A Sentence Similarity Measure Based on Conceptual Elements

Wendy Tan Wei Syn, Bong Chih How and Dayang Hanani Abang Ibrahim
Faculty Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia.
wendytws@siswa.unimas.my

***Abstract*—There has always been a growing interest in sentence similarity measure for practical NLP tasks using various state-of-art NLP methods. Some of the widely used methods in measuring sentence similarity are lexical semantics, deep learning, neural networks, ontology, statistical models, graph based model and etc. Based on our findings, one of the main drawbacks in using these methods is not able to resolve word ambiguity where one word can have different interpretations in different sentences. In this paper, we present a sentence similarity measure by representing the sentences in conceptual elements to measure the semantic similarity between sentences. We used Microsoft Paraphrase Corpus (MSR) and Quora question pairs dataset to evaluate the performance. The study concludes that we were able to use conceptual elements to measure sentence similarity with the highest micro averaged precision of 0.71.**

***Index Terms*—Sentence Similarity Measure; Concept; FrameNet.**

I. Introduction

Most of the sentence similarity measures derived from word's similarity, co-occurrence, word order, N-gram, synonym, antonym, and etc. However, two sentences that have different structure or even overlapped words can be semantically similar per se. For example *"I feel sad."* and *"My mood is down."*. While two sentences that shared 80% of identical words can be dissimilar. For example, *"I am Andy".* and *"I am sad."*. The sentences above obviously proved that bag of words (BOG) method has removed lots of detail yet less effective when the sentences contain ambiguous words conveying different meaning under different contexts. In order to find similar meaning sentences, hence semantically similar sentences, we need to go beyond word usage and sentence structure where a model can be trained to understand the concepts in the sentences.

A concept can be defined as "*a perceived regularity in events or objects, or records of events or objects, designated by a label"* [1]. Concepts is abstract. It is the mental representation of classes of things [2]. Concepts are represented with words or phrases. For example, the phrase of *"saves time"* represent the concept of efficiency.

When we want to understand a discussion, we are trying to grasp concepts using our background knowledge so that we can comprehend the statement. Here, concepts connect our past experience with the current interaction with the world [2]. Each concept is connected to one and another. We often discuss about common concept in our daily conversation. Throughout the conversation, we are using our own words when explaining something. For example, we might use different word such as *"wonderful!"*, *"fantastic!"*, but we are still conversing on the same concept which is expressing our feelings towards something.

In order to comprehence a sentence, besides trying to understand the each word's meaning, we have to capture the overall concept of the sentence as well, which are made up of words. As we know human have the ability to use memory as the inventory to structuring, classifying, and interpreting experiences [2], each particular word are associated in memory with particular frames (concept elements) [2]. For example, words such as *"buy"*, *"sell"*, *"pay"* able to activate the commercial event scenario in someone brain. Therefore it is crucial that in understanding a word's meaning requires knowing the whole scenario [2]. We might use the same word but referring to different other frames.

The same goes to when we want to identify if two sentences are similar, we should look at the similarity of concept besides overlapping words, syntactic similarity, or whether the sentence having the same subject-verb–object (SVO) or semantic role labelling (SRL). There are always possibilities that we might misunderstand the meaning if the above syntactic features are ambiguous. This illustrates the importance to focus on capturing the concept of a sentence in order to measure sentence similarity.

When we looking at the concepts, finding similar sentences mean finding sentences that are conceptually similar. When we represent the sentence as a concept, the words are categorized under a common concept which could help in sentence similarity measure. For example, *"This phone is easy to use"* and *"This phone is difficult to learn"*, the concept that we intend to capture for both of the sentences is the difficulty in using something. Throughout this paper, we will discuss on how to use conceptual elements in sentence to measure their similarity.

II. Problems

Recently, Google offshoot Jigsaw released a machine-learning-based service called Perspective which can be used to identify toxical comments to ensure the safety of Internet [3]. Perspective was trained from thousands of comments and was reported that the system tends to *"sensitized to particular words and phrases but not the meanings"* [3]. This clearly showed that the current AI approaches in understanding meaning in text remains a challenging issue especially when dealing with ambiguous scenarios.

Based on one of our experiment in using a computational model with the implementation of Latent Semantic Analysis (LSA) by Laudauer et. al. [4], we found out that in some cases the model failed to find related sentences which caused wrong classification. For example, one of the hiccup we run into was the following sentences *"Not to contradict myself, while Me functioned properly 80 percent of the time on my machine,*