Similitud textual automatizada: evaluación algorísmica en la detección de duplicados

Motger de la Encarnación, Joaquim
Universitat Oberta de Catalunya
Universitat Politècnica de Catalunya

Resumen

El procesamiento de lenguaje natural o *Natural Language Processing*(NLP) és una de las áreas del *Machine Learning* (ML) en pleno auge y con un gran potencial en distintas áreas. Consiste en la extracción de representaciones parciales de las características y normas que rigen el lenguaje natural a partir de la información textual, tales como la información sintáctica como semántica[1]. El objetivo es utilizar este conocimiento para aplicar análisis automatizados y obtener funcionalidades relacionadas con la generación de unidades textuales.

Una de estas áreas de aplicación es la Ingeniería de Requisitos (RE), que comprende las actividades y procesos de ingeniería del software centrados en el desarrollo, análisis, comunicación y gestión de los requisitos que describen un sistema[2]. La experiencia en desarrollo de proyectos software, especialmente dado el crecimiento del volumen de datos y las dimensiones con las que estos trabajan, ha generado problemáticas relacionadas con la gestión y mantenimiento de requisitos (información textual)[3]. Su gestión de manera manual supone una tarea tediosa que requiere una inversión de tiempo excesivamente elevada, siendo igualmente crucial para el correcto desarrollo de cualquier proyecto software. Entre estas problemáticas podemos destacar la detección y gestión de requisitos duplicados o de una gran similitud semántica que deriva en malas prácticas tales como redundancias en la información textual del proyecto o duplicidad de tareas.

Este es el punto de partida de esta tesis: aplicar la similitud textual automatizada, utilizando herramientas de PLN, para la detección de duplicados entre requisitos de un proyecto. El objetivo principal de esta investigación, por tanto, será en primer lugar un análisis exhaustivo del estado del arte del área de evaluación de similitud textual automatizada, basado en el estudio de distintos algoritmos de similitud, sus puntos fuertes y débiles, y sus aplicaciones. En segundo lugar, y orientado a un caso de uso de aplicación real, el desarrollo e implementación de un algoritmo específico que cumpla las características esenciales para garantizar una precisión considerada como aceptable, en base al estado del arte y al preproceso y postproceso de las unidades textuales de información (requisitos) con los que se cuente[4].

Esta investigación se realiza en el marco de OpenReq (https://openreq.eu/), un proyecto de investigación cuyo propósito es generar una plataforma con un *toolkit* de herramientas *open source* basadas en la recomendación y la toma de decisiones automatizadas en tareas relacionadas con la RE.

Keywords— ingeniería de requisitos, machine learning, lenguaje natural, similitud textual, detección de duplicados

Referencias

[1] Collobert, Ronan, Jason Weston, León Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research, http://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf.

[2] Jeremy Dick, Elizabeth Hull, Ken Jackson, Requirements Engineering, Springer, pp. 7–9, ISBN 978-3-319-61073-3