

# Tecnología de tutorización educativa basada en recursos multimedia semánticos

Miquel Canal Esteve, *Universidad Internacional Menéndez Pelayo*

Tutor: Yoan Gutierrez Vazquez, *Universidad de Alicante*

**Resumen:** En este trabajo se ha estudiado el efecto de usar tecnologías basadas en herramientas del procesamiento del lenguaje natural para mejorar la eficiencia de la tutorización docente cuando esta se complementa o es totalmente realizada con recursos multimedia. Se ha usado como ejemplo de uso la asignatura de Física de 2º de bachillerato, pero la tecnología aquí investigada es escalable a cualquier dominio de conocimiento que utilice recursos multimedia. El desarrollo de la investigación realizada ha consistido, en primer lugar, en la creación de una ontología (Physics Learning Object, PLO) como mecanismo para representar semánticamente el contenido docente multimedia. En segundo lugar, se han investigado y desarrollado técnicas de recomendación de contenidos multimedia basadas en semántica mediante el uso de tecnologías de procesamiento de lenguaje natural. Por último, con el fin de evaluar el desarrollo de dicha tecnología se ha desarrollado un asistente virtual en versión prototipo integrado en Telegram. Se ha testeado el asistente con 29 alumnos que lo han utilizado para resolver un ejercicio de física. Se les ha comparado con un grupo de control formado por 27 estudiantes que han resuelto el ejercicio pudiendo buscar libremente por internet. El impacto más significativo de esta investigación se ha observado en los estudiantes que tenían una nota media del curso baja, los cuales han mejorado su rendimiento (en comparación con el grupo de control) obteniendo un tamaño del efecto de  $d=2,33$  (alumnos de 4º de la ESO) y  $d=1,44$  (alumnos de 1º de bachillerato). Ambos resultados mejoran el estado del arte de investigaciones y tecnologías similares que en promedio tienen un  $d=0,76$ .

**Índice de términos:** e-learning, procesamiento del lenguaje natural, aprendizaje profundo, ontología, sistemas de tutorización inteligente, representación semántica de recursos multimedia

## I. INTRODUCCIÓN

Tras la pandemia producida por el COVID-19, el e-learning (aprendizaje electrónico) ha sido impulsado [1] y en consecuencia ha habido un aumento exponencial de contenidos educativos digitales, la mayoría de los cuales en formato vídeo como clases grabadas o generación de vídeos complementarios [2].

Este fenómeno acentúa una problemática que ya fue señalada en 2006 [3]: "La mayoría de los materiales didácticos no ofrecen posibilidad de reutilización debido a la falta de granularidad (es decir, de nivel de detalle) o posibilidades de acceso a diferentes dispositivos (PC, teléfono móvil, etc.) de manera eficiente".

Sin embargo, existen los conocidos como Sistemas de Tutorización Inteligentes (ITS) que se caracterizan por tener un elevado grado de granularidad y de hecho en [4], se demuestra que un aumento de granularidad implica una mejora en el rendimiento de los estudiantes. De hecho algunos de estos sistemas consiguen obtener en ciertas evaluaciones resultados similares a los obtenidos por un profesor particular según [5], donde se revisan 50 de ellos.

Sin embargo, en [6] se apunta que los ITS tienen los siguientes problemas:

- Las herramientas no son particularmente fáciles de usar.
- La creación de un ITS requiere mucho trabajo porque siempre se crea desde cero.
- El conocimiento y los componentes integrados en los ITS rara vez se pueden compartir o reutilizar.

En el mismo artículo [6] se plantea que la mejor solución ha esta problemática es el desarrollo de ontologías dado su carácter jerárquico que permite una buena organización, así como su carácter reutilizable.

Por otra parte, en [7] se muestra que una de las tareas que mejor realizan los docentes humanos en comparación con los ordenadores es la capacidad de indicar el contenido relacionado adecuado para que el alumno avance en su razonamiento, lo que técnicamente se conoce como *scaffolding*.

En resumen, la problemática actual consiste en un aumento de generación de recursos multimedia docentes [1] [2] y en la falta de tecnologías que sean capaces de reutilizar [6] y sugerir [7] adecuadamente el material multimedia didáctico, y que además sean fáciles de usar por los estudiantes y docentes [3] [6].

## II. MOTIVACIÓN

Dada la problemática descrita en la sección I, la hipótesis que se pretende investigar es que sería útil para los estudiantes que existiera una tecnología que permitiera estructurar y representar de manera reutilizable el contenido multimedia creado por los docentes, que además dicho contenido fuera accesible mediante un buscador sencillo de utilizar capaz de recomendar tanto imágenes, textos, audios, vídeos así como fragmentos de estos y que esta tecnología ofreciera siempre al estudiante

un contenido adicional relacionado con su exploración y/o pregunta en lenguaje natural.

### III. OBJETIVOS

Para poder realizar la propuesta mostrada en la sección II, se plantea el objetivo general de investigar y desarrollar una tecnología de tutorización educativa basada en recursos multimedia semánticos que permita facilitar el proceso de aprendizaje de los estudiantes.

Para llevar a cabo el objetivo general es necesario cumplir con los siguientes objetivos específicos:

En primer lugar, se debe realizar una revisión del estado del arte para ver si ya existe tal tecnología y en caso de que no exista se deberán construir las herramientas necesarias.

En segundo lugar, como se verá en la sección IV, no existen tales tecnologías, de manera que se deberá encontrar un método para representar los contenidos de forma semántica y relacionarlos para que puedan ser sugeridos.

En tercer lugar, se debe desarrollar un motor de búsqueda que pueda encontrar los recursos multimedia representados semánticamente.

En cuarto lugar, se debe desarrollar un algoritmo que haga de puente entre la representación semántica y un usuario que pueda utilizar dicha representación.

Por último, se debe realizar un experimento para estudiar la efectividad de la representación semántica de los contenidos.

### IV. ESTADO DEL ARTE

En esta sección se explica el estado del arte de la inteligencia artificial aplicada a la educación. En la sección IV-A, se explica que es un Sistema de Tutorización Inteligente (ITS) y se muestra cual es el estado del arte de esta temática. En la sección IV-B, se expone la manera de abordar de manera objetiva y cuantitativa el efecto de los ITS en los estudiantes. Por último, en la sección IV-C se hace una revisión de las herramientas del procesamiento del lenguaje natural que pueden ser utilizadas como motor de búsqueda en un repositorio semántico.

#### A. Inteligencia artificial aplicada a la educación: Sistemas de Tutorización Inteligentes (ITS)

Se ha observado que los alumnos consiguen mejores notas cuando son enseñados de manera individual en comparación con una clase convencional [4] [5] [8]. Esto significa que tienen el potencial para conseguir esos resultados. Sin embargo destinar un profesor para cada alumno es inviable en la mayoría de circunstancias. El reto es utilizar metodologías que permitan mejorar los resultados de una clase convencional sin los gastos que supone tener un profesor para cada alumno [8].

Los Sistemas de Tutorización Inteligentes son la herramienta más común de inteligencia artificial aplicada a la educación. Un Sistema de Tutorización Inteligente (ITS) es una tecnología de apoyo a la docencia que está constituida por los siguientes cuatro componentes [9] [10]:

- Un **modelo del dominio** que posee el conocimiento de la materia. Por ejemplo, todo el contenido sobre órbitas de la física de bachillerato.

- Un **modelo del alumno** en el que se registra la actividad del estudiante (sus respuestas, el tiempo que ha tardado en responderlas, etc.).
- Un **modelo pedagógico** que proporciona material específico a los estudiantes y responde a errores y preguntas como lo haría un maestro. Por ejemplo, establece que para resolver un problema de órbitas se pueden ofrecer seis pistas ordenadas de más fácil a más difícil.
- Una **interfaz** en la que el estudiante se comunica con el sistema.

Tanto el modelo de aprendizaje como el modelo pedagógico pueden utilizar herramientas de aprendizaje automático para permitir que estos modelos se adapten a las características de cada estudiante en función de su nivel sobre la materia en concreto y/o se adapten al contenido que el estudiante desea aprender utilizando sistemas de recomendación basados en motores de búsqueda para encontrar el contenido [10]. Este último punto conecta completamente con el área del procesamiento del lenguaje natural.

Existen varias opciones de ITS en el mercado y en los centros de investigación, cada una de ellas con sus características individuales. Sin embargo, se puede hacer una clasificación en tres grandes grupos en función del modelo pedagógico que se utiliza [10]:

- Secuencia de instrucciones y tareas paso a paso determinadas por el sistema, conocidos simplemente como ITS. Ejemplos de estos sistemas son: Mathia<sup>1</sup>, Assisments<sup>2</sup>, Area9 Lyceum's Rhapsode<sup>3</sup>, Century<sup>4</sup>, CogBooks<sup>5</sup>, iReady<sup>6</sup>, RealizeIt<sup>7</sup>, Smart Sparrow<sup>8</sup> y Summit Learning<sup>9</sup>.
- Instrucciones y tareas paso a paso determinadas por el sistema mediante un diálogo, conocidos como DBTS (*dialogue-based tutoring systems*): AutoTutor<sup>10</sup> y Watson Tutor<sup>11</sup>.
- El estudiante explora libremente las tareas y el material, conocidos como ELE (*Exploratory Learning Environments*): iTalk2Learn<sup>12</sup>, Betty's Brain<sup>13</sup> y ECHOES<sup>14</sup>.

Un meta-análisis de 50 ITSs concluyó que estos son efectivos [5], sin embargo, todos están enfocados en predecir cual es el siguiente mejor paso que debe dar el estudiante.

