GENERACIÓN AUTOMÁTICA DE QUIZZES PARA MUSEOS EMPLEANDO TÉCNICAS DE DEEP LEARNING

AUTOMATIC GENERATION OF QUIZZES FOR MUSEUMS USING DEEP LEARNING TECHNIQUES

TRABAJO FIN DE MÁSTER
CURSO 2020-2021

AUTOR
ALLINSON DE LAS NIEVES BORROTO FONSECA

DIRECTOR
PEDRO ANTONIO GONZÁLEZ CALERO

MÁSTER EN INTERNER DE LAS COSAS
FACULTAD DE INFORMÁTICA
UNIVERSIDAD COMPLUTENSE DE MADRID

GENERACIÓN AUTOMÁTICA DE QUIZZES PARA MUSEOS EMPLEANDO TÉCNICAS DE DEEP LEARNING

AUTOMATIC GENERATION OF QUIZZES FOR MUSEUMS USING DEEP LEARNING TECHNIQUES

TRABAJO DE FIN DE MÁSTER EN INTERNET DE LAS COSAS
DEPARTAMENTO DE INGENIERÍA DEL SOFTWARE E INTELIGENCIA ARTIFICIAL

AUTOR
ALLINSON DE LAS NIEVES BORROTO FONSECA

DIRECTOR
PEDRO ANTONIO GONZÁLEZ CALERO

CONVOCATORIA: JUNIO 2021
CALIFICACIÓN: 8

MÁSTER EN INTERNER DE LAS COSAS
FACULTAD DE INFORMÁTICA
UNIVERSIDAD COMPLUTENSE DE MADRID

7 DE JULIO DE 2021

DEDICATORIA

A mis padres,

A la memoria de mis abuelos.

AGRADECIMIENTOS

A Pedro, por su ayuda y apoyo durante toda la etapa de desarrollo de este trabajo.

RESUMEN

Generación automática de quizzes para museos empleando técnicas de Deep Learning.

En este trabajo se presenta el diseño de un algoritmo que permite obtener respuestas incorrectas a preguntas de donde se conoce la contestación correcta, empleando tecnologías de Web Semántica y técnicas de Deep Learning. Se plantea como entorno de explotación un minijuego, que se encuentra desplegado en el Museo de Ciencias Naturales de Madrid. El juego se basa en las áreas de las exposiciones como la evolución de la vida en la Tierra, que abarcan desde los primeros microorganismos hasta el Homo Sapiens, por medio de colección de fósiles, esqueletos, reconstrucciones e ilustraciones que recrean la vida en la Tierra en sus diferentes eras o etapas.

Sobre esta base, se plantea la posibilidad de ampliar la programación del juego, en cuanto a un algoritmo que permita definir cuestionarios de tipo pregunta de opción múltiple, en donde habrá una respuesta correcta y tres incorrectas, con posibilidad de establecer niveles de conocimiento. La experiencia irá dirigida a famosos paleontólogos y sitios paleontológicos. Para esto, se emplean tecnologías de la Web Semántica como DBpedia y su variante SpotLigh, Wikidata, Wikipedia, Yago y técnicas de Deep Learning como es el caso de Word2Vect.

Palabras clave

Reconocimiento de Entidades Nombradas, Word2Vect, DBpedia, DBpedia SpotLigh, Yago Wikicats, Museo Nacional de Ciencias Naturales.

ABSTRACT

Automatic generation of quizzes for museums using Deep Learning techniques

This paper presents the design of an algorithm that allows to obtain incorrect answers to questions where the correct answer is known using Semantic Web technologies and Deep Learning techniques. A mini-game, which is deployed in the Museum of Natural Sciences of Madrid, is proposed as an operating environment. The game is based on the areas of exhibitions such as the evolution of life on Earth, ranging from the first microorganisms to Homo Sapiens, through a collection of fossils, skeletons, reconstructions, and illustrations that recreate life on Earth in its different eras or stages.

On this basis, the possibility of extending the programming of the game, regarding the methodology for defining questionnaires question multiple choice type, where there will be a correct answer and three incorrect arises. The experience will be aimed at famous paleontologists and paleontological sites. For this, Semantic Web technologies are used as DBpedia and its variant Spotligh, Wikidata, Wikipedia, Yago and Deep Learning techniques as in the case of Word2Vect.

