Pitfall of Google Tri-Grams Word Similarity Measure

Linda Wong Lin Juan, Bong Chih How, Johari Abdullah and Lee Nung Kiong
Faculty Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia.
15020282@siswa.unimas.my

***Abstract*—This paper describes and examines Google Trigram word similarity based on Google n-gram dataset. Google Tri-grams Measure (GTM) is an unsupervised similarity measurement technique. The paper investigates GTM's word similarity measure which is the state-of-the art of the measure and we eventually reveal its pitfall. We test the word similarity with MC-30 word pair dataset and compare the result against the other word similarity measures. After evaluation, GTM word similarity measures is found significantly fall behind other word similarity measure. The pitfall of GTM word similarity is detailed and proved with evidences.**

***Index Terms*—Google Tri-grams; Pitfalls, Sentence Similarity, Text Similarity; Trigrams; Unsupervised; Word Similarity.**

I. INTRODUCTION

Text is composed of words and phrases. The two measures commonly used to gauge if two given text are similar are text similarity and text relatedness. Text similarity quantifies closeness of two texts. On the other hand, text relatedness is the degree of how two texts relate to each other. Theoretically, text relatedness is a function of word relatedness. Text relatedness measures are methods to quantify the relatedness of two texts while text similarity measures are methods that are used to identify how similar the texts to each other. According to Mihalcea Rada in guidebook of social science [1], there is an obvious relatedness between two phrases like *"We own a pet"* and "*I love animals*", even though they are obviously dissimilar. Text similarity and relatedness are two of the important area in the field of natural language processing and they are widely applied in real life like, detecting plagiarism [2], automatic question answering [3] that return candidate answers by evaluating textual data and information retrieval [4] as in searching for related articles based on the keywords like Google and Yahoo search engines.

To date, text similarity is computed by using word and phrase similarity. TrWP [5] is an unsupervised text similarity approach using both word and phrase similarity. It is a Bag-of-Word-and-Phrase (BoWP) approach where phrase-pair (unigram vs bi-gram or bi-gram vs bi-gram) are used to computes the text similarity. It adopts Sum-Ratio (product of sum and ratio between minimum and maximum of two numbers) to capture the strength of association between two overlapping Google n-grams based on the statistics in the Google n-gram dataset of overlapping n-grams associated with the two compared texts[5].

There is no lack of literatures since researchers like Landauer [6], Mihalcea [7], Li et. al [8], and Lin[9] wo have produced various text similarity measures. Well-known works like LSA[6] uses Singular Value Decomposition (SVD) to analyse the statistical relationships among words to find the semantic representation of words in a reduced dimensional space. To derive similarity, corresponding word vectors are computed of its cosine angle to obtain the text similarity. On the other hand, Li et al. [8] proposed a method that computes text similarity based on corpus statistics and syntactic information. The approach has also considered sequence of words of a text as it carries useful information and specific meaning. Liu [10] proposed a novel approach to compute short text similarity by considering semantic information, word order and the contribution of different parts of speech in a sentence. The overall sentence similarity is derived from a weighted combination of the distance between sub sequences.

In 2012, Islam [11] has reported that their proposed text similarity--Google Tri-grams Measure (GTM)--has outperformed many well-performed text similarities. The state-of-the-art of GTM measure is Google Tri-grams word similarity measure. Hence in this paper, we intend to detail how GTM word similarity works and at the same time, to highlight the pitfall of the measure. Lastly, we will present some evidences to verify the pitfall.

II. GOOGLE TRI-GRAMS WORD SIMILARITY MEASURE (GTM)

Google Trigrams Similarity Measure (GTSM) is a distributional method that uses a Google n-gram corpus dataset to find the inherent properties of similarity between texts. In general, GTSM has two main components: trigram word similarity and text similarity. Trigram The word similarity component is to derive word-word similarity which is the fundamental component that is required to derive the sentence similarity. The word-word scores are aggregated to deliver a score to represent text similarity. In this paper, we examine GTM word similarity.

The word similarity in GTM is derived through Google n-grams's tri-grams dataset. It takes into consideration all the tri-grams that begins and ends with the given pair of words regardless of their order. In additional, the most frequent unigram of each word is used to normalize the mean frequency of the tri-grams. The algorithm of the word similarity is described in detail in the following.

Given two words, w_a and w_b,

Step 1: First, obtain the sum of unigram frequency from Google unigram dataset, which is represented as F_{max}.

Step 2: Obtain the frequency of unigram w_a as $f(w_a)$, and w_b as $f(w_b)$ from Google unigram dataset.

Step 3: Between the unigram frequency of w_a and unigram frequency of w_b, choose the frequency of