Se han identificado dos elementos que de mejorarse pueden complementar estas tecnologías. Por un lado, los componentes que constituyen estas tecnologías típicamente son difíciles de reutilizar (entre distintos niveles académicos o entre distintas asignaturas, por ejemplo) [3] [6] y por otro lado, ninguna de

<sup>1</sup><https://www.carnegielearning.com/solutions/math/mathia>

<sup>2</sup><https://new.assisments.org>

<sup>3</sup><https://area9lyceum.com>

<sup>4</sup><https://www.century.tech>

<sup>5</sup><https://www.cogbooks.com/technology>

<sup>6</sup><https://www.curriculumassociates.com/products/i-ready>

<sup>7</sup><https://realizeitlearning.com>

<sup>8</sup><https://www.smartsparrow.com>

<sup>9</sup><https://www.summitlearning.org>

<sup>10</sup><http://ace.autotutor.org/IISAutor/index.html>

<sup>11</sup><https://www.ibm.com/watson/education/pearson>

<sup>12</sup><https://www.italk2learn.com>

<sup>13</sup><https://wp0.vanderbilt.edu/oele/bettys-brain>

<sup>14</sup><https://www.ucl.ac.uk/ioe/research-projects/2020/sep/echoes-project>

ellas incorpora un buscador de fragmentos de vídeo, es decir que pueda encontrar un contenido exacto de 30 segundos en un vídeo de 2 horas que haya sido generado en la grabación de una clase (práctica habitual desde el COVID-19 [2]), en una conferencia subida a YouTube, o en otra plataforma.

Por esta razón se propone el desarrollo de un ITS fundamentado en una tecnología basada en recursos multimedia semánticos que haga uso de una ontología como repositorio y dotar a esta tecnología de un motor de búsqueda para que el estudiante pueda encontrar el contenido (en formato texto, imagen, vídeo, audio, documento o página web o una combinación de varios) preciso que necesita para realizar una tarea.

### B. Métrica para evaluar un ITS: Tamaño del efecto

En 1984 fue propuesto en [8] por primera vez el que paso a denominarse "problema de las dos sigmas" [10].

Este "problema" consiste en comparar dos distribuciones de notas de dos conjuntos de estudiantes, de los cuales, uno ha sido enseñado de manera estándar (en una clase convencional) y el otro mediante una metodología concreto o un ITS determinado. Tomando la desviación estándar ( $\sigma$ ) de la distribución de notas de la clase convencional se puede calcular cuantas veces se tiene que desplazar sigma para llegar a la media de la distribución del otro conjunto de estudiantes. Este factor recibe el nombre de tamaño del efecto (*effect size* en inglés), matemáticamente se puede definir de la siguiente forma y se puede ver un ejemplo gráfico en la Figura 1:

$$d = \frac{\text{Media experimental} - \text{Media control (convencional)}}{\sigma_{\text{control}}}$$

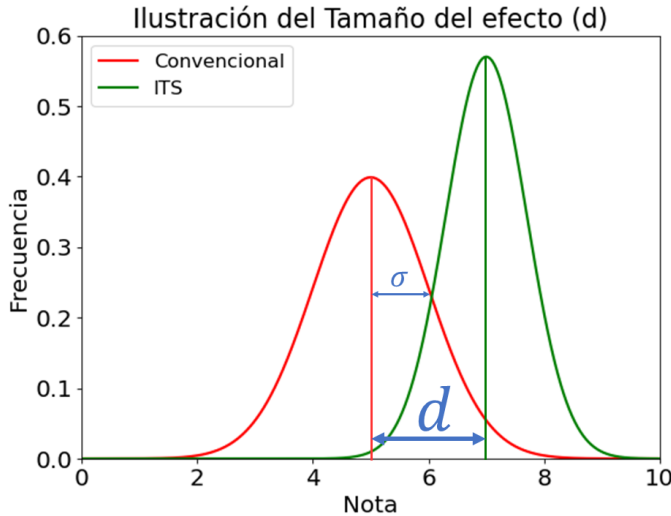


Fig. 1. El tamaño del efecto es una forma de cuantificar la distancia entre dos distribuciones de notas obtenidas por estudiantes que han aprendido mediante diferentes mecanismos. En este ejemplo el grupo representado por la línea roja fue enseñado de manera convencional (grupo clase con 30 alumnos) mientras que el grupo representado por la línea verde fue enseñado mediante un ITS. En este ejemplo se tendría un tamaño del efecto de  $d = 2$  dado que la  $\sigma = 1$  y las medias experimentales y de control son 7 y 5 respectivamente.

El problema de las dos sigmas refleja el hecho de que la mejor herramienta para enseñar algo a un alumno es un

profesor particular, el cual cuenta con un tamaño del efecto de  $d = 2$  según [8]. Sin embargo, estudios más recientes muestran que sería más adecuado hablar de un  $d = 0,79$  [5] y además esta métrica depende de cómo se esté evaluando a los estudiantes, de que tarea estén haciendo, que curso académico, etc. [5].

No obstante, el fenómeno que trata de ilustrar el problema de las dos sigmas es que la tutorización individual es la óptima para el aprendizaje de un estudiante y que lo que deberían hacer los sistemas ITS es acercarse al máximo posible al rendimiento conseguido de esta forma [8]. Estudios recientes muestran que los ITSs ya están obteniendo valores similares a la tutorización individual [5] obteniendo una media de un tamaño del efecto de  $d = 0,76$ .

Cabe decir que esta métrica también se puede utilizar para comparar ITSs entre ellos o distintas metodologías de aprendizaje independientemente de los ITSs.

### C. Procesamiento del Lenguaje Natural: Transformers

Con la finalidad de estudiar la tecnología del presente trabajo, el uso de tecnologías de procesamiento del lenguaje natural es necesario para poder comprender las preguntas en lenguaje natural que realizan los usuarios y poder recomendarles los recursos multimedia más a fines. A continuación se muestra un análisis de algunas de ellas.

Desde el artículo *Distributed Representations of Words and Phrases and their Compositionality* [11] publicado por Google en 2013 en el que se proponía el modelo word2vec para representar palabras mediante vectores de dimensiones mucho más reducidas que en la metodología anterior (*one-hot encoding*) no solo se consiguió reducir la dimensionalidad sino también capturar la relación semántica entre las palabras del vocabulario. La gran mayoría de avances en el campo del procesamiento del lenguaje natural siguen esta línea que se puede generalizar dentro del término de *transformers*, haciendo referencia al conjunto de algoritmos que utiliza esta tecnología.

En el presente trabajo se ha utilizado la parte del *embedding* de los *transformers* (véase Figura 6), de manera que se han codificado los textos (los escritos por el usuario y los de las instancias de la base de datos) mediante esta tecnología con la finalidad de realizar el cálculo de la similitud semántica [12] para ofrecer a los usuarios las respuestas adecuadas a sus consultas.

Existen varios modelos que se basan en esta tecnología y en este trabajo se han elegido probar los siguientes, que han sido evaluados con la métrica descrita en V-A3c para la selección del que mejor funcione: *paraphrase-xlm-r-multilingual-v1*, *paraphrase-multilingual-mpnet-base-v2* y *bert-base-multilingual-cased* porque según [13] *paraphrase-multilingual-mpnet-base-v2* es el modelo de mejor calidad de entre los modelos plurilingües. Por otra parte, se escogió *bert-base-multilingual-cased* porque es un modelo entrenado con textos de wikipedia. Dado que el dominio es de un ámbito científico-técnico sobre el cual wikipedia está bien dotado, se pensó que podría funcionar bien. Por último, *paraphrase-xlm-r-multilingual-v1* es un modelo que no destaca por encima de

los demás, pero se utilizó para tener una referencia con la que comparar los otros dos modelos plurilingües.

## V. METODOLOGÍAS Y TÉCNICAS UTILIZADAS

El objetivo principal descrito en la sección III es investigar y desarrollar una tecnología de tutorización educativa basada en recursos multimedia semánticos que permita facilitar el proceso de aprendizaje de los estudiantes. Dado que tras la revisión del estado del arte (sección IV) se ha identificado que no existe tal tecnología, se requiere la construcción de la misma.

En esta sección se expone la metodología que se ha seguido para construir dicha tecnología, así como la metodología planteada para realizar los experimentos.

### A. Marco general y flujo de información

En la Figura 2 se muestra el marco general que ha sido planteado para la construcción de una tecnología de soporte semántico de recursos multimedia aplicada a la docencia. Se han definido 5 componentes básicos (Ontología, Consultor, Relevancia, Visualización e Interfaz) que son gestionados por un componente central (Controlador).

A continuación se explica en que consiste cada componente y cómo debe ser el flujo de la información desde un punto de vista genérico para que sea reproducible en cualquier dominio. De manera complementaria se muestra la aplicación práctica de dos temas de la asignatura de física de 2º de bachillerato, ya que ha sido el contenido con el que han sido evaluados 56 alumnos (véase sección VI-C).

- **Ontología:** Su función es almacenar y organizar los recursos multimedia semánticos.
- **Consultor:** Su función es recuperar la información del repositorio semántico (la ontología) y transformarla a un formato tratable por los componentes Relevancia y Visualización. Una ontología suelen estar en formato RDF, de manera que las consultas se pueden hacer en SPARQL o un lenguaje similar.
- **Relevancia:** Su función es encontrar los recursos multimedia etiquetados semánticamente y disponibles en la base de datos. Debe tener un motor de búsqueda y un método para ordenar los resultados.
- **Visualización:** Su función es procesar el contenido del repositorio semántico y transformarlo para que sea presentable a través de la interfaz.
- **Interfaz:** Su función es hacer de puente entre el texto introducido por el usuario y la tecnología de soporte semántico. Puede ser una interfaz web, un servicio de mensajería instantánea (como Telegram, Facebook Messenger, What's App, etc.), una aplicación móvil desarrollada para dicha tarea, etc.
- **Controlador:** Su función es conectar los componentes Relevancia y Visualización con la Interfaz.