Keywords

Named Entity Recognition, Word2Vect, DBpedia, DBpedia SpotLigh, Yago Wikicats, National Museum of Natural Sciences.

ÍNDICE DE CONTENIDOS

Dedicatoria III

Agradecimientos V

Resumen VII

Abstract IX

Índice de contenidos X

Índice de figuras XII

Índice de tablas XV

Capítulo 1 - Introducción 1

1.1 Planteamiento General 1

1.2 Motivación 2

1.3 Objetivos 3

1.4 Plan y estructura de trabajo 4

Capítulo 2 - Descripción de métodos y tecnologías aplicadas 7

2.1 DBpedia 7

2.2 DBpedia SpotLight 9

2.3 Word2Vect 11

2.4 WikiData 13

2.5 Stanza 14

2.6 Estudio de trabajos relacionados 15

Capítulo 3 - Obtención de distractores 18

3.1 Obtención del corpus 18

3.1.1 Identificación de títulos de páginas de Wikipedia. 20

3.1.2 Filtrado de las entidades. 24

3.1.3 Procesado del corpus....25
3.2 Entrenamiento de la red neuronal Word2Vect....26
3.3 Adquisición de distractores....27
3.3.1 Obtención de candidatos....27
3.3.2 Filtrado de los candidatos a distractores....29
3.3.3 Agrupación por niveles de dificultad....31
3.3.4 Sugerencia de nuevos cuestionarios....35
3.4 Presentación de ejemplos y análisis de resultados....37
Capítulo 4 - Conclusiones....45
Capítulo 5 -....49
General Planning....49
Motivation....50
Objectives....51
Work plan and structure....52
Capítulo 6 -....55
Bibliografía....57
Apéndices....63

ÍNDICE DE FIGURAS

Figura 2-1Ilustración de la arquitectura actual de suministro de datos de DBpedia [12]. 8

Figura 2-2 Representación gráfica del modelo CBOW y del modelo Skip-Gram [16]. 12

Figura 2-3 Propagación hacia adelante de la red neuronal en Word2Vect [17]. 13

Figura 2-4 Diagrama de un elemento de Wikidata [18]. 14

Figura 2-5 Descripción general de la canalización de NPL de la red neuronal de Stanza [19]. 15

Figura 3-1 Interfaz de usuario gráfica, Menú Inicial. 19

Figura 3-2 Texto introducido por el experto. 20

Figura 3-3 Selección de Categorías para la recopilación del corpus. 22

Figura 3-4 Selección del umbral de similitud. 23

Figura 3-5 Presentación de un ejemplo de un cuestionario. 28

Figura 3-6 Candidatos obtenidos en el proceso para el cuestionario número 2 de ejemplo. 30

Figura 3-7 Ejemplo para la lematización y aplicación de los POS TAG. 32

Figura 3-8 Ejemplo de representación de las distancias de dos documentos [33]. 33

Figura 3-9 Ejemplo de aplicación del método wmdistance entre oraciones [34]. 34

Figura 3-10 Resultados para un cuestionario de dos niveles. 35

Figura 3-11 Resultados para un cuestionario de tres niveles. 35

Figura 3-12 Opción de nuevos cuestionarios para complementar la aplicación, ejemplo para Edwin H. Colbert. 36

Figura 3-13 Opción de nuevos cuestionarios para complementar la aplicación, ejemplo para Othniel Charles March. 37

Figura 3-14 Primeros 10 distractores para el cuestionario número dos sobre paleontólogos. 42

Figura 3-15 Primeros 10 distractores para el cuestionario número uno sobre sitios de relevancia paleontológica. 43

ÍNDICE DE TABLAS

Tabla 2-1 Ejemplo de descripción de entidad devuelta por DBpedia SpotLight. 10

Tabla 3-1 Modelo de consulta a DBpedia, opción annotate. 21

Tabla 3-2 Consultas SparQL de @types y @subjects a DBpedia. 21

Tabla 3-3 Consulta Wikidata para obtener títulos de páginas de Wikipedia relacionadas a la categoría. 23

Tabla 3-4 Estadística para los cuestionarios sobre paleontólogos. 39

Tabla 3-5 Estadística para el cuestionario sobre sitios destacados para la paleontología. 41

Capítulo 1 - Introducción

1.1 Planteamiento General

La utilización de las nuevas tecnologías es clave para mejorar la experiencia del usuario en los museos, así como aprovechar y explotar oportunidades en nuevas formas de actividades culturales interactivas. Las instituciones de la memoria necesitan mantener, por no decir reforzar, su atractivo y el interés de sus visitantes, especialmente para las nuevas generaciones.

