Alineación de glosarios del dominio

Domain Glossary Alignment

Paola Grijalva Arriaga[1,2][0000-0003-2616-417X], Leandro Antonelli [2][0000-0003-1388-0337] and Pablo Thomas[3][0000-0001-9861-987X]

[1] Escuela de Ingeniería en Computación e Informática, Facultad de Ciencias Agrarias, Universidad Agraria del Ecuador, Av. 25 de Julio y Pio Jaramillo, Guayaquil, Ecuador
[2] LIFIA - Facultad de Informática, Universidad Nacional de La Plata, 50 esquina 120, Buenos Aires, Argentina
[3] LIDI - Facultad de Informática, Universidad Nacional de La Plata, 50 esquina 120, Buenos Aires, Argentina
pgrijalva@uagraria.edu.ec
leandro.antonelli@lifia.info.unlp.edu.ar
pthomas@lidi.info.unlp.edu.ar

Resumen. Durante la etapa de especificación de requerimientos se hace uso de descripciones conceptuales del Universo del Dominio (UdD) a través de varias representaciones como los Glosarios. Las organizaciones tienen documentos que permiten a los ingenieros de requerimientos obtener una descripción de procesos y procedimientos que aportan al conocimiento del dominio. Por otro lado, organizaciones externas poseen también documentación que aporta al mismo dominio, pero no necesariamente utilizando los mismos términos o vocabulario. Es por ello, que es necesario una alineación de los vocabularios para poder compararlos y obtener una descripción más completa y consistente. Los glosarios son herramientas que se utilizan durante esta etapa y permiten la descripción del dominio. El objetivo de este trabajo doctoral es definir un proceso de alineación de glosarios del dominio representados a partir del Léxico Extendido del Lenguaje (LEL), integrando heurísticas y métodos semánticos, léxicos, con el fin de hallar similitudes, diferencias u omisiones. Se realiza una investigación documental de la literatura con respecto a las técnicas utilizadas para alinear dominios específicos y los utilizados para alcanzar la completitud en glosarios. Luego, se realizará la creación del método que contemple heurísticas e integre técnicas de similitud semánticas y léxicas. La evaluación del proceso se realizará mediante un experimento utilizando un caso de estudio. Y a través de un caso de estudio se validará el resultado del proceso con la alineación de LELs de dominios relacionados.

Abstract. During the requirements specification stage, conceptual descriptions of the Domain Universe (UdD) are used through various representations such as Glossaries. Organizations have documents that allow requirements engineers to obtain a description of processes and procedures that contribute to domain knowledge. On the other hand, external organizations also have documentation that contributes to the same domain, but not necessarily using the same terms or vocabulary. For this reason, it is necessary to align the vocabularies in order to compare them and obtain a more complete and consistent description. Glossaries are tools used during this stage and allow the description of the domain. The objective of this doctoral work is to define a process of aligning domain glossaries represented from the Extended Language Lexicon (LEL), integrating heuristics and semantic, lexical methods, in order to find similarities, differences or omissions. A documentary research of the literature is carried out regarding the techniques used to align specific domains and those used to achieve completeness in glossaries. Then, the creation of the method that contemplates heuristics and integrates semantic and lexical similarity techniques will be carried out. The evaluation of the process will be carried out through an experiment using a case study. And through a case study, the result of the process will be validated with the alignment of LELs of related domains.

1. Introducción/Motivación

La definición de requerimientos es una construcción gradual desde el estudio del dominio del problema hasta la captación de los diferentes requerimientos de las partes interesadas. La completitud es una característica deseada, sin embargo, es muy difícil lograr. Ridao et al. [1] indican que obtener un modelo de requerimientos completo es una meta inalcanzable, y la sola estimación del grado de completitud alcanzado es muy difícil. Una de las herramientas utilizadas y muy necesaria durante la fase de elicitación de requerimientos, son los glosarios. Martin Glinz [2] define un Glosario como una colección de definiciones de términos que son relevantes en algún dominio, que contiene referencias cruzadas, sinónimos, homónimos, siglas y abreviaturas. Arora et al.[3] indican que un glosario es una parte importante de cualquier documento de requerimientos de software. Hace explícitos los términos técnicos en un dominio y proporciona definiciones para ellos, ayudando a mitigar la imprecisión y la ambigüedad. Actualmente, es muy común la interacción de los sistemas informáticos, en donde cierto sistema brinda servicios para que consuman otros sistemas. En este marco, contar con dos LEL, en donde cada uno describa un dominio diferente, pero con perspectiva de que haya un borde de conexión, permite identificar fronteras entre los dos dominios de forma tal de que permita la interoperabilidad. En este caso, es crítico identificar la intersección, superposición y solapamiento de los glosarios del dominio, para que los sistemas se comuniquen y puedan intercambiar información.