Los componentes Relevancia, Visualización y Controlador se pueden desarrollar con cualquier lenguaje de programación (python, R, Java, etc.) apto para realizar las funciones descritas.

A continuación se muestra como debe ser el flujo de la información (véase Figura 2):

- (1): El usuario introduce una consulta en formato texto.
- (2): El texto es enviado al componente Controlador mediante una interfaz que puede ser una interfaz web, un servicio de mensajería instantánea (como Telegram, Facebook Messenger, WhatsApp, etc.), una aplicación móvil desarrollada para dicha tarea, etc.
- (3): Se envía el texto al componente de Relevancia para identificar el mejor recurso (instancia) que se le puede ofrecer al usuario. para ello accede al repositorio semántico mediante el componente Consultor (4) y (5).
- (6): Si el usuario ha realizado una consulta, Relevancia devuelve la/s instancia/s de la/s mejor/es respuesta/s, en función de lo que decida el desarrollador (en el caso del presente trabajo se devuelven 3). Si el usuario ha seleccionado un recurso, Relevancia devuelve la instancia de dicho recurso.
- (7): Se envía la/s instancia/s al componente visualizador para preparar una infografía. Si el usuario ha seleccionado una instancia, se le ofrecerá la explicación correspondiente y contenido relacionado (imagen *a*). Si ha escrito un texto libremente se le ofrecerá las tres instancias que mayor puntuación hayan obtenido en Relevancia (imagen *b*). En el presente trabajo se ha realizado el diseño utilizando el formato html.
- (8) y (9): Si el usuario ha seleccionado una instancia, se hace una llamada al Consultor para obtener las instancias sugeridas.
- (10), (11) y (12): Se envía la infografía al usuario.

A continuación se detallan cada uno de los componentes.

1) **Base de datos (ontología):** Creación de la ontología Physics Learning Object (PLO), en español: Objetos de Aprendizaje de Física.

Existen varias metodologías para desarrollar una ontología, como por ejemplo BSDM [14], KADS [15], IDEF5 [16], etc, pero la escogida en este trabajo es la conocida como METHONTOLOGY [17] porque hace posible construir ontologías desde los conceptos hasta la implementación, incluyendo los principios de evaluación que se deben seguir.

A continuación se detallan los pasos seguidos aplicando esta metodología: planificación; especificación; adquisición de conocimiento, conceptualismo, reutilización e integración, formalización e implementación y evaluación.

*a) Planificación:* La fase de planificación consta de tres tareas: En primer lugar, estudiar el estado del arte de las representaciones semánticas. En segundo lugar, analizar las necesidades de los usuarios. Por último, examinar las dependencias entre las componentes de la ontología.

Estas tres tareas deben ser iteradas después de aplicar las fases de especificación y evaluación (secciones V-A1b y V-A1e). En el desarrollo de la ontología PLO<sup>15</sup> se tardaron tres meses en crear la primera versión que fue iterada dos meses más tarde y finalizó con una última iteración un mes después.

*b) Especificación:* Esta fase consiste en construir un esquema de la ontología utilizando lenguaje natural teniendo como referencia un conjunto de preguntas de competencia, las cuales son preguntas que podría hacerse un usuario y la respuesta debe encontrarse en el contenido de la ontología.

<sup>15</sup><https://github.com/miquelcanalesteve/tfm/blob/main/plo.v.0.4.owl>



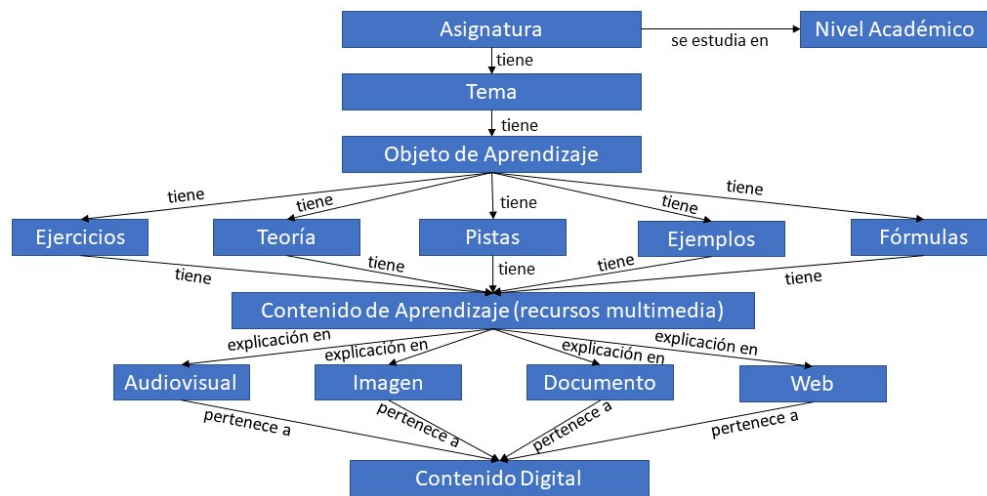


Fig. 3. Arquitectura de la ontología Physics Learning Object (PLO).

dar situaciones similares en todos los elementos del nivel 4. Se mantuvieron fijos los tres primeros niveles y se fijaron los niveles de Contenido de Aprendizaje y Contenido Digital a 98 y 99 respectivamente para dejar espacio para expandirse a los elementos del nivel 4.

d) *Formalización e implementación*: Esta fase consiste en definir la ontología utilizando un lenguaje formal que recoja el contenido desarrollado en las fases anteriores. Para desarrollar la ontología PLO se utilizó el programa Protégé Desktop v.5.5.0<sup>16</sup>. Según la nomenclatura de este programa, cada concepto (cada uno de los elementos de la Figura 3) se convirtió en una clase. Cada relación en una Propiedad del Objeto y cada atributo en una Propiedad de los Datos. La estructura completa de la ontología fue almacenada en un archivo OWL.

e) *Evaluación*: Hay cierto desacuerdo en cual es la forma en la que se debe evaluar una ontología [18] [19] [20] [21] [22], pero la tendencia actual es aceptar que el propósito principal de una evaluación es validar que el modelo conceptual encaja con el contenido. Por lo tanto la evaluación consiste básicamente en comprobar que no haya errores de construcción. Para ello se definen métricas cualitativas y cuantitativas. Las primeras corresponden a las preguntas de competencia y las segundas a un seguido de cálculos. En la sección VI-A se detallan los resultados obtenidos una vez aplicadas las métricas de evaluación de la calidad de la ontología resultante.

2) *Consultor*: Uno de los componentes principales del marco general (véase la Figura 2) es el Consultor. Mediante este componente se pueden obtener instancias y atributos de las instancias que serán utilizados por los componentes Relevancia y Visualización.

En el trabajo desarrollado se ha resuelto de la siguiente manera: Se ha subido la ontología al repositorio en línea rdf4j<sup>17</sup> y desde un código escrito en python se ha integrado la librería SPARQLWrapper que permite utilizar el lenguaje

SPARQL desde python para realizar consultas al repositorio, pues SPARQL es el lenguaje más usado para acceder al contenido de una ontología almacenada en un archivo OWL [23], como es el caso.

3) *Relevancia*: El componente Relevancia (véase Figura 2) tiene dos funciones principales. Por un lado, identificar si un usuario ha realizado una consulta o si ha seleccionado una instancia, por lo que se deberán utilizar herramientas del procesamiento del lenguaje natural.

Si el usuario ha seleccionado una instancia, Relevancia sencillamente transmite esa instancia al Controlador. Si ha realizado una consulta, Relevancia debe ordenar las instancias que se ofrecen como respuesta y seleccionar un número de ellas para ofrecérselas al usuario (en el caso de este trabajo se le han ofrecido 3 respuestas a cada consulta). Todo lo que se muestra a continuación en esta sección hace referencia a la metodología utilizada para desarrollar el motor de búsqueda capaz de realizar tal tarea.

Así pues, la metodología utilizada en el presente trabajo para identificar y ordenar las instancias ha consistido en el uso de un motor de búsqueda que combina un motor de búsqueda basado en embeddings, dado que se ha comprobado su efectividad [12], y otro basado en heurística (informática tradicional) que consiste en un recuento de palabras iguales, ya que se quiere analizar el hecho de que si las instancias de la ontología han sido bien etiquetadas mediante atributos por alguien que conoce como habla el usuario, este motor de búsqueda puede ser muy preciso.

Una vez obtenidos los resultados se genera una clasificación que ordena las instancias obtenidas mediante un sistema de puntuación que combina la similitud semántica (obtenida con el motor basado en IA) y el recuento de palabras iguales (obtenido del motor basado en heurística). A continuación se detalla la metodología de cada uno de estos procesos.