Según los Estatutos del Consejo Internacional de Museos (ICOM), adoptados por la 22a Asamblea General en Viena, Austria, el 24 de agosto de 2007, la definición de museo, es la siguiente, "es una institución permanente, sin ánimo de lucro, al servicio de la sociedad y su desarrollo, abierta al público, que adquiere, conserva, investiga, comunica y expone el patrimonio material e inmaterial de la humanidad y su entorno con fines de educación, estudio y disfrute" [1]. Jette Sandahl, curadora danesa que encabeza la comisión del ICOM, plantea que la definición actual "no habla el idioma del siglo XXI" al ignorar las demandas de la "democracia cultural". Según el artículo publicado por la propia curadora [2], se debe diseñar una nueva definición que se ajuste a los museos del siglo XXI, que reconozca su existencia de sociedades diversas y cambiantes, y que apoye en el desarrollo de nuevos paradigmas.

Es evidente que las instituciones de la memoria se han dado cuenta de las complejidades del mundo actual. Tener que hacer frente a las nuevas preferencias de los visitantes, rivalizando con el cine, los teatros, los centros comerciales, y un reto aún mayor, el internet y todo lo que implica el acceso a la información, desde un dispositivo electrónico. La función de conservar y preservar el patrimonio sigue siendo lo más importante, lo único que ha cambiado es la forma en que los trabajadores de los museos y otros especialistas deciden acercarse al público [3].

El término "Web Semántica", según la definición de World Wide Web Consortium (W3C), es la visión del W3C al referirse a la Web de datos enlazados. Las tecnologías de Web Semántica permiten a las personas crear almacenes de datos en la Web, construir

vocabularios y escribir reglas para manejar información. Los datos enlazados están potenciados por tecnologías como RDF, SPARQL, OWL y SKOS [4].

Otra manera de entender el significado de Web Semántica, puede ser la definida en [5], de forma abstracta como la alianza entre la representación del conocimiento y las herramientas lo suficientemente potentes como para permitir razonar sobre los datos en línea. De forma estricta, la alianza entre los lenguajes y las aplicaciones de la web semántica para compartir, analizar y procesar datos.

Actualmente, la Web Semántica y sus beneficios son explotados para usos de disímiles tipos en los museos, como es el caso del Grafo de Conocimiento del Museo del Prado, construido sobre estándares de la Web Semántica y de acuerdo con los principios de la Web de Datos Enlazados [6]. El British Museum creó una nueva versión de su base de datos basada en la Web Semántica [7].

La publicación como datos abiertos está creciendo en los últimos años por parte de todo tipo de organizaciones como ayuntamientos, museos e instituciones relacionadas con el patrimonio cultural. Por lo que este proyecto se suma a esta línea de estas iniciativas y pretende animar a que se sumen a la revolución y digitalización de las instituciones de la memoria.

1.2 Motivación

Las recientes investigaciones sobre la narración digital interactiva, la personalización, la adaptabilidad, y la realidad mixta, junto con los sistemas que permiten la movilidad, prometen no sólo hacer más atractivos los sitios del patrimonio cultural, sino también proporcionar nuevos medios para transmitir el conocimiento, la interpretación y el análisis [8].

Con los recientes avances de la Web Semántica y las tecnologías de Deep Learning, se abre un nuevo campo de aplicación, debido a la existencia de grandes cantidades de información digital. El caso en el que se basa este trabajo es en el Museo Nacional de Ciencias Naturales (MNCN), uno de los museos de Historia Natural más antiguos de Europa y el más importante de España. En MNCN actualmente ya se encuentra en explotación, un proyecto para promover el aprendizaje informal, a través de un juego del tipo búsqueda del tesoro. El juego Enigma MNCN es una búsqueda del

tesoro para dispositivos móviles [9], donde se combina los divertido, por medio de minijuegos, con lo didáctico a través de la lectura de escritos establecidos en la historia del juego.