El objetivo de este trabajo doctoral es definir un proceso de alineación de glosarios del dominio representados a partir del Léxico Extendido del Lenguaje (LEL), con el fin de hallar similitudes, diferencias u omisiones. Este proceso, brindaría solución a

tres situaciones. Por un lado, permite mejorar la calidad de los LELs, ya que permite reutilizar un LEL ya construido para enriquecer uno nuevo. Por otro lado, este proceso, permitiría analizar descripciones de dominios diferentes, buscando un borde de conexión, para identificar fronteras entre los dos dominios que permita la interoperabilidad. Finalmente, la alineación de dos descripciones del mismo dominio representados en diferentes glosarios permite unificar el lenguaje como ocurre por ejemplo cuando se necesita en iniciativas de certificación de calidad donde la organización debe alinear la descripción de sus prácticas a las descripciones de los estándares. La Figura 1 describe la arquitectura de la iniciativa propuesta.

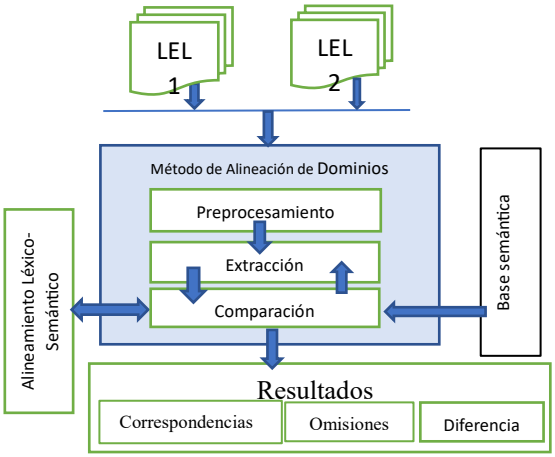

Fig. 1: Arquitectura propuesta

En la figura 1 se puede observar cómo sería el funcionamiento del método de alineación. Internamente estaría compuesto de un preprocesamiento que permite preparar el documento para que pueda ser analizado por las medidas de similitud léxica-semántica, la extracción que contendrá heurísticas y la comparación que estará relacionado con las medidas léxicos- semánticas y una base semántica con términos del dominio. Como resultado del proceso, se espera las correspondencias, omisiones y diferencias entre los LELs.

2. Estado del Arte

Se revisa información relacionada con proyectos de alineación de dominios, utilizando glosarios de dominio LEL, así como también trabajos referentes a la alineación utilizando ontologías basados en el uso de la semántica, léxica y estructural. En el caso de los Glosarios LEL, se ha encontrado documentación donde la alineación consiste en el desarrollo de heurísticas, en técnicas de inspección que permiten mejorar la completitud y disminuyen las omisiones como se puede observar

en [3],[4],[5],[6],[7]. Por otro lado, utilizando alineación de ontologías se encuentra trabajos que permiten alinear dominios utilizando técnicas de similitud semántica, estructural y léxica, como se puede observar en [8],[9],[10],[11],[12],[13].

3. Metodología y enfoque de investigación

La metodología para el desarrollo del proyecto doctoral se puede observar en la Figura 2.

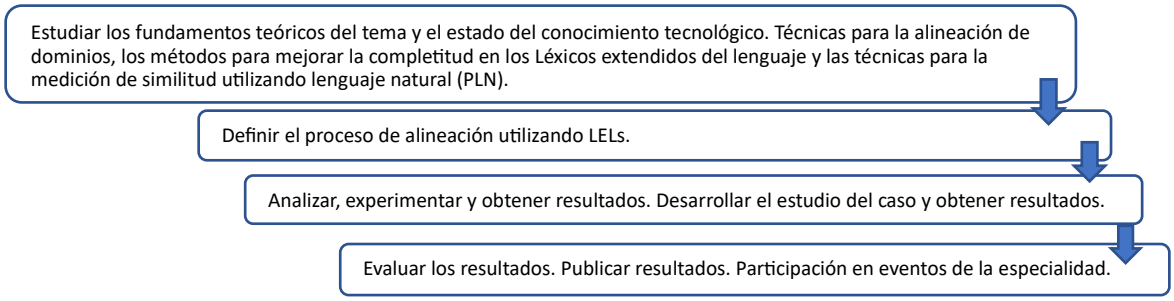

Fig. 2. SEQ Figura * ARABIC 2 Metodología de investigación y desarrollo

Primero, se realizará una investigación documental de la literatura con respecto a la temática, caracterizando el estado actual de la ciencia con respecto a los métodos utilizados para la alineación de dominios y los utilizados para alcanzar la completitud en los léxicos extendidos del lenguaje. Luego, se desarrollará el proceso de alineación integrando las técnicas que se adecuen a las naturales del LEL. La evaluación del proceso de alineación de LELs a través de la experimentación, obteniendo como resultados las heurísticas y técnicas de alineación que brinden mejores resultados, con la ejecución del proceso. Luego, se aplicará el modelo a un caso de estudio, utilizando LELs relacionados a un mismo dominio para identificar sus similitudes, diferencias u omisiones. Finalmente, con la colaboración de un ingeniero de requerimientos y usuarios expertos en el área se validará los resultados obtenidos del modelo.