Cabe decir que sea cual sea el motor de búsqueda que se utilice, la búsqueda en el repositorio semántico debe hacerse utilizando como fuente de texto aquel texto que pertenezca a uno o varios atributos. En esta investigación se ha hecho uso de tres atributos que estaban definidos en todas las instancias

<sup>16</sup><https://protege.stanford.edu/>

<sup>17</sup><https://rdf4j.org>



Clase	Propiedades de los datos						
purl:Asignatura	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords			
plo:ContenidoDeAprendizaje	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords			
plo:CDAAudiovisual	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords	dma:linkStart	dma:linkEnd	
plo:CDADocumento	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords	plo:paginaInicio	plo:paginaFin	
dcterms:CDAImagen	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords			
plo:CDAWeb	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords			
dcterms:ContenidoDigital	plo:nivel	dcterms:autor	schema:url	dcterms:tipo			
plo:Ejemplo	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords	plo:enunciado	plo:resultado	
plo:Ejercicio	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords	plo:enunciado	plo:resultado	
plo:Formula	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords	plo:formula	physics:simbolo	physics:unidad
plo:Motivacion	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords			
purl:NivelAcademico	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords			
dbpedia:ObjetoDeAprendizaje	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords			
plo:Pista	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords			
plo:Tema	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords			
plo:Teoria	plo:nivel	dcterms:titulo	dcterms:descripcion	schema:keywords			

Fig. 4. Listado de clases y sus atributos correspondientes de la ontología PLO.

Dominio	Propiedad del Objeto	Rango
Ejemplo, Ejercicio, Formula, Motivacion, ObjetoDeAprendizaje, Pista, Teoria	plo:explicacionEnAudiovisual	CDAAudiovisual
	plo:explicacionEnDocumento	CDADocumento
	plo:explicacionEnImagen	CDAImagen
	plo:explicacionEnWeb	CDAWeb
	plo:relacionadoConEjemplo	Ejemplo
	plo:relacionadoConEjercicio	Ejercicio
	plo:relacionadoConFormula	Formula
	plo:relacionadoConMotivacion	Motivacion
	plo:relacionadoConObjetoDeAprendizaje	Aprendizaje
	plo:relacionadoConPista	Pista
	plo:relacionadoConTeoria	Teoria
Asignatura	plo:seEstudiaEn	NivelAcademico
Asignatura	plo:tieneTema	Tema
Tema	plo:tieneObjetoDeAprendizaje	ObjetoDeAprendizaje

Fig. 5. Listado de propiedades de los objetos (conexiones entre clases) de la ontología PLO.

(tal como se puede ver en la Figura 4): título, descripción y keywords (palabras clave).

a) *Motor de búsqueda basado en modelos de embeddings*: Existen una gran variedad de modelos de procesamiento del lenguaje natural que pueden ser utilizados como motores de búsqueda. Concretamente, de entre aquellos que utilizan embeddings, los modelos basados en *transformers* [11] son los más utilizados actualmente [24]. Su uso como motores de búsqueda se basa en utilizar la métrica *similitud* entre la entrada de texto del usuario y el contenido del repositorio semántico.

Así pues, para seleccionar las mejores instancias en un motor de búsqueda basado en embeddings se puede hacer uso de la *similitud* [12]. Se puede ver un esquema de esta metodología en la Figura 6 la cual muestra que el proceso consiste en codificar por un lado la consulta del usuario y por otro uno de los tres atributos (título, descripción o keywords) de una instancia del repositorio semántico. Este proceso se repite para cada uno de estos tres atributos, de manera que finalmente se obtiene un valor de la similitud que es la suma

de tres similitudes.

Se puede ver la lógica exacta de este proceso en el siguiente pseudocódigo:

```

Paso 1: textoEmb = Codificar el texto de
          entrada con el encoder del
          transformer.
Paso 2: nodosEmb = Codificar todos los
          nodos de la misma manera que
          en el Paso 1.
Paso 3: Crear una lista vacia scoreML = []
          y llenara de tantos 0s como nodos.
Paso 4: Por cada i en titulo, descripcion
          y keywords de un nodo:
          Por cada j en (0, len(nodos)):
              sim=similitud(textoEmb,
                             nodosEmb[i][j])
              scoreML[j] = scoreML[j]+sim

```

b) *Motor de búsqueda basado en heurística*: Por otra parte, el motor de búsqueda basado en heurística se basa en

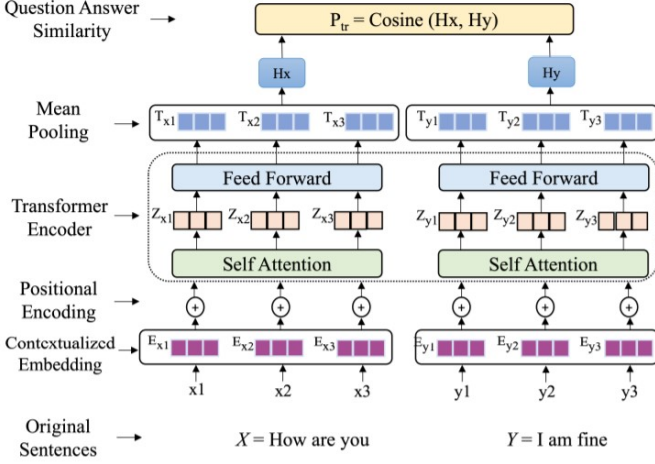


Fig. 6. Imagen extraída de [12]. Esquema del funcionamiento del cálculo de la similitud entre dos textos mediante un *transformers*. En el presente trabajo  $X$  sería el texto introducido por el usuario y  $Y$  el texto del atributo título, descripción o keywords. La similitud total se obtiene con la suma de las similitudes de los tres atributos.

contar el número de palabras iguales. Teniendo en cuenta que los usuarios no siempre siguen las reglas gramaticales cuando escriben se han pasado todas las palabras a minúsculas y se han eliminado las tildes. También se han eliminado las *stopwords* (palabras que se repiten mucho pero carecen de significado por sí mismas como preposiciones, artículos, etc.). Por lo tanto, el motor de búsqueda basado en heurística sigue la lógica que se muestra en el siguiente pseudocódigo:

```

Paso 1: textoToken = Tokenizar el texto de
          entrada por palabras. Ponerlas
          en minúscula, eliminar los acentos
          y eliminar los stopwords.
Paso 2: nodosToken = Tokenizar todos los
          nodos de la misma manera que
          en el Paso 1.
Paso 3: Crear una lista vacía score = []
          y llenarla de tantos 0s como nodos.
Paso 4: Por cada palabra en textoToken:
          Por cada i en titulo,
          descripción y keywords de
          un nodo:
          Por cada j en
          (0, len(nodos)):
              score[j] = score[j] + el
              número de palabras
              iguales.

```

c) *Evaluación intrínseca: Métrica para seleccionar el mejor motor de búsqueda:* Dado que no hay un motor de búsqueda estándar que sea exageradamente mejor que los otros [13], en función del contexto uno puede resultar más útil que otro, es necesario definir una métrica para compararlos. Además de utilizar esta métrica para comparar también el motor de búsqueda basado en heurística y evaluar un motor de búsqueda que combine varios motores.

Así pues, para comparar los modelos se ha utilizado la métrica acierto definida de manera similar a la que se emplea

en [25] para comparar el rendimiento de diferentes algoritmos en la clasificación de documentos. En [25] se define el acierto como el cociente entre los documentos bien clasificados y el número total de documentos. En el caso del presente trabajo se ha definido el acierto tal y como se muestra en la Fórmula 1:

$$\text{Acierto} = \frac{\text{consultas bien respondidas}}{\text{número total de consultas}} \quad (1)$$

Donde las consultas son preguntas que podría hacer un usuario. Se definieron un total de 31 consultas junto con la respuesta correcta correspondiente a una de las instancias de la ontología. Si el motor de búsqueda era capaz de responder correctamente se le otorgaba una puntuación de 1 y sino de 0.

En el presente trabajo se decidió que el motor de búsqueda ofrecería al usuario hasta 3 respuestas a cada consulta. Se considero que si una de las tres respuestas es correcta se consideraba acierto, es decir se otorgaba 1 punto.

En la sección VI-B1 se detallan los resultados obtenidos con los algoritmos utilizados.

d) *Motor de búsqueda que combina embeddings y heurística:* Una vez definidos los dos mecanismos para ordenar las instancias que mejor resuelven una consulta realizada por un usuario se deben juntar en una sola clasificación.

No se encontró ninguna referencia en la literatura que combinara dos motores de búsqueda, uno de los cuales este basado en embeddings y otro en heurística, de manera que a continuación se propone la metodología que se utilizó para el presente trabajo.

Dicha metodología consiste en realizar los siguientes pasos: En primer lugar, normalizar (acotar entre 0 y 1) los resultados obtenidos con los dos métodos (la similitud ( $P_{IA}$ ) y el recuento de palabras ( $P_H$ )). En segundo lugar, definir un peso para cada uno de los métodos y combinarlos en una función matemática que combine los dos resultados normalizados junto a sus pesos. En el presente trabajo se ha realizado una combinación lineal de manera que la puntuación total se puede definir como se muestra en la Fórmula 2:

$$P_{total} = w_H P_H + w_{IA} P_{IA} \quad (2)$$

Donde  $P_{total}$  es la puntuación total,  $w_H$  es el peso del motor basado en heurística,  $P_H$  es la puntuación del motor basado en heurística,  $w_{IA}$  es el peso del motor basado en IA y  $P_{IA}$  es la puntuación obtenida con el motor basado en IA.

En tercer lugar, se deben probar varios pesos evaluándolos con la métrica acierto definida previamente (sección V-A3c).

e) *Evaluación extrínseca: Métrica para evaluar el rendimiento del motor de búsqueda:* Por otra parte, se puede evaluar de manera extrínseca si el motor de búsqueda funciona correctamente al testear toda la tecnología con estudiantes y preguntándoles cómo de fácil les ha resultado encontrar la información. En la sección V-B se detalla como se debe plantear el experimento para contrastar los resultados con el grupo de control y en la sección VI-B2 se muestran los resultados.



