library. As parameters we use the learning rate igual 0,01 and 30 epoches of interations. After model trained, the goals of this work is, from a new document the model going to capable to identify the MWE. We use as the threshold, the frequency greater than two coocorrency in the document to select the bigram.

6 RESULTS AND CONCLUSIONS

With the model trained, we to perform the test dataset and obtained an accuracy of 80,58%, a precision of 80,56% and revocation of 100%. After that we use the model to identify other diferents papers and to show the word cloud of bigrams founded. The Figure 1 shows an example.

Figure 1 – word cloud with the bigrams.

Source: by the authors.

7 Recommentations for future work

The result obtained by RNN model it showed good responses if consider that the dataset size. Actually, this experiment it was an essay as concept proof. As future works we recommend using a bigger dataset for to create a greater bigrams dictionary to supply to the model more data to improve the learning and to become the model more generalized.

References

CALZOLARI, Nicoletta FILLMORE, Charles J.; GRISHMAN, Ralph, IDE. Nancy;

DIAS, Gaël ; LOPES, José Gabriel Pereira ; GUILLORÉ, Sylvie. Mutual expectation: a measure for multiword lexical unit extraction. In Proceedings of Vextal, 1999.

ERDEN, Berna. Identification of Verbal Multiword Expressions Using Deep Learning Architectures and Representation Learning Methods. Master Dissertation in Computer Engineering, Bogaziçi University, 2015.

EVERT, Stefan ; KREEN, Brigitte. 2005. Using small random samples for the manual evaluation of statistical association measures. Computer Speech and Language, 19(4):450–466.

KLYUEVA, Natalia, DOUCET, Antoine and STRAKA, Milan. Neural Networks for Multi-Word Expression Detection. Proceedings of the 13th Workshop on Multiword Expressions MWE, 2017.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich An introduction to information retrieval. Ed. Cambridge online, 2009.

MIKOLOV, Tomás; KARAFIÁT, Martin.; BURGET, Lukás; CERNOCKÝ, Jan Honza. and KHUDANPUR, Sanjeev. Proceedings of the 11th Annual Conference of the International Speech Communication Association , page 1045--1048. ISCA, 2010.

PEARCE, Darren. A comparative evaluation of collocation extraction techniques. Em of the Third (LREC 2002), Las Palmas, Canary Islands, Spain, May, 2002.

PECINA, Pavel ; SCHLESINGER, Pavel. Combining Association Measures for Collocation Extraction. In ACL'06, page 652, 2006.

PORTELA, Ricardo Jorge Rosa ; MAMEDE Nuno ; BATISTA, Jorge. Multiword Identificação. In Terceiro Simpósio de Informáctica Portugal pp. 110-199, 2011.

RANCHOLD, Elisabete M. O lugar das expressões 'fixas' na gramática do Português. in Castro, I. and I. Duarte (eds.), Razão e Emoção, vol. II, Lisbon: INCM, pp. 239-254, 2003.

RAMISCH, Carlos. Multiword terminology extraction for domain specific documents. Dissertação – Mathématiques Appliqueées, École Nationale Supérieure d'Informatiques, Grenoble, 2009.

SAG, Ivan A. ; BALDWIN, Thimothy ; BOND, Francis ; COPESTAKE, Ann ; FLICKINGER, Dan. Multiword expression: a pain in the neck for nlp. Em Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing CICLing-2002), volume 2276 of (Lecture Notes in Computer Science), pp. 1–15, London, UK. Springer-Verlag.

SARMENTO, Luís. Simpósio Doutoral Linguateca 2006. Disponível em: http://www.linguateca.pt/documentos/SimposioDoutoral2005.html: out. 2011

SILVA, Ferreira J. LOPES Pereira G. A local maxima method and fair dispersion normalization for extracting multi-word units from corpora. (1999). Sixth meeting on Mathematics of Language, pp. 369-381.

SILVA, Edson Marchetti; SOUZA, Renato Rocha. Information retrieval system using multiwords expressions (MWE) as descriptors. JISTEM - Journal of Information Systems and Technology Management Vol. 9, No. 2, Mai/Aug. 2012, pp.213-234.

VARIS, Dusan, KLYUEVA, Natalia Improving a Neural-based Tagger for Multiword Expression Identification. 2018

VILLAVICENCIO, Aline ; RAMISCH, Carlos; MACHADO, André; CASELI, Helena de Medeiros; FINATTO, Maria José. Identificação de expressões multipalavra em domínios específicos. Linguamática, v. 2, n. 1, p. 15-33, abril, 2010.

ZAMPIERI, Nicolas; SCHOLIVET, Manon; RAMISCH, Carlos; FAVRE, Benoit. Veyn at PARSEME Shared Task 2018: Recurrent Neural Networks for VMWE Identification. Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 290–296 Santa Fe, New Mexico, USA, 2018.

ZAREMBA, Wojciech; SUTSKEVER, Ilya; VINYALS, Oriol. Recurrent neural network regularization. In Proc. ICLR. 2015.

ZHANG, Wen; YOSHIDA, Taketoshi; TANG, Xijin; HO, Tu-baq. Improving effectiveness of mutual information for substantival multiword expression extraction. Expert Systems with Applications, Elsivier, v. 36, 2009.