4. Plan de Evaluación

Mediante un experimento se evaluará las técnicas de alineación de similitud semánticas-léxicas aplicadas en glosarios LEL, seleccionando en cada ámbito las técnicas con mayor precisión en sus resultados. Para su ejecución, se formarán conjunto de símbolos de cada LEL y en cada uno se definirán subconjuntos con la descripción correspondiente a cada símbolo. Para cada ámbito léxico o semántico se aplicará a los conjuntos y subconjuntos las medidas de distancia y similitud utilizadas en este contexto. Para la ejecución del experimento, se utilizará librerías estadísticas y de aprendizaje automático incluidas en Python.

A través de un caso de estudio se validará el proceso de alineación en base a los resultados obtenidos de la ejecución. Esta validación será realizada por un ingeniero de requerimientos y expertos en el área, a través de una IDE que facilitará visualización de resultados. El caso de estudio se enmarcá en el contexto de acreditación de las universidades y escuelas politécnicas del Ecuador (IES), que periódicamente son evaluadas para obtener la certificación que acredite la calidad mínima requerida para su funcionamiento. Para cada proceso de acreditación el Consejo de aseguramiento de la calidad (CACES) proporciona a las IES el modelo de evaluación, con el cuál serán evaluadas. Se considerará por un lado el LEL desarrollado del manual de políticas y procesos de un área sujeta a evaluación, como es el caso de investigación de la IES. Y el otro LEL estará relacionado con la función sustantiva de investigación del Modelo de evaluación para el aseguramiento de la calidad.

5. Resultados preliminares

Se realiza una investigación documental de la literatura con respecto a la temática, caracterizando el estado actual de la ciencia con respecto a las técnicas utilizadas para la alineación de dominios y los utilizados para alcanzar la completitud en los léxicos extendidos del lenguaje. La tabla 1 se observa la distribución de la clasificación documental seleccionada para análisis, las medidas más utilizadas en las diferentes técnicas y cómo han sido utilizadas estas técnicas.

Tabla 1: Distribución de la Clasificación

Herramienta	Técnica de alineación	Medidas utilizadas	Métodos
Ontología	Similitud léxica [11] Similitud Lingüística [7] Similitud semántica [21] Similitud estructural [9]	**Léxico:** Jaro Winkler, Levenshtein Distance, Jaro Distance, Hamming Distance. Medidas Rougel-L; **Semántico**: Euclidean, Coeficiente de Dice, Coeficiente de Jaccard, Resnik (res), Lin (lin) Jiang y Conrath, Leacock y Chodorow (lch), Wu y Palmer (wup), Path length (path) **Estructural:** similitud Herencia (MSI) Hermanos (MSS)	Experimentos y casos de estudios.**Semántico y léxico:** mapeos de referencia, ampliación de herramientas o prototipos, indexación estructural, emparejamiento **Estructural:** Biblioteca de patrones
Glosarios	Semánticos [5] Estructurales [13] Sintácticos [9]		Experimentos y Casos de Estudios con aplicación de Heurísticas, Variantes de Inspección y Mapas conceptuales

Los resultados de la revisión bibliográfica sobre las técnicas para la alineación de dominios utilizando glosarios y ontologías, muestran que entre las más utilizadas están a nivel semántico, léxico y estructural. Los estudios han concentrado mayor esfuerzo en el uso de ontologías para encontrar la similitud entre diferentes dominios,

utilizando varias técnicas como Euclidean, Manhathan, Coeficiente Jaccad con expansión en sinónimos, similitud coseno, empleándolas en proyectos relacionados de alineación utilizando máquinas de traducción, construcción automática de resúmenes, atribución de autoría, pruebas de lectura comprensivas, recuperación de información, que necesitan medir el grado de similitud entre dos textos dados. Por otro lado, con los glosarios de dominio (LEL), los trabajos se han concentrado en mejorar la completitud y la calidad de estos, mejorando las heurísticas y técnicas estructurales sobre el mismo glosario en análisis. Los trabajos de la revisión bibliográfica realizada se enfocan a encontrar similitudes, omisiones y errores dentro de un solo contexto.