4) **Visualización:** El último componente de la metodología general planteada (véase Figura 2) es el Visualizador, el cual consiste en generar la información necesaria para que el contenido se envíe al usuario a través de la interfaz. Se puede utilizar un lenguaje como HTML para combinar imágenes o vídeos con texto y dotar a este último de formato.

Se ha planteado una metodología en la que siempre se le ofrece al usuario una infografía. Se deben definir al menos dos tipologías de infografía puesto que se plantea una interacción con el usuario que puede ser o bien ofrecerle el contenido de valor de la instancia y de sus instancias relacionadas (infografía (a) de la Figura 2) o bien ofrecerle una o varias instancias como resultado de su búsqueda (infografía (b) de la Figura 2).

5) **Interfaz:** En el caso de estudio de este trabajo se utilizó Telegram dado que cuenta con una librería para python bien documentada, pero se podría haber utilizado cualquier otra interfaz.

## B. Evaluación con usuarios

Típicamente, la metodología que se utiliza para analizar un ITS mediante un grupo de control es darles a este grupo el mismo contenido que a los estudiantes del grupo que utiliza el ITS, pero sin las funcionalidades de este. Es decir, se les da la información por escrito o en un vídeo o mediante algún otro método tradicional [5].

Idealmente se habría utilizado esta metodología pero no se hizo debido a que dada la magnitud de la tecnología que se desarrolló solo fueron implementadas instancias para cubrir una parte de dos temas del contenido de física de 2º de bachillerato. En este contenido solo había dos problemas implementados junto con sus correspondientes pistas, fragmentos de vídeo, ejemplos, teoría y fórmulas.

Ya que solo se tenía la posibilidad logística de experimentar con los estudiantes durante una hora, se consideró que si se daba esa cantidad de información en un documento al grupo control, estos estudiantes tendrían la ventaja de que debían buscar entre poco contenido y el grupo del ITS tendría la desventaja de que debía familiarizarse con una nueva herramienta.

Además, tal y como se ha expuesto en la sección I, la intención del desarrollo de esta tecnología es mejorar el rendimiento de los estudiantes que se encuentran ante una cantidad masiva de información desorganizada en internet.

Así pues, dado que el propósito de esta investigación es observar el efecto de utilizar una tecnología basada en la estructura y la implementación del contenido por expertos, se decidió realizar la metodología que se describe a continuación.

1) **Metodología propuesta:** En primer lugar, se deben hacer dos grupos estratificados según las notas. En segundo lugar, se les debe dar una prueba evaluable a ambos grupos. En tercer lugar, a un grupo se le da acceso al ITS (y se le niega la posibilidad de utilizar cualquier otra herramienta como Google o similar) y al otro grupo se le permite navegar libremente por internet para que sean buscadas allí las respuestas a las consultas (grupo control). En cuarto lugar, siguiendo parte de la metodología utilizada en [27], se les

pide a los usuarios que rellenen una encuesta de satisfacción donde se les pregunta sobre características que se quieran medir como por ejemplo si han encontrado el contenido que buscaban (evaluación extrínseca del motor de búsqueda). Por último, se puede complementar la información de las encuestas utilizando entrevistas para profundizar en los detalles, así como registrar la actividad de los usuarios mientras utilizaban el ITS.

La prueba evaluable puede ser un tipo test, un ejercicio o cualquier otra tarea que se pueda cuantificar. En el caso de estudio se les dio un problema de física del nivel de las pruebas de acceso a la universidad de Cataluña (PAU) (al que se le cambiaron los datos del enunciado y se modificó el contexto para que los estudiantes del grupo control no pudieran encontrar el ejercicio navegando por la red) y se definieron unas pautas de corrección muy similares a las que se siguen en las PAU, pero evaluando el ejercicio sobre 10 puntos para que los resultados fueran más fáciles de interpretar, aunque la métrica más importante es el tamaño del efecto, la cual solo muestra la diferencia de la distribución.

En el caso de estudio, se planteó el experimento para ser realizado en el colegio Salesians de Mataró, el cuál facilitó la posibilidad de realizar los experimentos en una clase de 1º de bachillerato (correspondiente a los estudiantes del bachillerato científico y tecnológico) y una clase de 4º de la ESO (correspondiente con los estudiantes que realizaban la asignatura de Física y Química) y se planteó un ejercicio distinto a cada uno tratando de que el nivel de dificultad se adaptara a sus conocimientos.

En la sección VI-C se muestran los resultados obtenidos con un total de 56 alumnos. Se medirá el tamaño del efecto comparando los estudiantes que prueben la tecnología con los del grupo control. Se espera observar un tamaño del efecto similar o superior al obtenido en el estado del arte, que es de 0,76.

## VI. RESULTADOS, EVALUACIÓN Y DISCUSIÓN

En esta sección se muestran los resultados obtenidos en cada una de las partes descritas en la metodología (sección V). En primer lugar se muestran los resultados de la ontología (sección VI-A) que consta de una evaluación de 30 preguntas de competencia (análisis cualitativo) y el cálculo de varias métricas cuantitativas. A continuación se muestran las evaluaciones de los motores de búsqueda (sección VI-B). Por último se muestra la evaluación con estudiantes del ITS propuesto en esta investigación (sección VI-C).

Referente a la evaluación con los usuarios, ya que el sistema ITS tiene un formato similar a un chatbot se le puso el nombre de Sofia para que la interacción fuera más amigable. Se puede acceder a dicho chatbot desde el buscador de Telegram escribiendo *Sofia3FisicaBot*, pero el servidor que sostiene la infraestructura está en un ordenador local y el chatbot solo está operativo cuando el ordenador también lo está.

A partir de ahora, cuando se haga referencia al grupo que interactuó con Sofia se le llamará grupo Sofia para diferenciarlo del grupo de control.

## A. Evaluación de la ontología

### 1) Métricas cualitativas: Preguntas de competencia:

Las preguntas de competencia son consultas realizadas en SPAQRL sobre la ontología que sirven para validar que esta está bien estructurada. Se realizaron 30 preguntas que permitieron validar dicha estructura. En la Tabla V se pueden ver dos ejemplos.

Tal y como se comenta en la sección V-A1, en cada una de las tres iteraciones se utilizaron las preguntas de competencia para identificar (y posteriormente corregir) errores y así validar la usabilidad de la ontología.

2) Métricas cuantitativas: En las Tablas II y IV se muestran un conjunto de métricas que sirvieron para tener un conocimiento más profundo de la ontología. Algunas métricas han sido calculadas manualmente y otras se calculan automáticamente con el programa Protégé, el cual ha sido utilizado para construir la ontología (véase sección V-A1d). Los cálculos de las métricas manuales se ha basado en los realizados en [26].

Los resultados de la Tabla II muestran que la ontología presenta una cantidad notable de axiomas que se distribuyen en diferentes categorías, es decir, clases, propiedades e instancias.

Tabla II  
MÉTRICAS CALCULADAS MEDIANTE PROTEGE Y MANUALMENTE

Métrica	Resultado
Protege	
Axiomas	2447
Recuento de axiomas lógicos	2067
Recuento de clases	16
Recuento de propiedades de objeto	25
Recuento de propiedades de datos	18
Recuento de instancias	296
Recuento de subclases	4
Manual	
Relaciones Taxonómicas	29
Otras Relaciones No-Taxonómicas	36
Clases Raíz N.	3
Clases Intermediarias N.	2
Clases de hojas N.	11
Relaciones Equivalentes N.	0
Clases Rehusadas	5
Propiedades de Objeto Rehusadas	6
Propiedades de Datos Rehusadas	11

Los resultados de la Tabla IV muestran que hay un equilibrio entre relaciones taxonómicas (es decir, jerárquicas) y no taxonómicas. También se puede ver que se ha hecho un buen reuso de otras ontologías pues en promedio 2,38 de cada tributos de una clase son reutilizados. Otro elemento a destacar es que hay 4,69 relaciones por clase, es decir que es una ontología que cuenta con bastantes relaciones.

## B. Evaluación de los motores de búsqueda

1) Evaluación intrínseca de los motores de búsqueda: Por un lado, se ha construido un motor de búsqueda basado en heurística cuyo funcionamiento se ha explicado en la sección V-A3b.

Por otro lado, se han utilizado modelos de embeddings basados en *transformers* que han sido comentados en la sección V-A3a.

Los resultados de cada modelo de manera individual se muestran en la Tabla III. El sistema de puntuación se ha detallado en la sección correspondiente de la metodología (véase la sección V-A3c), el cual básicamente consiste en el cociente entre el número de consultas bien respondidas y el número total de consultas (véase Fórmula (1)). Se definieron un total de 31 consultas.

Tabla III  
EVALUACIÓN INTRÍNSECA DE LOS MOTORES DE BÚSQUEDA. LA MÉTRICA ACIERTO ES EL COCIENTE ENTRE EL NÚMERO DE CONSULTAS BIEN RESPONDIDAS Y EL NÚMERO TOTAL DE LAS 31 CONSULTAS QUE FUERON DEFINIDAS

modelo	Acierto
bert-base-multilingual-cased	0,10
paraphrase-xlm-r-multilingual-v1	0,55
paraphrase-multilingual-mpnet-base-v2	0,71
contador de palabras	0,84
paraphrase-multilingual-mpnet-base-v2 + contador de palabras	0,87

Se puede observar que el mejor modelo es el contador de palabras, basado en heurística. Se sospecha que se ha obtenido este resultado porque el dominio era pequeño y las anotaciones semánticas (que son los textos de los atributos título, descripción y keywords) han sido bien descritas, pero se prevé que en un dominio más amplio un motor de búsqueda basado en *transformers* dará mejores resultados.