1.3 Objetivos

El objetico general de este trabajo es crear un procedimiento semiautomático, en donde los expertos tengan interacciones puntuales, empleando tecnologías de la Web Semántica y técnicas de Deep Learning. Con un texto de entrada, una pregunta, que deberá tener relación con dicho relato, y conociendo la respuesta correcta, generar opciones a posibles contestaciones incorrectas o como serán denominas en el trabajo distractores. Los distractores son el grupo de entidades, en nuestro caso paleontólogos o lugares de importancia en la paleontología, que tendrán la función de confundir a los jugadores cuando deban seleccionar la respuesta correcta, en preguntas que presentan cuatro opciones.

El término experiencia estará definido como el marco de aplicación del trabajo y en el cual se realizaron las pruebas del algoritmo, preguntas de opción múltiple. Las interrogantes permitidas en nuestra experiencia serán aquellas en las que las respuestas sean paleontólogos y lugares que figuran como relevantes en esta rama de las ciencias naturales.

Como objetivos específicos se plantean los siguientes:

- ***Estudio de las herramientas utilizadas***: profundizar en las características de cada una de las librerías y plataformas relacionadas con la Web Semántica y técnicas de Deep Learning empleadas en el presente trabajo.
- ***Crear un corpus a medida con la temática*:** realizar el diseño de un algoritmo que permita la recopilación de artículos de Wikipedia, que tengan relación con la experiencia a desarrollar. El proceso debe incorporar una etapa de filtrado y técnicas de procesamiento de textos, con los objetivos de crear un vocabulario homogéneo y aumentar las coocurrencias en el corpus.

- ***Entrenar un modelo con redes neuronales*:** generar un modelo con un vocabulario amplio y que se acople a la experiencia, empleando el algoritmo Word2vcet, con el objetivo final de calcular las similitudes entre entidades.
- ***Predecir distractores de acuerdo con la experiencia que indique el experto*:** recopilar las entidades relacionadas con cada una de las respuestas correctas de los cuestionarios definidos por el especialista. Filtrar y mostrar cada conjunto de opciones, en donde cada distractor esté relacionado con la contestación correcta.
- ***Generar niveles de conocimientos para los usuarios*:** establecer niveles de dificultad teniendo en cuenta las similitudes entre los distractores y la respuesta correcta, y el nivel de inlinks o popularidad de cada entidad.
- ***Evaluar los resultados obtenidos*:** valorar el comportamiento y resultados, al someter el algoritmo completo, a diferentes experiencias y cuestionarios.

1.4 Plan y estructura de trabajo

Para conseguir cada uno de los objetivos planteados anteriormente, lo primero es tener claro sobre que experiencia se desea trabajar, y como abordarla. En el caso de este proyecto, se tratarán dos tipos de respuestas, personas vinculados a la paleontología y lugares con significado para dicha rama de la ciencia. Con lo cual, el proceso de redactar el texto que alimentará el entrenamiento del modelo debe ser realizado por un especialista, que deberá tener muy claras las características de los cuestionarios que pretende diseñar. Esta fase es la más importante, y repercutirá en el resto de las etapas.

En el siguiente enlace, https://drive.google.com/drive/folders/1RiaWRKMcrBwR6QmiMWSiV4BFZ1117_aE?usp=sharing se encuentra la programación de la aplicación utilizada como apoyo para este trabajo, en donde se permite su uso en local. En dicha carpeta compartida en Drive, si se desea indagar en los cuestionarios, se encuentran los mismo con sus resultados por etapas. Y en los casos que desee ejecutarlo en línea, también se encuentra la versión en Google Colab. Todo bajo licencia de código abierto, tanto las librerías como el

desarrollo de la aplicación en general. De igual manera, se cuenta con un archivo *Readme.txt*, en donde se detalla la estructura de la carpeta.

Este trabajo se encuentra organizada en cuatro capítulos, el primero, que es el actual, va dirigido a realizar una introducción que recoge el planteamiento general, motivaciones y los objetivos trazados en el proyecto. El segundo, describe cada una de las tecnologías y herramientas que son empleadas, presentado un breve resumen de trabajos relacionados con la temática del proyecto. El capítulo tres consta de la implementación de los objetivos planteados, profundizándose en cada una de las etapas del proceso, culminando con el análisis de los resultados obtenidos. Por último, las conclusiones, en donde se realiza un balance sobre el trabajo de manera integral.

