P- ISSN 1991-8941 E-ISSN 2706-6703 *Journal of University of Anbar for Pure Science (JUAPS)* Open Access
2019,13 (1) :50-54

A Technique for Discovering Similarities between Texts Based on Extracting Features from the Text

Alaa Abdalqahar Jihad [1*] , Mortadha M. Hamad [2]

[(1)] Computer Center, University of Anbar, it.alaa.heety@uoanbar.edu.iq.
[(2)] College of Computer Sciences and IT , University of Anbar, mortadha61@yahoo.com.

ARTICLE INFO

Received: 22 / 4 /2019
Accepted: 29 / 8 /2019
Available online: 19/7/2022

DOI: 10.37652/juaps.2022.171876

Keywords:
Similarity.
Text Processing.
Pattern Recognition.
Extraction.
Semantic Textual Similarity (STS).
Natural Language Processing (NLP).

ABSTRACT

The discovery of the similarity between two texts is very important and useful in many applications. The similarity between texts is the core research area of dataset, data warehouse, and data mining. This paper provides a framework that gives a similarity between two input texts based on pattern recognition and the use of approximate string matching; there is a weight that affects the proportion of similarity. The search compares the similarity of two texts without adherence to the grammar or the use of synonyms or meanings of words. Preliminary results showed the benefit of extracting some of the features in the discovery of the similarity between the texts.

Introduction:

Text classification is a key technique in text mining, which indicates the task of assigning text documents to one or more pre-defined categories. This is a direct concept of automated learning, which means advertising a group of labeled categories as a method of document representation and a statistical classification of a coach with a training group[1].

For a long time, the text similarity measures are used in natural language processing (NLP) applications [2]. Many algorithms exist to determine the similarities between texts.

The problem of text analysis has been discussed intensively in the last few years in various fields of application related to the text such as: retrieval of information, text classification, subject tracking, document clustering, question generation, answering questions, recording short answers, and topic detection[3].

* Corresponding author at: Department of Chemistry, Ibn-Al-Haithem College of Education for pure science. University of Baghdad, Iraq.E-mail address: *it.alaa.heety@uoanbar.edu.iq* ,

Large feature sets are a challenge in text classification problems that need to be addressed for better performance [4]. This paper proposes a framework for comparing two texts without using synonyms or grammar.

Related Works:

Many of the research focused on the discovery of similarities between the texts.

Atish Pawar and Vijay Mago [9], presented a methodology dealing with the integration of semantic similarity and corpus statistics to calculate the semantic similarity among words and sentences, pursue the edge-based methodology using a lexical database and test procedure on both standard and human similarity data sets.

Yue Wang et al [10], proposed a way to compute the semantic similarity between sentences based on a hybrid approach. This approach uses Bidirectional Long Short-Term Memory Networks (BLSTM) and Convolutional Neural Networks (CNN) to extract the semantic features of the text, learn to represent each sentence with attention to the level of the word, collect the vigilant representation and feed it into the output layer to calculate similar sentence scores.

Suhad M. and Aseel Q. [7]. This work presents many methods that are used in the similarity of texts. The proposed system finds the similarity between two Arabic texts using hybrid similarity measurement techniques: Semantic similarity measure, Cosine similarity measure and N-gram. They designed an Arabic SemanticNet that stores keywords for a particular field (such as computer science), and through this SemanticNet it is possible to find similarities between words according to specific equations. They use the Cosine and N-gram similarity measures to find similar character sequences.

Aminul Islam and Diana Inkpen [5]. They provide an approach for measuring the semantic similarity of texts by a corpus-based measure for the word similar to a semantic and a normalized and edit version of the Longest Common Subsequence (LCS) text matching algorithm. They handle the similarities among two sentences or two short texts.

Some Techniques Which Exist to Discover Similarities:

There are many techniques that work to detect the similarities and approximate matching between the texts, such as:

Q-grams: utilized in approximate text matching by sliding a window of length q over the characters of a text to make a various 'q' length grams for coordinating a match is then evaluated as a number of q-gram matches inside the second text over conceivable q-grams [12].

Cosine Similarity: is a typical vector which is based on similarity measure similar to dice is coefficient where the input text is converted into vector space so that the Euclidean cosine principle can be utilized to decide similarity.

Dice Coefficient: a similarity scale is based on the measure (0-1) where the similarity measure is characterized as twice the number of terms common to compared entity's divided by the total number of terms in both tried entities[13].

Challenges of Discovering:

Calculating similarities between texts that have been written in one language or multiple languages form one of the most important difficulties confronting the natural language processing. There are many of the challenges and problems which face the process of discovering similar texts. We give in this paragraph some of them, such as:

- Two texts written electronically, and between them are small and simple adjustments and need to focus for the discovery .
- Spelling mistakes.
- Synonymy of words.
- Submission and delay in sentences and words.
- Use of punctuation.