6. Etapa doctoral

Temprana

7. Conclusiones

Los resultados de la revisión sobre las técnicas para la alineación de dominios utilizando glosarios y ontologías, muestran que entre las más utilizadas están a nivel semántico, léxico y estructural. Los estudios han concentrado mayor esfuerzo en el uso de ontologías para encontrar la similitud entre diferentes dominios, utilizando varios métodos como Euclidean, Manhathan, Coeficiente Jaccad con expansión en sinónimos, similitud coseno, en proyectos relacionados a máquinas de traducción, construcción automática de resúmenes, atribución de autoría, pruebas de lectura comprensivas, recuperación de información, que necesitan medir el grado de similitud entre dos textos dados. Por otro lado, utilizando LELs, se busca mejorar la completitud y la calidad de estos, mejorando las heurísticas y técnicas estructurales, creando mapas conceptuales con la información concerniente del dominio enfocados a encontrar similitudes, omisiones, errores dentro de un solo LEL. Cómo trabajo futuro, se creará un método de alineación de dominios, que recepte los LELs a ser alineados. Creando heurísticas e integrando medidas de similitud léxicas y semánticas y conexión a base semántica, los cuales serán probados mediante un prototipo para estos fines.

Referencias

1. M. Ridao y J. H. Doorn, «Estimación de completitud en modelos de requisitos basados en lenguaje natural», *WER 2006 - 9th Work. Requir. Eng.*, pp. 146-152, 2006, [En línea]. Disponible en: http://www.inf.puc-rio.br/wer/WERpapers/artigos/artigos_WER06/ridao.pdf.
2. M. Glinz, «A glossary of requirements engineering terminology», *Stand. Gloss. Certif. Prof. Requir. Eng. Stud. Exam, Version*, vol. 1, n.º May, p. 56, 2011.
3. C. Arora, M. Sabetzadeh, L. Briand, y F. Zimmer, «Automated Extraction and Clustering of Requirements Glossary Terms», *IEEE Trans. Softw. Eng.*, vol. 43, n.º 10, pp. 918-945, oct. 2017, doi: 10.1109/TSE.2016.2635134.

4. C. Litvak, G. Rossi, y L. Antonelli, «Conflict Management in the Collaborative Description of a Domain Language (S)», jul. 2018, n.º July, pp. 524-577, doi: 10.18293/SEKE2018-106.
5. A. Sebastián, G. D. S. Hadad, y E. Robledo, «Inspección centrada en Omisiones y Ambigüedades de un Modelo Léxico», *CIbSE 2017 - XX Ibero-American Conf. Softw. Eng.*, pp. 71-84, 2017.
6. A. Sebastián, G. D. S. Hadad, y D. Raffo, «Evaluación de Variantes de Inspección en un Modelo Léxico», *An. do WER 2019 - Work. em Eng. Requisitos*, pp. 1-6, 2019.
7. G. D. S. Hadad, C. S. Litvak, y J. H. Doorn, «Heurísticas para el modelado de requisitos escritos en lenguaje natural», *CACIC 2014 - XX Congr. Argentino Ciencias la Comput.*, 2014, [En línea]. Disponible en: http://sedici.unlp.edu.ar/handle/10915/42337.
8. D. R. Chandranegara y R. Sarno, «Ontology Alignment using combined similarity method and matching method», en *2016 International Conference on Informatics and Computing (ICIC)*, 2016, n.º Icic, pp. 239-244, doi: 10.1109/IAC.2016.7905722.
9. G. P. Kuntarto, Y. Alrin, y I. P. Gunawan, «The Key Role of Ontology Alignment and Enrichment Methodologies for Aligning and Enriching Dwipa Ontology with the Weather Concept on the Tourism Domain», en *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, oct. 2019, pp. 1-6, doi: 10.1109/ICICoS48119.2019.8982437.
10. N. Karam, A. Khiat, A. Algergawy, M. Sattler, C. Weiland, y M. Schmidt, «Matching biodiversity and ecology ontologies: challenges and evaluation results», *Knowl. Eng. Rev.*, vol. 35, p. e9, mar. 2020, doi: 10.1017/S0269888920000132.
11. B. B. Alhassan, «Extending an Ontology Alignment System with a Lexical Database», *Sci. Res. J.*, vol. III, n.º I, pp. 12-17, 2015.
12. I. Ouali, F. Ghozzi, R. Taktak, y M. S. Hadj Sassi, «Ontology Alignment using Stable Matching», *Procedia Comput. Sci.*, vol. 159, pp. 746-755, 2019, doi: 10.1016/j.procs.2019.09.230.
13. H. Karimi y A. Kamandi, «A learning-based ontology alignment approach using inductive logic programming», *Expert Syst. Appl.*, vol. 125, pp. 412-424, jul. 2019, doi: 10.1016/j.eswa.2019.02.014.