Capítulo 2 - Descripción de métodos y tecnologías aplicadas

En este capítulo serán descritos todas las tecnologías involucradas durante el desarrollo de este trabajo. Además de realizar un breve resumen de sus funcionalidades, estructura y principales características, se menciona la función que ocupa en el algoritmo. De igual manera, se presenta una abreviada recopilación de trabajos previos que utilizan la Web Semántica, específicamente Wikipedia, como fuente de información para desarrollo de aplicaciones. Los mencionados artículos también tratan temas basadas en cuestionarios de preguntas y respuestas de opción múltiple, recopilación de textos, reconocimiento de entidades nombradas y procesamiento de lenguaje natural.

2.1 DBpedia

DBpedia es un proyecto que ofrece la posibilidad de acceder a la información creada en varios proyectos de Wikimedia, como el caso Wikipedia. Permite a los usuarios la extracción de datos estructurados, multilingües y de forma gratuita, mediante las tecnologías de la Web Semántica y datos enlazados o crowdsourcing.

La versión en inglés describe actualmente 6,6 millones de entidades, de las cuales 4,9 millones tienen resúmenes, 1,9 millones tienen coordenadas geográficas y 1,7 millones de representaciones. En total, 5,5 millones de recursos se clasifican en una ontología coherente, que consta de 1,5 millones de personas, 840.000 lugares, 496.000 obras (incluidos álbumes de música, películas y videojuegos), 286.000 organizaciones, 306K especies, 58K plantas y 6K enfermedades. El número total de recursos en DBpedia en inglés es de 18M que, además de los 6.6M de recursos, incluye 1.7M de conceptos (categorías) de Simple Knowledge Organization System (SKOS), 7.7M de páginas de redireccionamiento, 269K de páginas de desambiguación y 1.7M de nodos intermedios [10].

Desde su creación en 2007, la DBpedia no ha dejado de publicar datos abiertos en RDF, extraídos de varios proyectos de Wikimedia mediante un complejo sistema de software llamado DBpedia Information Extraction Framework (DIEF) [11]. El proyecto

emplea Resource Description Framework (RDF) como un modelo de datos flexibles para la representación de la información que se extrae y publica en la web, a través de conocidas propiedades estandarizadas. La actual estructura del proyecto se basa en un conjunto de datos DBpedia RDF que se aloja y publica mediante OpenLink Virtuoso. La infraestructura Virtuoso proporciona acceso a los datos RDF de DBpedia a través de un punto final SPARQL, junto con el soporte HTTP para los GET estándar de cualquier cliente web para representaciones HTML o RDF [12], tal como se muestra en la figura 2.1.

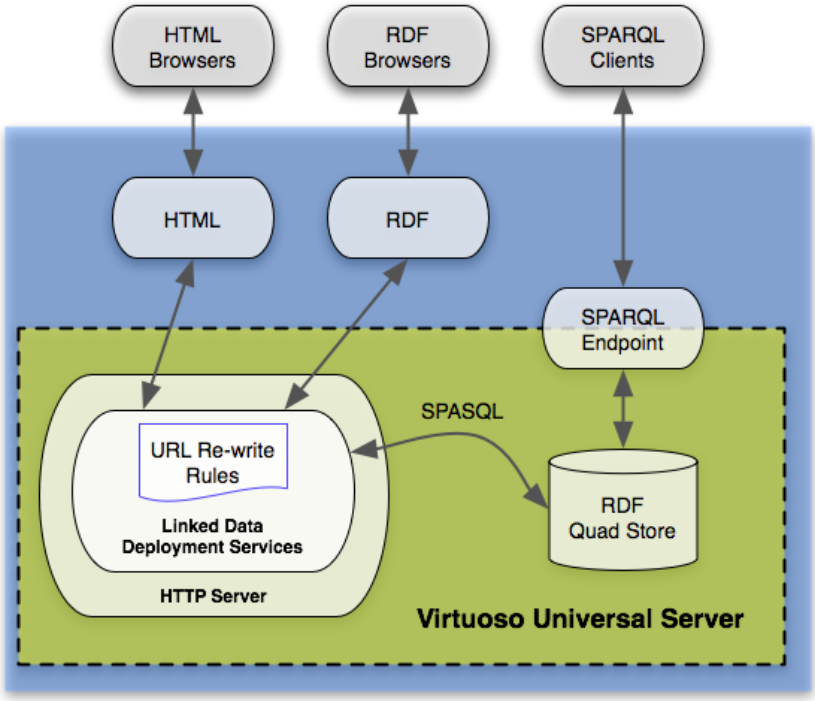

Figura 2-1Ilustración de la arquitectura actual de suministro de datos de DBpedia [12].

