Estudio comparativo sobre la aplicación de librerías PLN para reconocimiento de entidades en la validación de requerimientos

[1]Sonia Santana, [1]Lucrecia Perero, [1]Noelia Rodriguez, [2]Alejandro Fernandez, [2]Leandro Antonelli

[1]Facultad de Ciencias de la Administración - Universidad Nacional de Entre Ríos
sonia.santana, ramona.perero [@uner.edu.ar], romi2022@gmail.com,
[2]Laboratorio de Investigación y Formación en Informática Avanzada (LIFIA),
Facultad de Informática, Universidad Nacional de La Plata
alejandro.fernandez, leandro.antonelli [@lifia.info.unlp.edu.ar]

Resumen

Este artículo presenta los avances de las líneas de investigación iniciadas en la Universidad Nacional de Entre Ríos (UNER) que tiene como objetivo desarrollar un proceso de validación de requerimientos utilizando, entre otras técnicas, procesamiento de lenguaje natural y entornos colaborativos [1]. Como parte de las actividades previstas en este proyecto se realizaron dos revisiones bibliográficas para conocer las principales tendencias de técnicas de procesamiento del lenguaje natural (PLN) y de entornos colaborativos en la validación de requerimientos. Además, se analizaron distintas librerías de PLN para el Reconocimiento de Entidades Nombradas (por sus siglas en inglés, NER) en la validación de requerimientos.

Palabras clave: validación de requerimientos, técnicas de procesamiento de lenguaje natural, reconocimiento de entidades nombradas.

Contexto

El presente PID 7070 se encuadra en la línea de investigación "Ingeniería de Software" y es un proyecto en conjunto entre Facultad de Ciencias de la Administración de la UNER y la Facultad de Informática de la Universidad Nacional de la Plata (UNLP). La línea de investigación es establecida como prioritaria desde la carrera Licenciatura en Sistemas de la Facultad de Ciencias de la Administración de la UNER. Se adecua, además, a una de las prioridades de la UNER considerando que es un proyecto aplicado a la investigación sobre Tecnologías de la Información y la Comunicación [1].

Introducción

La validación de requerimientos es el proceso encargado de verificar que la especificación de requerimientos se ajuste al sistema deseado por el cliente [2]. El objetivo es alcanzar que todos los requerimientos del sistema se hayan establecido sin ambigüedades, incoherencias, omisiones y errores [3] para evitar problemas de costos y retrasos adicionales en etapas posteriores del ciclo de desarrollo del software [4]. Definir y validar los requerimientos del sistema son las actividades más importantes en el desarrollo del software. A menudo requieren la colaboración de múltiples partes interesadas que tienen diferentes necesidades y perspectivas. En muchos casos, es especialmente difícil facilitar la recopilación de requerimientos de manera eficiente y eficaz en un entorno jerárquico y, al mismo tiempo, animar a las partes interesadas del sistema a compartir libremente sus ideas y opiniones [5].

Para organizar, conservar y recuperar la información pertinente al documento de especificación de requerimientos de software, muchas veces es necesario un sistema automatizado de extracción de información [6]. Si bien la información en la especificación de requerimientos tiene una estructura semántica coherente para los humanos, las computadoras no suelen ser capaces de comprenderla. Sin embargo, la información puede ser utilizada en trabajos de etiquetado y clasificación, brindando información valiosa sobre procesos, entidades y otros aspectos del desarrollo de software. La información obtenida de la especificación de requerimientos, se puede utilizar para trabajar con el NER. El NER permite identificar elementos o entidades de un texto y clasificarlas en un conjunto predefinido de tipos tales como persona, organización y ubicación, permite conseguir que los requerimientos sean correctos y disminuye las posibilidades de encontrar errores durante el proceso de validación, obteniendo un producto de calidad [7].

Asimismo, el NLP4RE es un área de investigación y desarrollo que busca aplicar tecnologías del PLN (técnicas, herramientas y recursos) a diferentes tipos de documentos de requerimientos para respaldar una variedad de tareas de análisis lingüístico realizadas en varias fases de RE [8]. Desde finales de la década de 2000, el NLP4RE se ha convertido en un área de investigación que atrae a investigadores de la comunidad de Ingeniería de Requerimientos en general.

