Project proposal: incorporating logical constraints in AI-based text similarity searches

Overview

In many information retrieval tasks, we are interested in finding similar sentences while strictly enforcing logical rules between the query and retrieved entries. For example, in the context of searching for prior art for in patent application, the sentence "The invention is a machine that has 4 wheels but it is not a car.", should not be similar to any sentence related to cars. Nevertheless, modern text similarity approaches based on sentence embedding would not respect this condition.

The situation arises frequently in patent searches and in iterative searches when the user provides feedback based on previously retried results.

Goals

1. Understand the limitations of embedding-based similarity in the context of enforcing logic.
2. Suggest methods to improve searches.

Specifications

One relatively simple yet promising approach is to combine negation and other logical operators within the search, in a manner similar to standard search engine interface and database searches. For example, a good starting point is to augment vector similarity with logic deduced from the wording in the sentence such that sentences not meeting the logic criteria are excluded from the list of retrieved results.

Supervisor

The project will be supervised by Dr. Alon Kipnis.

Milestones

I. Obtain data

We will use patent publication data. The data will be provided by SenseIP.

II. Characterize failure of embedding-based similarity

Here we are looking for different examples demonstrating the issue with vanilla embedding-based similarity searches.

III. Literature review of logical operators in search

IV. Proposing and testing a new method

V. Evaluating the new method over benchmarks and the examples from Milestone II.