Dentro de los estándares de RDF se encuentra Uniform Resource Identifier (URI), el cual se emplea para la representación de cada entidad de DBpedia utilizando el prefijo dbr y la forma *http://dbpedia.org/resource/Name,* donde el *name* proviene de la URL del artículo fuente de Wikipedia, que tiene la forma *http://en.wikipedia.org/wiki/Name*. Como una mejora desde 2011 se agregó la posibilidad de consultar por idioma de la siguiente manera, *http://xx.dbpedia.org/resource/Name*, donde xx corresponde al idioma de Wikipedia al que se hace referencia. A partir de la versión 3.8, se comienza a utilizar Internationalized Resource Identifier (IRI), para todos los idiomas exceptuando el inglés, por temas de compatibilidad con versiones anteriores.

El tipo de contenido de Wikipedia más valioso para la extracción de la DBpedia son los infoboxes. Se utilizan con frecuencia para enumerar los hechos más relevantes de un artículo como una tabla de pares atributo-valor [13]. En su definición se emplean una gran variedad de plantillas que permiten especificar una serie de atributos, como es el caso de personas, lugares, libros, organizaciones, hechos, películas, etc. Como muchos editores no siguen las recomendaciones planteadas por Wikipedia, provoca que pueda existir un atributo que se refiera al mismo valor, pero se define sintácticamente diferente, por ejemplo, dbr: birthplace y dbr: placeofbirth. Este problema es muy importante que se tenga en cuenta para el trabajo con DBpedia. En su representación se emplea la siguiente estructura http://dbpedia.org/property/, con el prefijo dbp, para representar las propiedades extraídas de infobox en bruto.

Desde el lanzamiento de DBpedia 3.7, la ontología se representa por medio de un gráfico acíclico dirigido, no un árbol. Las clases pueden tener múltiples superclases, lo cual es importante para las asignaciones a schema.org. La ontología de DBpedia contiene actualmente alrededor de 4.233.000 instancias, y se puede consultar a través del end-point DBpedia SPARQL. Y se representa http://dbpedia.org/ontology/ y el prefijo dbo, por ejemplo, dbo: abstract, dbo: birthDate, dbo: birthPlace.

En este trabajo DBpedia es herramienta principal para la adquisición y procesamiento del corpus, y de igual manera se utilizará como parte del mecanismo de selección, filtrado y clasificación de candidatos a distractores.

2.2 DBpedia SpotLight

DBpedia SpotLigth es un servicio en línea donde a partir de un texto se logra relacionar las superficies semánticas con entidades de DBpedia. Las superficies somáticas son palabras o conjunto de ellas, que de acuerdo con el contexto del texto que se consulta se atribuye mayor importancia en la narrativa. La plataforma tiene como principal tarea, lo que se conoce como Named-Entity Recognition (NER). La conforman herramientas de detección de entidades o identificación de superficies semánticas, resolución de nombres, en donde diferentes formas de superficie pueden conducir a la misma entidad, dependiendo del contexto.

La herramienta posee tres tipos de opciones de consultas básicas, se puede acceder a ellos desde Scala / Java API, un servicio web REST y desde una interfaz de usuario en la web (HTML / Javascript). Existen una serie de parámetros de configuración que se pueden usar para influir en las funciones de anotación y desambiguación, así como el formato de salida de la consulta: text/html, application/xhtml+xml, text/xml, application/json. Las funciones básicas son:

- /annotate: devuelve texto/entidades de una URL, localización, mapeo de candidatos, desambiguación y vinculación/estadísticas.
- /spot: devuelve candidatos a una URL.
- /candidates: localización y mapeo de candidatos a una URL

De las tres funcionalidades en este trabajo de emplea */annotate* para el reconocimiento y asociación de superficies semánticas relevantes a entidades de DBpedia, y la consulta */candidates* para la desambiguación de entidades de DBpedia.