De todas formas, se observó que algunas de las consultas en las que fallaba el motor de búsqueda basado en heurística, el modelo *paraphrase-multilingual-mpnet-base-v2* (*mpnet*) acertaba. De manera que se probó de combinar ambos buscadores.

El cómo realizar dicha combinación ha sido descrito al detalle en el apartado de metodología (sección V-A3d), donde se ha expuesto que se utilizaría el sistema de puntuación con aquellos pesos que maximizaran la Fórmula 2.

Tras probar las siguientes combinaciones de pesos (1, 1), (2, 1), (3, 1), (1, 1.5), (1, 2) y (1, 3) donde la primera coordenada es  $w_{IT}$  y la segunda es  $w_{mpnet}$  se llegó a la conclusión de que el mejor resultado era combinar el motor basado en heurística y el modelo de machine learning con una combinación de pesos (1, 2), dando lugar así a una puntuación de 0,87, tal y como se muestra en la Tabla III.

2) Evaluación extrínseca del motor de búsqueda: Evaluación con usuarios: Tal y como se ha comentado en la sección V-A3e, la evaluación extrínseca consistió en preguntarles tanto al grupo Sofía como al grupo control como de comprendidos se habían sentido. Dado que se buscaba ver la efectividad del motor de búsqueda la pregunta al grupo Sofía se realizó de la siguiente forma: "Si has hecho al menos una pregunta, valora del 1 al 10 cómo de comprendido te has sentido"<sup>18</sup> y al grupo control se le preguntó: "Del 1 al 10, ¿Cómo de fácil te ha sido encontrar un respuesta apropiada buscándola libremente por internet?"

Los resultados se muestran en la Figura 7 y se puede ver claramente que los usuarios percibieron que encontraban el

<sup>18</sup>El 90% de los usuarios hicieron al menos una pregunta. Se hace esta distinción porque los usuarios también podían navegar el contenido mediante los botones sin hacer uso del buscador.

Tabla IV  
MÉTRICAS OBJETIVAS DE LA ONTOLOGÍA PLO

Métrica	Fórmula	Resultado
Subclases (promedio de subclases)	$\frac{\sum_{i=1}^c s(i)}{c}$	0,95 [0,15]
Relaciones taxonómicas por clase	$\frac{\sum_{i=1}^c r(i)}{c}$	2,06 [0,5]
Relaciones no taxonómicas por clase	$\frac{\sum_{i=1}^c r_{not}(i)}{c}$	2,19 [0,10]
Relaciones semánticas reutilizadas por clase	$\frac{\sum_{i=1}^c reused\_rel(i)}{c}$	0,19 [0,3]
Atributos reutilizados por clase	$\frac{\sum_{i=1}^c reused\_prop(i)}{c}$	2,38 [0,5]
Profundidad promedio de herencia por clase	$\frac{\sum_{i=1}^c max(path(i))}{c}$	2,38 [0,5]
Densidad de las propiedades	$\frac{\sum_{i=1}^c n\_att(i) + n\_rel(i)}{c}$	2,69 [0,5]
Densidad de la herencia	$\frac{\sum_{i=1}^c sc(i)}{c}$	0,25 [0,4]
Densidad de las relaciones	$\frac{\sum_{i=1}^c tax\_rel(i) + sem\_rel(i)}{c}$	4,69 [1,13]

contenido de manera más fácil utilizando Sofía que navegando por internet. Concretamente, la media de la valoración de los usuarios del grupo Sofía fue de un 6,66 sobre 10 en frente de un 4,67 sobre 10 del grupo control, suponiendo una diferencia de 1,99 puntos.

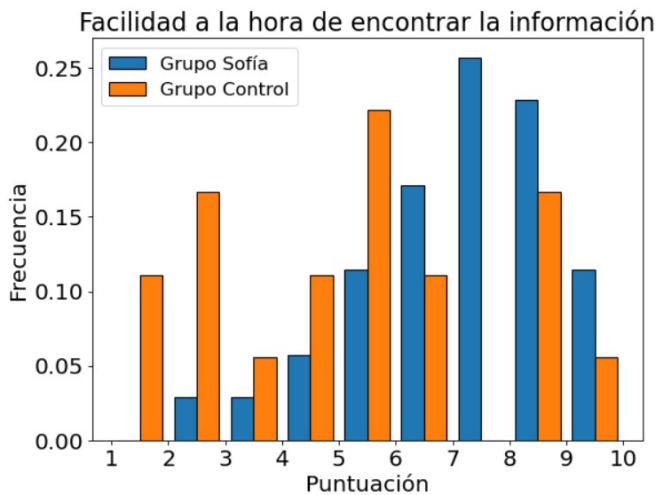


Fig. 7. Distribución de la valoración que dieron los estudiantes del grupo control (naranja) y el grupo Sofía (azul) a la pregunta de cómo de fácil les había resultado encontrar la información navegando por internet (grupo control) o utilizando a Sofía (naranja).

### C. Evaluación del experimento realizado con usuarios

En esta sección se muestra el análisis de los resultados del ejercicio que resolvieron los estudiantes con la ayuda de Sofía o de la búsqueda libre por internet. Se puede ver una imagen del funcionamiento de Sofía en las imágenes (a) y (b) de la Figura 2 (esquina superior derecha).

Tal y como se ha explicado en la sección V-B, donde se detalla la metodología seguida para realizar el experimento, la métrica principal es el tamaño del efecto (descrito en la sección IV-B), la cual se puede resumir como una métrica que cuantifica la diferencia en la distribución de las notas obtenidas al realizar una tarea evaluable.

Se observó que al segmentar los estudiantes en función de la nota media del curso de la asignatura de física, aquellos estudiantes que tenían una media más baja mostraron ser mucho más efectivos en la resolución del problema planteado al usar a Sofía que el grupo control. Por esta razón, primero se muestran los resultados generales y a continuación los resultados al aplicar esta segmentación.

La muestra de estudiantes fue de 56 alumnos, 30 de los cuales de 4º de la ESO y 26 de 1º de bachillerato. De estos 56, 29 pertenecieron al grupo Sofía y 27 al grupo control. Los estudiantes de 4º de la ESO resolvieron un ejercicio del tema de órbitas gravitatorias y los de 1º de bachillerato del átomo de Bohr.

1) *Resultados generales:* En la Tabla VI se muestran las notas obtenidas por los estudiantes de 4º de la ESO y 1º de Bachillerato, tanto del grupo Control como del grupo Sofía. Se puede observar que la nota media está por debajo de 5 en ambos grupos. La nota del ejercicio era sobre 10. Se atribuyen unas notas tan bajas debido a que se les pidió que resolvieran un ejercicio que estaba muy por encima de sus conocimientos (los ejercicios eran de selectividad). No obstante, el objetivo no era que aprendieran a resolver el ejercicio que se les planteó sino observar la diferencia entre el grupo control y el grupo Sofía.

Referente a este último punto, se puede observar que en 4º de la ESO la nota media es muy similar y el tamaño de efecto muy cercano a 0. Se recuerda que los ITSs del estado del arte tienen un promedio de un tamaño de efecto de  $d = 0,76$  [5].

En el caso de 1º de bachillerato se puede observar que hay una diferencia de 0,73 puntos en la nota media y que el tamaño del efecto es de 0,19, que sigue estando por debajo del estado del arte.

En la primera columna de la Figura 10 se muestra la distribución de notas que se resumen en la Tabla VI que acaban de ser comentadas. Se puede observar que las distribuciones de notas son muy similares entre el grupo control y el grupo Sofía, tal y como refleja el valor de la métrica tamaño del efecto, cercana a 0.

Tabla V  
EJEMPLOS DE CONSULTAS A LA ONTOLOGÍA EN SPARQL

Consulta	¿Qué ejercicios están relacionados con el Periodo?			
Tipo de usuario	Profesor o estudiante			
SPARQL	<pre> PREFIX schema: &lt;https://vocab.org/aiiso/schema/&gt; PREFIX dc: &lt;https://dublincore.org/specifications/dublin-core/dcmi-terms/&gt; PREFIX dbpedia: &lt;https://dbpedia.org/resource/&gt; PREFIX physics: &lt;http://physics.kg.ac.rs/physics/2013/03&gt; PREFIX purl: &lt;http://purl.org/vocab/aiiso/schema&gt; PREFIX dcterms: &lt;http://purl.org/dc/terms/&gt; PREFIX schema1: &lt;https://schema.org/&gt; PREFIX dma: &lt;https://w3id.org/media/dma&gt; PREFIX plo: &lt;http://www.semanticweb.org/plo&gt; PREFIX rdf: &lt;http://www.w3.org/1999/02/22-rdf-syntax-ns&gt; SELECT DISTINCT ?ejercicio WHERE{ ?ejercicio ?tienePista ?clusterPistas. ?clusterPistas ?tienePista ?pistas. ?pistas ?relacionadoConTeoria ?teoria. ?ejercicio rdf:type plo:Ejercicio. ?teoria rdf:type plo:Teoria. FILTER regex(str(?teoria), 'Periodo')} </pre>			
Resultado	ind_Ejercicio_1 ind_Ejercicio_2 ind_Ejercicio_3			
Consulta	¿Cuál es el frame del video y la url para la teoría de la aceleración centrípeta?			
Tipo de usuario	Profesor o estudiante			
SPARQL	<pre> PREFIX schema: &lt;https://vocab.org/aiiso/schema/&gt; PREFIX dc: &lt;https://dublincore.org/specifications/dublin-core/dcmi-terms/&gt; PREFIX dbpedia: &lt;https://dbpedia.org/resource/&gt; PREFIX physics: &lt;http://physics.kg.ac.rs/physics/2013/03&gt; PREFIX purl: &lt;http://purl.org/vocab/aiiso/schema&gt; PREFIX dcterms: &lt;http://purl.org/dc/terms/&gt; PREFIX schema1: &lt;https://schema.org/&gt; PREFIX dma: &lt;https://w3id.org/media/dma&gt; PREFIX plo: &lt;http://www.semanticweb.org/miquel&gt; PREFIX rdf: &lt;http://www.w3.org/1999/02/22-rdf-syntax-ns&gt; SELECT DISTINCT ?url ?linkStart ?linkEnd WHERE{ ?teoria dcterms:title ?tituloTeoria. ?teoria ?tieneContenidoDeAprendizajeAudiovisual ?contenidoDeAprendizajeAudiovisual. ?contenidoDeAprendizajeAudiovisual dma:linkStart ?linkStart. ?contenidoDeAprendizajeAudiovisual dma:linkEnd ?linkEnd. ?contenidoDeAprendizajeAudiovisual ?tieneCD ?contenidoDigital. ?contenidoDigital schema1:url ?url. FILTER regex(str(?tituloTeoria), 'Aceleración Centrípeta')} </pre>			
Resultado		url	linkStart	linkEnd
		https://www.youtube.com/watch?v=e1ZqoJvUNCs	777	1149