Zhao realizó el primer estudio de mapeo sistemático sobre el panorama de la investigación NLP4RE para comprender el estado de publicación, el estado de la investigación empírica, el enfoque de la investigación, el estado del desarrollo de herramientas y, finalmente, su uso de las tecnologías de PLN. La investigación muestra que el 67,08% de los estudios de NLP4RE son propuestas de solución que se evaluaron mediante un experimento de laboratorio o una aplicación de ejemplo, mientras que el 7,18% de los estudios se evaluaron en un entorno industrial, lo que destaca una falta general de evaluación industrial de los resultados de investigación de NLP4RE. Al mismo tiempo, los estudios analizados se centran en la fase de análisis, con la detección de defectos de calidad como tarea central de análisis lingüístico y la especificación de requerimientos como tipo de documento comúnmente procesado. Además, se encontraron un total de 130 nuevas herramientas para respaldar una variedad de tareas de análisis lingüístico, pero solo 17 de estas herramientas están disponibles para descargar. Se utilizaron 231 tecnologías de PLN diferentes, que comprenden 140 técnicas, 66 herramientas y 25 recursos, pero solo una cuarta parte se usan con frecuencia y las tecnologías de PLN más populares son las léxicas o sintácticas, como los etiquetadores POS, analizadores sintácticos, y WordNet [9].

Según Schmitt, el NER juega un papel clave en la detección y clasificación de entidades en aplicaciones de PLN. Ha demostrado que, a pesar de la disponibilidad de distintos softwares NER, sigue siendo difícil para los profesionales de la PLN identificar clara y objetivamente qué software funciona mejor. Una razón de esto es que la mayoría de los estudios existentes carecen de transparencia para permitir la

reproducibilidad de los experimentos, agregando que los distintos estudios de evaluación, es decir, evaluar un mismo software basado en un mismo corpus, a menudo conducen a resultados diferentes, a veces de manera sustancial. Schmitt realizó una evaluación de cinco herramientas NER, StanfordNLP, NLTK, OpenNLP, SpaCy y Gate, basado en dos corpus distintos, CoNLL 2003 y GMB. Los resultados muestran que StanfordNLP supera al resto de herramientas, de manera significativa respecto a CoNLL en un 30% más de rendimiento, lo que se esperaba debido a que el clasificador predeterminado fue entrenado en ese corpus, y de manera menos significativa con respecto a GMB en un 15%, siendo StanfordNLP incluso menos eficiente que NLTK al etiquetar la "Ubicación" para el cual ninguno de los softwares evaluados recibió aprendizaje [10].

Continuando con una de las líneas de investigación y en el marco del proyecto se propone identificar y clasificar entidades de texto de la especificación de requerimientos mediante el uso de librerías para determinar si los requerimientos son correctos y lograr disminuir las posibilidades de encontrar errores durante el proceso de validación, obteniendo un producto de mayor calidad.

Líneas de Investigación, Desarrollo e Innovación

I. Técnicas de PLN.
II. Entornos Colaborativos.
III. Librerías de PLN para NER.

Los resultados obtenidos / esperados se pueden resumir en:

Con el desarrollo de las actividades previstas, se ha avanzado en las distintas líneas de investigación. Se ha realizado el estudio de técnicas de procesamiento de lenguaje natural en la validación de requerimientos, mediante una revisión bibliográfica de las principales tendencias de las técnicas de PLN en la validación de requerimientos publicados en los últimos 13 años. Se obtuvieron las siguientes contribuciones, las técnicas de PLN [11]:

- Se centran en el análisis semántico, es decir, en el significado contextual de las palabras y en la validación de requerimientos.
- Validan la propiedad de correctitud de los requerimientos utilizando patrones en la especificación de requerimientos y se aplican en la etapa inicial del desarrollo del software.
- Proporcionan diferencias en el enfoque de dominio en la validación de requerimientos con diversos grados de éxito.
- Requieren personal con experiencia en Lenguaje Natural Controlado (LNC) debido a que aplican algoritmos de aprendizaje automático en la validación de requerimientos.
- Evidencian el escaso uso de indicadores de rendimiento para el control y seguimiento de los defectos en los requerimientos, así como también objetivos a ser evaluados por dichos indicadores.

Además, se ha elaborado el estudio de entornos colaborativos en la validación de requerimientos, mediante una revisión bibliográfica desde el año 2010 al año 2023 de las principales tendencias de entornos

colaborativos en la validación de requerimientos del software, obteniendo las siguientes contribuciones [12]:

- Utilizan herramientas sincrónicas o asincrónicas para proponer explicaciones y formulaciones alternativas de un concepto en el proceso de colaboración.
- Vinculan las necesidades y especificaciones de los grupos de trabajo para satisfacerlas mediante herramientas de colaboración.
- Buscan aumentar la calidad del producto de software y reducir la complejidad de la tarea de validación de requerimientos.
- Unifican los puntos de vista, promueven la comunicación y colaboración efectiva entre las partes interesadas además de integrar criterios y comprender los requerimientos del software.