El servicio posee opciones de configuración como, *confidence*, umbral para el nivel de confianza en el mapeo de las entidades. Las respuestas a las consultas contienen parámetros como *support* (número mínimo de enlaces de entrada de las entidades candidatas), o un conjunto de tipos, *types,* de entidades que deben estar asociadas en el conjunto de datos DBpedia. En la tabla 1, se muestra un ejemplo de respuesta a una consulta */annotate* en DBpedia SpotLigth para el caso de la superficie semántica Georges Cuvier, conocido por ser considerado el padre de la paleontología.

@URI: http://dbpedia.org/resource/Georges_Cuvier
@support: 1144
@types: Http://xmlns.com/foaf/0.1/Person, Wikidata:Q901, Wikidata:Q5, Wikidata:Q24229398, Wikidata:Q215627, DUL:NaturalPerson, DUL:Agent, Schema:Person, DBpedia:Person, DBpedia:Agent, DBpedia:Scientist
@surfaceForm: Georges Cuvier
@offset: 295
@similarityScore: 1.0
@percentageOfSecondRank: 0.0

Tabla 2-1 Ejemplo de descripción de entidad devuelta por DBpedia SpotLight.

2.3 Word2Vect

Word2Vec es una herramienta de Natural Language Processing (NPL), el algoritmo emplea redes neuronales de dos capas y para crear un modelo que permite aprender asociaciones de palabras, a partir de un gran volumen de texto. En donde las palabras se representan como vectores y se capturan las características semánticas, estudiando la co-ocurrencia de palabras en la entrada.

Word2vec fue creado, patentado y publicado en 2013, por un equipo de investigadores de Google, dirigidos por Tomas Mikolov, recogidos en los artículos son [14], [15]. Word2Vec es una técnica sin supervisión alimentada por una fuente de datos sin etiqueta, es decir, texto sin formato. Estos artículos propusieron dos métodos para aprender representaciones de palabras:

- Continuous Bag-Of-Words (CBOW): predice la palabra del medio en función de las palabras del contexto circundante. El contexto consta de unas pocas palabras antes y después de la palabra actual (intermedia) y donde el orden de las palabras en el contexto no es importante.
- Continuous Skip-Gram: predice palabras dentro de un cierto rango antes y después de la palabra actual en la misma oración.

En el modelo CBOW, las representaciones distribuidas del contexto o de las palabras circundantes se combinan para predecir la palabra del medio. En el modelo Skip-Gram la representación distribuida de la palabra de entrada se utiliza para predecir el contexto [16].

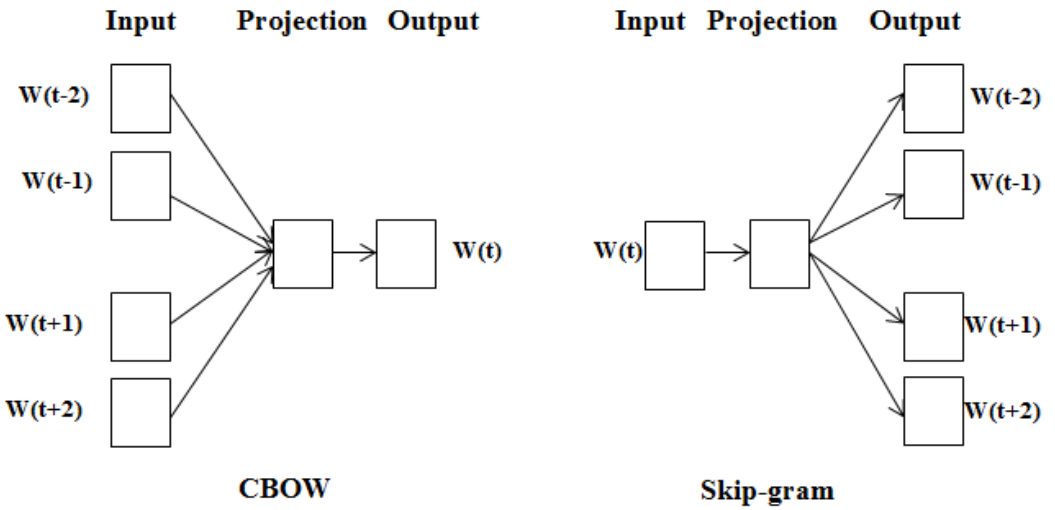

Figura 2-2 Representación gráfica del modelo CBOW y del modelo Skip-Gram [16].