Applications of Text :

There are several fields that can play an important role directly or indirectly such as:

- Sentiment analysis.
- Natural language understanding.
- Machine translation.
- Summarize the document.
- Categorize short answers.
- Retrieve and extract information.
- Web Search.
- Correcting the answers to the essay questions.
- Remove Duplicate Records in Dataset.

Proposed Work :

The proposed work in general is illustrated in Fig (1), where the original text and the secondary text are entered in preprocessing, and then extracting the features of the texts. After extracting the features they are compared and then given a ratio to each feature.

The following paragraph describes the preprocessing steps**.**

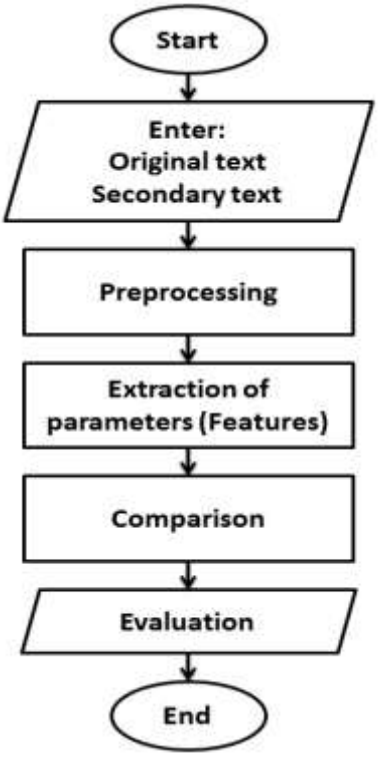

Fig. 1: The proposed work.

Algorithm (1): Preprocessing steps

Input: Raw text
Output: Filtered text

1. Remove all the stop words.
2. Convert letters to lowercase.
3. Remove every letter or a group of letters in the end of words, such as (ed, ing, tion, sion…).
4. Remove the completely different words.

The following are features that extracted from the text, which we suggested in the work, whose impact is examined in text comparison:

- Number of Words.
- Number of Characters.
- Number of Letters.
- Number of Match Words: Match words 100%.
- Number of Different Words.
- Number of Match Words and Match in Location: Locations of match words 100%.
- Number of Different Letters:
- Number of Match Letters: The letters and their repetition
- Number of Letters Match in Location: The number of letters (in the sentence) corresponds to the same letter in other sentences
- When delete Vowel letters with all the above details (such as Subword Models and N-gram overlap)

These values are calculated and the ratio of similarity is extracted as in the following steps:

Algorithm (2): Text Analysis

Inputs: Two text

Output: The ratio of comparison, similar or dissimilar

- Calculate the features specified for both texts.
- Give each feature a certain ratio and use of statistical factors.
- Compare the ratios of both texts using Parameters and weight.
- If close, texts are similar.
- If the ratios are far away, the texts are not the same.

Results the Discussion

The proposed work was created using C# language. Simple known statistics are used in comparison with their use after deleting some characters (vowels). Where the features are extracted again after deletion.

By changing the weight of each feature several solutions were obtained as shown in table (1) when weights are equal and table (2) when changing weights according to the most influential.

The length and shortness of the text affects the result where the longer the text is, the more accurate the result.

The weight can be changed for each feature, where data and other ratios will be obtained for similarity. Note feature (Number of Match in Location) is the highest and the most important effect, followed by feature (Number of Match Words) and then (Number of Different Words).

Table 1: Effect of each feature when weights are equal.

Feature	Weight	Match text	Match except some words lack	Match words with changing locations	Equal in word number	Equal in word number with some matching	Mixed (Real text)
Number of Words	10%	10	9.09	10	8.701	10	8.888
Number of Characters	10%	10	9.056	10	7.397	9.406	8.802
Number of Letters	10%	10	9.053	10	7.228	9.528	8.8
Number of Match Words	10%	10	9.09	10	1.688	3.766	6.666
Number of Different Words	10%	10	10	10	2.987	3.766	7.777
Number of Match in Location	10%	10	0	2.207	0	0	0.555
Number of Different Letters	10%	10	9.053	10	6.678	8.192	8.64
Number of Match Letters	10%	10	0.378	10	0.499	0.567	3.68
Number of Letters Match in Location	10%	10	9.053	10	7.194	9.511	8.8
Delete Vowel letters	10%	10	7.78	9.22	6.011	6.592	8.004
Ratio		100%	72.557%	91.428%	48.386%	60.932%	70.615%

Table 2: Effect of each feature when changing weights according to the most influential.