Por último, se ha avanzado en el análisis de librerías de procesamiento de lenguaje natural para el reconocimiento de entidades en la validación de requerimientos, obteniendo las siguientes contribuciones, la mayoría de las librerías analizadas [13]:

- Son multilenguaje.
- Tienen eficiencia y facilidad de uso moderada.
- Trabajan con modelos pre-entrenados y externos.
- Son de código abierto y poseen buena documentación.

En el marco del proyecto se espera, además:

- Avanzar en la capacitación continua de los miembros de la línea de investigación.
- Avanzar en el aprendizaje de librerías de PNL para el reconocimiento de entidades teniendo como finalidad aplicarlas al proceso de Validación de Requerimientos.
- Avanzar en el estudio de procesos de entrenamiento de cada herramienta.

Formación de Recursos Humanos

Este estudio prevé la formación e iniciación en actividades de investigación de cinco alumnos de la carrera de Licenciatura en Sistemas, el comienzo de un proyecto de Trabajo Final y la continuidad de un trabajo de tesis de maestría en la Facultad de Informática de la UNLP.

Referencias

1. Santana S., Perero L., Fernandez A., Antonelli L.: Proceso de validación de requerimientos aplicando técnicas de procesamiento de lenguaje natural en un entorno colaborativo, in Libro de Actas XXV Workshop de Investigadores en Ciencias de la Computación, pp. 444-448, ISBN 978-987-3724664, (2023).
2. Sommerville, I.: Software engineering, 9a ed. Boston: Pearson, (2011).
3. Pressman, R. S.: Software engineering: A practitioner's approach, 5a ed. Boston, Mass: McGraw Hill, (2001).
4. Santana, S. R., Antonelli, R. L., Thomas, P. J.: Evaluación de metodologías para la validación de requerimientos. In XXVII Congreso Argentino de Ciencias de la Computación (CACIC), (2021).
5. A. Fruhling, L. Steinhauser, G. Hoff, C. Dunbar: Designing and Evaluating Collaborative Processes for Requirements Elicitation and Validation, 40th Annual

Hawaii International Conference on System Sciences (HICSS'07), Waikoloa, HI, USA, pp. 15-15, (2007).
6. Apache OpenNLP: Apache OpenNLP Developer Documentation. Recuperado de https://opennlp.apache.org/docs/2.3.0/manual/opennlp.html, (2023).
7. Malik, G., Çevik, M., Khedr, Y., Parikh, D., Başar, A.: Named Entity Recognition on Software Requirements Specification Documents. Proceedings of the Canadian Conference on Artificial Intelligence, (2021).
8. Zhao L., Alhoshan W., Ferrari A., Letsholo K., Ajagbe M., Chioasca E., Batista Navarro R.: Natural Language Processing for Requirements Engineering: A Systematic Mapping Study. ACM Comput. Surv. 54, 3, Article 55, 41 pages, (2021).
9. Dalpiaz F., Ferrari A., Franch X., Palomares C.: Natural Language Processing for Requirements Engineering: The Best Is Yet to Come, in IEEE Software, vol. 35, no. 5, pp. 115-119, (2018).
10. Schmitt X., Kubler S., Robert J., Papadakis M., LeTraon Y.: A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate, Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, pp. 338-343, (2019).
11. Santana S., Perero L., Rodriguez N., Antonelli L., Fernandez A.: Estudio de Técnicas de Procesamiento de Lenguaje Natural en la Validación de Requerimientos, XXIX Congreso Argentino de Ciencias de la Computación (CACIC), pp. 352-362, ISBN 978-987-9285-51-0, (2023).
12. Santana S., Perero L., Rodriguez N., Fernandez A., Antonelli L.: Estudio de entornos colaborativos en la validación de requerimientos, XXI Jornadas de Administración e Informática, Universidad Nacional de Entre Ríos, Facultad de Ciencias de la Administración, ISBN 978-950-698-578-3, (2023).
13. Costas A., Dri A., García Gelsi R., Hernández L., Horta F., Medvedovsky F., Tano N.: Análisis de librerías de procesamiento de lenguaje natural para el reconocimiento de entidades en la validación de requerimientos, XXI Jornadas de Administración e Informática, Universidad Nacional de Entre Ríos, Facultad de Ciencias de la Administración, ISBN 978-950-698-578-3, (2023).