Tabla VI  
RESULTADOS DE LOS EXPERIMENTOS. SE MUESTRA LA NOTA MEDIA QUE OBTUVIERON AL RESOLVER EL EJERCICIO QUE SE LES PROPUSO Y EL TAMAÑO DEL EFECTO DEL GRUPO SOFÍA RESPECTO AL GRUPO DE CONTROL

Nivel académico	Grupo	Nota media	Tamaño del efecto (d)
4º ESO	Control	2,63	0,07
	Sofía	2,79	
1º BACH	Control	3,8	0,19
	Sofía	4,53	

2) *Resultados segmentados por la nota media del curso en la asignatura de física:* Pese a que aparentemente los resultados obtenidos están por debajo del estado del arte, si se realiza un análisis más profundo haciendo una segmentación entre los alumnos que tenían una nota promedio de la asignatura alta o baja, se observa una mejora significativa del rendimiento en aquellos estudiantes que partían de notas medias más bajas.

El criterio para definir si una nota es alta o baja ha sido hacer dos grupos del mismo tamaño cada uno y considerar la nota intermedia como el umbral.

En la Tabla VII se muestra solamente el tamaño del efecto y se puede observar que en 4º de la ESO se obtiene un valor de 2,30 para los estudiantes que tenían una nota media baja. En el caso de 1º de bachillerato se puede observar un tamaño del efecto de 1,44. Ambos resultados están por encima del tamaño del efecto promedio del estado del arte que es de 0,76. Triplicando este resultado en el caso de 4º de la ESO y duplicándolo en el caso de 1º de bachillerato.

Por otra parte, se puede observar que de los estudiantes que tenían notas medias altas obtuvieron peores resultados aquellos que utilizaron a Sofía en comparación al grupo de control, ya que el tamaño del efecto es negativo, lo que significa que la distribución de las notas se ha desplazado hacia la izquierda, es decir, notas más bajas.

En la Figura 10 se muestran los resultados descritos en esta

sección y resumidos en la Tabla VII. En la columna central se pueden observar las distribuciones de las notas del ejercicio de los alumnos que tenían una nota media alta del curso y en la columna de la derecha se pueden observar las distribuciones de los alumnos que tenían una nota media baja. Se puede observar en esta última columna como claramente las notas se han desplazado hacia la derecha.

Tabla VII  
RESULTADOS DE LOS EXPERIMENTOS AL SEGMENTAR LOS ESTUDIANTES EN FUNCIÓN DE LA NOTA MEDIA DEL CURSO DE LA ASIGNATURA DE FÍSICA. SE MUESTRA EL TAMAÑO DEL EFECTO DEL GRUPO SOFÍA RESPECTO AL GRUPO DE CONTROL

Nivel académico	Nota promedio del curso	Tamaño del efecto ( $d$ )
4º ESO	Baja	<b>2,30</b>
	Alta	-0,28
1º BACH	Baja	<b>1,44</b>
	Alta	-0,04

3) *Entrevistas, encuestas con los usuarios y registro de su actividad:* Tal y como se ha explicado en V-B1, con la finalidad de complementar los datos obtenidos se realizó una encuesta a todos los usuarios y un total de 20 entrevistas preguntándoles sobre el efecto que habían experimentado en su rendimiento al utilizar Sofía para resolver el ejercicio propuesto.

La información obtenida mediante las encuestas reflejó que las características más valoradas por los estudiantes fueron las siguientes (se les ofreció una selección múltiple en la que podían escoger hasta 10 elementos): Un 81% de se sentía muy satisfecho teniendo una herramienta que les ofreciera fórmulas rápidamente y bien organizadas, un 68% valoró positivamente las pistas para resolver los ejercicios, un 54% que hubiera siempre contenido relacionado y un 40% poder preguntar libremente.

Mediante las entrevistas y un análisis de la actividad de los usuarios (puesto que fue registrada en un documento xls) se observó que los estudiantes de 4º de la ESO que obtuvieron mejores notas usando a Sofía estuvieron más tiempo viendo ejemplos, tal como muestra la Figura 8. En cambio, en 1º de bachillerato (véase Figura 9), la característica destacada entre los alumnos que mejor resolvieron el ejercicio fueron las pistas.

Por último, cabe destacar que varios alumnos manifestaron haberse sentido como si tuvieran un profesor particular que les acompañaba en el proceso y sentían que esta herramienta les resultaba útil. De hecho, el 86% de los usuarios manifestaron que les gustaría que su colegio les ofreciera una herramienta de estas características y un 14% opinó que tal vez. Nadie considero que no quería esta tecnología.

## VII. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo se ha demostrado que una tecnología de tutorización educativa basada en recursos multimedia semánticos accesibles desde un buscador y que además esta ofrezca siempre al usuario un contenido adicional relacionado con su exploración y/o pregunta en lenguaje natural mejora el rendimiento de los estudiantes que aparentemente tienen más dificultades, es decir, aquellos que tienen una nota media del

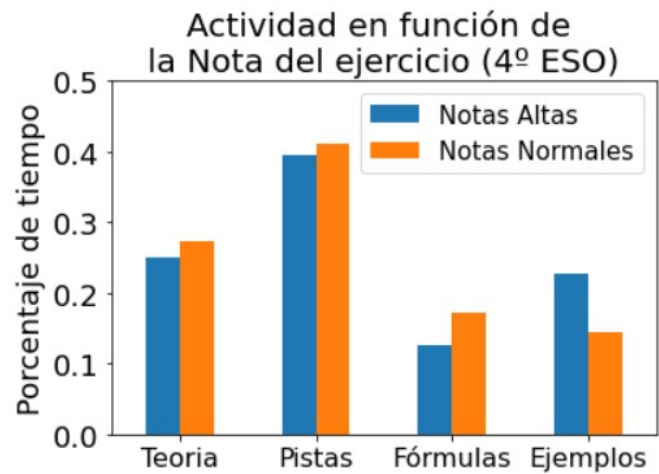


Fig. 8. Porcentaje de tiempo promedio que dedicaron los estudiantes de 4º de la ESO a cada uno de los elementos de la ontología que contenían contenido. Se puede ver que, en promedio, aquellos que dedicaron más tiempo a los ejemplos resolvieron mejor el ejercicio.

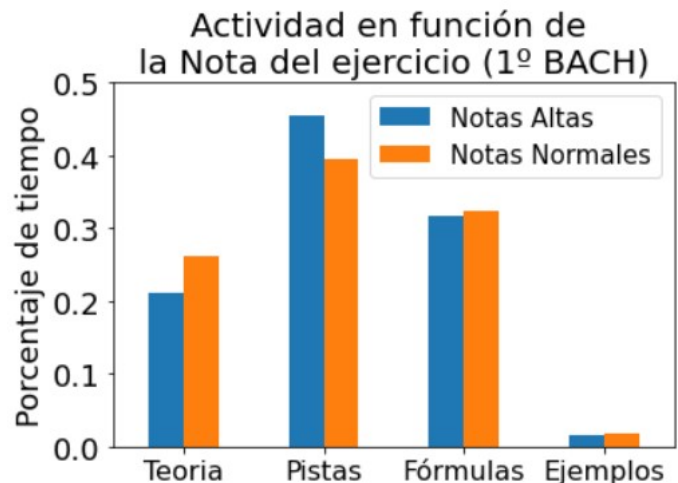


Fig. 9. Porcentaje de tiempo promedio que dedicaron los estudiantes de 1º de bachillerato a cada uno de los elementos de la ontología que contenían contenido. Se puede ver que, en promedio, aquellos que dedicaron más tiempo a los ejemplos resolvieron mejor el ejercicio.

curso baja en comparación con sus compañeros. La evidencia se encuentra en los resultados que se muestran en la Tabla VII donde se tiene un tamaño de efecto de 2,30 para los estudiantes de 4º de la ESO y de 1,44 para los de 1º de bachillerato. Obteniendo así un resultado, en ambos casos, superior al promedio del estado del arte de tecnologías similares donde se obtiene un 0,76 [5].