Feature	Weight	Match text	Match except some words lack	Match words with changing locations	Equal in word number	Equal in word number with some matching	Mixed (Real text)
Number of Words	4%	4	3.636	4	4	3.272	3.555
Number of Characters	3%	3	2.716	3	3	2.726	2.640
Number of Letters	4%	4	3.621	4	4	3.683	3.52
Number of Match Words	18%	18	16.363	18	18	0.467	12
Number of Different Words	15%	15	15	15	15	-2.357	11.666
Number of Match in Location	20%	20	0	4.415	4.415	0	1.111
Number of Different Letters	8%	8	7.242	8	8	5.604	6.912
Number of Match Letters	10%	10	0.378	10	10	0.154	3.68
Number of Letters Match in Location	13%	13	11.769	13	13	13.962	11.44
Delete Vowel letters	5%	5	3.349	4.22	4.22	1.69	3.473
Ratio		100%	64.078%	64.078%	83.636%	29.223%	60.00%

After the experiment we found the effect of the features sequentially as follows:

1. Number of Match in Location.
2. Number of Match Words.
3. Number of Different Words.
4. Number of Letters Match in Location.
5. Number of Match Letters.
6. Number of Different Letters.
7. Delete Vowel letters.
8. Number of Letters.
9. Number of Words.
10. Number of Characters.

Conclusion:

The initial processing of text was very important in reduce total processing time and extracting features. Preliminary results showed the benefit of extracting some features and pattern recognition in discovering similarities between texts. There are features that are more influential than others in discovering similarities. In the future, we can use the words stemming in processing. And discover the document similarities using semantic analysis approach of text.

References:

[1] F. Sebastiani, “Machine learning in automated text categorization,”ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[2] Mohammad A. Al-Ramahi , Suleiman H. Mustafa, .2012. N-Gram-Based Techniques for Arabic Text Document Matching; Case Study: Courses Accreditation. Abhath AL-Yarmouk: "Basic Sci. & Eng, 21(1), pp: 85-105.

[3] M.K.Vijaymeena1 and K.Kavitha, A Survey on Similarity Measures in Text Mining, Machine Learning and Applications, Machine Learning and Applications: An International Journal (MLAIJ) 3(1), (2016) pp. 19-28.

[4] Harrag, F., Al-Qawasmah, E., 2010. Improving Arabic text categorization using neural network with SVD. JDIM 8 (4), 233–239.

[5] AMINUL I. and DIANA I., Semantic Text Similarity Using Corpus-Based Word Similarity

and String Similarity, University of Ottawa, ACM Transactions on Knowledge Discovery from Data, Vol. 2, No. 2, July 2008.

[6] Cer, Daniel & Diab, Mona & Agirre, Eneko & Nigo Lopez-Gazpio, ˜ & Specia, Lucia. (2017). SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation. 10.18653/v1/S17-2001.

[7] Suhad M., Aseel Q., Finding the Similarity between Two Arabic Texts, Iraqi Journal of Science, 2017, Vol. 58, No.1A, pp: 152-162.

[8] V Sharapova, E & V Sharapov, R. (2018). The problem of fuzzy duplicate detection of large texts. 270-277. 10.18287/1613-0073-2018-2212-270-277.

[9] Pawar, Atish & Mago, Vijay. (2018). Calculating the similarity between words and sentences using a lexical database and corpus statistics.

[10] Wang, Yue & Di, Xiaoqiang & Li, Jinqing & Yang, Huamin & Bi, Lin. (2018). Sentence Similarity Learning Method based on Attention Hybrid Model. Journal of Physics: Conference Series. 1069. 012119. 10.1088/1742-6596/1069/1/012119.

[11] Ramaprabha, J & Das, Sayan & Mukerjee, Pronay. (2018). Survey on Sentence Similarity Evaluation using Deep Learning. Journal of Physics: Conference Series. 1000. 012070. 10.1088/1742-6596/1000/1/012070.

[12] E. Ukkonen. Approximate string matching with q-grams and maximal matches. Theor. Comput. Sci., 92(1):191–212, 1992.

[13] Dice, L. R., Measures of the amount of ecologic association between species, Ecology, 26:297-302, 1945.

تقنية لاكتشاف التشابه بين النصوص المستندة على استخراج الميزات من النص

علاء عبدالقهار جهاد[1] ، مرتضى محمد حمد[2]

[1]جامعة الانبار، مركز الحاسبة الالكترونية
[2]جامعة الانبار، كلية علوم الحاسوب وتكنولوجيا المعلومات

المستخلص:

إن اكتشاف التشابه بين نصين مهم جدا ومفيد في العديد من التطبيقات. التشابه بين النصوص هو مجال البحث الأساسي في مجموعة البيانات ومستودع البيانات والتنقيب عن البيانات. توفر هذه الورقة إطارًا يعطي تشابهًا بين نصين مدخلين استنادًا إلى التعرف على الأنماط واستخدام مطابقة تقريبية للنص، هناك وزن يؤثر على نسبة التشابه. يقارن البحث التشابه بين نصين دون التقيد بالقواعد اللغوية أو استخدام المرادفات أو معاني الكلمات. اظهرت النتائج الاولية فائدة استخلاص بعض الميزات في اكتشاف التشابه بين النصوص.