Explainability feedback tool for Natural Language Processing

Sourced by David Martinez Iraola from IBM Research Australia

About IBM Research Australia

IBM Research – Australia is one of IBM’s 12 global research labs. Opened in 2011, the Lab combines both research and development. IBM Research – Australia work closely with clients to understand their business challenges, innovation opportunities and how solutions can be applied in the marketplace. The research pursued at our Lab focuses on conceiving, designing and building next-generation systems that will transform health and life sciences, analytics, resource management, and financial services.

Project Overview

In this project we propose to develop a feedback tool for interpretable outputs for Natural Language Processing tasks. The feedback tool will have access to two main components: (i) prediction engine, and (ii) explainability module. The prediction engine performs automatic annotation tasks over a given textual input (e.g. identify drug adverse event (DAE) mentions in a medical article). The explainability module is given a textual input and a prediction, and it generates an explanation by linking the input text with existing manually annotated data.

Here are some examples. Given two sentences:

(1) “After discontinuing captopril and starting systemic steroids, her eosinophilia and radiographic abnormalities both resolved.”

(2) "Their parkinsonism improved on discontinuation of metoclopramide therapy.”

**Prediction module**

Task: identify drug adverse event (DAE) mentions in a medical article

Input: sentence (1), DAE, Knowledge base

Output: "eosinophilia" and "radiographic abnormalities"

**Explainability module**

Task: identifies the most semantically similar sentences in its knowledge base, and explains the predictions by linking the input and the existing annotations.

Input: sentence (1), "eosinophilia" and "radiographic abnormalities", existing annotated data (which is (2)), and a knowledge base

Output: The justification for the annotation in (1):

- Input Sentence (1) is similar to sentence (2) "Their parkinsonism improved on discontinuation of metoclopramide therapy”:

“After discontinuing" in (1) has the same meaning as “on discontinuation of” in (2)

“captopril” in (1) and “metoclopramide therapy” in (2) are both Interventions

“both resolved” in (1) is similar to “improved” in (2)

Since "metoclopramide therapy" is known to be a DAE in Sentence (2), the explanation module concludes that "eosinophilia" and "radiographic abnormalities" should be annotated in Sentence (1).

\* User interface and feedback collection

The goal of the user interface is to provide the predictions and explanations from the previous modules in a user-friendly way, and collect feedback from the user on whether the given predictions and explanations are accurate. Both the prediction and explanation models are based on machine learning, and they may make mistakes. The user should have easy ways to flag these errors, which will in turn be used to improve the system in further iterations.

The user interface may cater to different purposes. It may be used by potential clients to estimate the reliability of a machine learning system, for developers to identify sources of error, or for manual annotators to contribute knowledge to the system.

Intellectual Property

This project is intended to be open sourced under the MIT licence for all parties to share and use.