También se ha demostrado que en el dominio de la ontología desarrollada, un motor de búsqueda que combine un motor basado en heurística con un motor de búsqueda basado en *transformers* funciona mejor que uno que solo utilice una de estas dos tecnologías. En la Tabla III se puede ver como se consigue un acierto de 0,87 al utilizar las dos tecnologías conjuntamente en vez de un 0,84 de utilizar solo la mejor de ellas. Estos resultados se han visto corroborados por la evaluación de los usuarios, los cuales han sentido comprendidos con una



## Distribución de notas del ejercicio del experimento

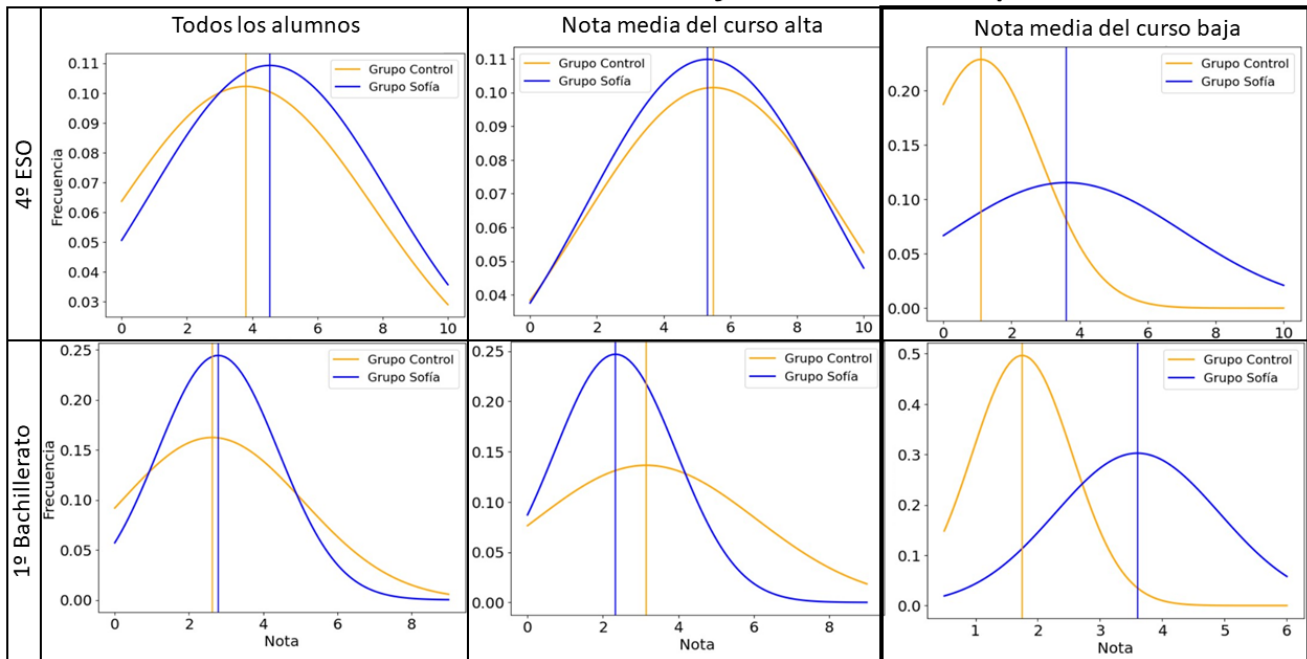


Fig. 10. Distribución gaussiana de las notas de los estudiantes obtenidas al realizar el ejercicio que se les pidió para testear la tecnología propuesta en este trabajo. La línea vertical muestra la media. En color azul se muestra el grupo Sofía y en naranja el grupo control. Los resultados se muestran segmentados en función de la nota media del curso escolar y del curso académico. En la primera fila se encuentran los datos de los alumnos de 4º de la ESO y en la segunda la de los alumnos de 1º de bachillerato. En la primera columna se muestra el conjunto de todos los estudiantes. En la segunda columna el conjunto de los estudiantes que tenían una nota media alta del curso escolar en la asignatura de Física y en la tercera el conjunto de los estudiantes que tenían una nota media baja. Se puede observar que en la tercera columna (gráficos remarcados con un marco más grueso) hay un aumento en el rendimiento al utilizar Sofía en aquellos alumnos que tenían una nota media baja. La distribuciones gaussianas se han realizado utilizando 100 compartimentos en cada una de ellas.

valoración de 6,66 en comparación con un 4,67 que es la valoración promedio que ha dado el grupo de control a como de comprendidos se han sentido buscando la información en internet.

Por lo tanto, se puede afirmar que una tecnología de tutorización educativa basada en recursos multimedia semánticos dotada de un buscador y un sistema que siempre ofrezca contenido relacionado ayuda a la tutorización docente.

De cara a futuros trabajos se plantea la posibilidad de ampliar el marco propuesto en la Figura 2 con un componente Estudiante, el cual almacene la información del estudiante y utilice esos datos para entrenar un sistema de recomendación que combinado con el motor de búsqueda ofrezca un contenido personalizado a cada estudiante.

### AGRADECIMIENTOS

Me gustaría dedicar mi más sincero agradecimiento a mi tutor, el Dr. Yoan Gutierrez, por haber sido un guía impecable desde el primer día hasta el último. Espero que lo iniciado en este trabajo no finalice con este proyecto. También quisiera agradecer a mi familia y amigos por su apoyo constante e incansable. Añadir un especial agradecimiento a las profesoras Càrol Estudillo y Cristina Salas por la orientación en la creación de la ontología y el soporte logístico brindado, juntamente con la directiva del colegio Salesians de Mataró, donde se realizaron los experimentos.

### REFERENCES

- [1] Ammar Y. Alqahtani, Albraa A. Rajkhan, *E-Learning Critical Success Factors during the COVID-19 Pandemic: A Comprehensive Analysis of E-Learning Managerial Perspectives*, Education Sciences, 2020
- [2] Shivangi Dhawan, *Online Learning: A Panacea in the Time of COVID-19 Crisis*, Journal of Educational Technology Systems, Vol. 49(1) 5–22, 2020
- [3] J. M. Gascueña et al. *Domain Ontology for Personalized E-Learning in Educational Systems*, Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06), 2006
- [4] K. VanLehn, *The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems*, Educational Psychologist, 46:4, 197–221, 2011
- [5] J. Kulik and J. D. Fletcher, *Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review*, Review of Educational Research, 86, 2015
- [6] Jacqueline Riichi Mizoguchi. *Using Ontological Engineering to Overcome Common AI-ED Problems*, International Journal of Artificial Intelligence in Education, 11, 107–121, 2000
- [7] Michelene T.H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, Robert G. Hausmann, *Learning from human tutoring*, Cognitive Science 25, 471–533, 2001
- [8] Benjamin S. Bloom, *The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring*, Educational Researcher, Vol. 13, No. 6, pp. 4–16, 1984
- [9] D. Schiff, *Out of the laboratory and into the classroom: the future of artificial intelligence in education*, AI Society, 2021
- [10] W. Holmes, M. Bialik and C. Fadel, *Artificial Intelligence in Education*, Center for Curriculum Redesign
- [11] T. Mikolov et al., *Distributed Representations of Words and Phrases and their Compositionality*, Google Inc. <https://arxiv.org/abs/1310.4546>
- [12] Md Tahmid Rahman Laskar, Jimmy Huang, Enamul Hoque, *Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task*, Proceedings of the 12th Conference on Language Resources and Evaluation, pages 5505–5514, 2020

- [13] SentenceTransformers Documentation, *Pretrained Models* [https://www.sbert.net/docs/pretrained\\_models.html](https://www.sbert.net/docs/pretrained_models.html)
- [14] IBM, *Introduction to business system development method*, Tech. Rep. GE19-5387-01, International Business Machines Corporation, 1990.
- [15] D. Tansley, C. Hayball, *Knowledge-Based Systems Analysis and Design: A Kads Developer's Handbook*, British Computer Society practitioner series, Prentice Hall, 1993.
- [16] KBSI, *Knowledge based systems incorporated*, Tech. rep., Wright-Patterson Air Force Base, Ohio, 1994.
- [17] M. Fernández-López, A. Gómez-Pérez, N. Juristo, *Methontology: From ontological art towards ontological engineering*, in: *Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series*, American Association for Artificial Intelligence, pp. 33–40, 1997
- [18] A. Gómez-Pérez, Ontology evaluation, in: *Handbook on Ontologies*, Springer, pp. 251–273, 2004
- [19] S. Tartir, I.B. Arpinar, M. Moore, A.P. Sheth, B. Aleman-Meza, *OntoQA: Metric-Based Ontology Quality Analysis*, IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources, pp. 45–53, 2005
- [20] H. Yao, A.M. Orme, L. Etzkorn, Cohesion metrics for ontology design and application, *J. Comput. Sci.* 1 (1) 107–113, 2005
- [21] E. Blomqvist, A. Öhgren, K. Sandkuhl, *Ontology Construction in an Enterprise Context: Comparing and Evaluating Two Approaches*, ICEIS (3), pp. 86–93, 2006
- [22] V. Cross, A. Pal, *An ontology analysis tool*, *Int. J. Gen. Syst.* 37 (1) 17–44, 2008
- [23] E. Sirin, B. Parsia, *SPARQL-DL: SPARQL query for OWL-DL*, in: *CEUR Workshop Proceedings*, vol. 258, 2007
- [24] Chris Edwards, *The best of NLP*, Communications of the ACM, Volume 64, Issue 4, 2021
- [25] R. Karwa V. Honmane, *Building Search Engine Using Machine Learning Technique*, International Conference on Intelligent Computing and Control Systems (ICCS), 2019
- [26] Y. Gutiérrez, D. Tomás and I. Moreno, *Developing an ontology schema for enriching and linking digital media assets*, Future Generation Computer Systems, 2019
- [27] Hei-Chia Wang, Chien-Wei Hsu, *Teaching-Material Design Center: An ontology-based system for customizing reusable e-materials*, *Computers Education* 46, 458–470, 2006