Dada las características de este trabajo, el modelo más adecuado es Skip-Gram debido a que se tiene una pregunta y se conoce la respuesta correcta, y lo que se busca es calcular las probabilidades del resto de palabras del vocabulario que se acerquen a la de entrada. En la Figura 2-3, muestra que la capa oculta no tiene función de activación y en cambio la de salida aplica una función softmax, que consigue que todos los valores estén entre 0 y 1, y que la suma de todos los valores sea 1, es decir, que sea una distribución de probabilidades.

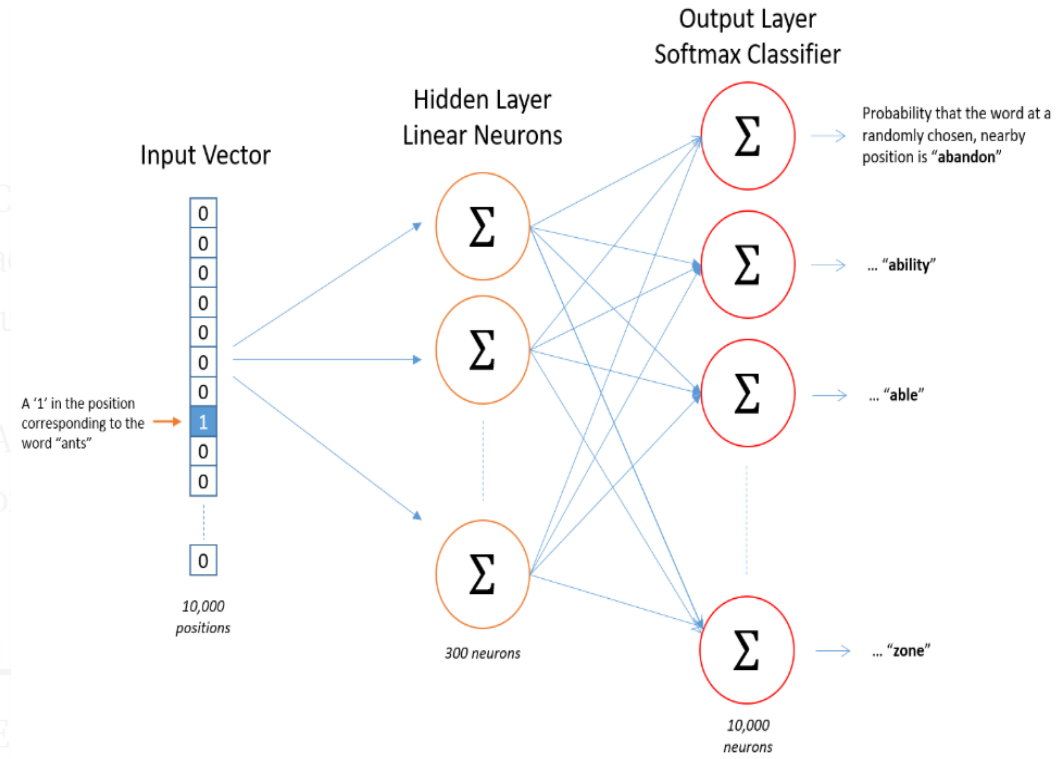

Figura 2-3 Propagación hacia adelante de la red neuronal en Word2Vect [17].

El modelo resultante será una representación vectorial de las palabras del corpus, donde una palabra y otras que son semánticamente equivalente, tengan valores de probabilidad similares Con lo cual, se encuentran vectorialmente representadas más cerca, que las palabras que tienen menos o nula relación con la palabra primaria. Este algoritmo será utilizado para identificar la similitud de la respuesta correcta respecto a los distractores, empleando las probabilidades del modelo entrenado en las dimensiones de las palabras contenidas en la pregunta. En la Sección 3.2.2 se encuentra descrito este proceso.

2.4 WikiData

Wikidata es un repositorio que almacena información estructurada y alimentada de forma colaborativa por parte de la Fundación Wikimedia, junto a proyectos como Wikipedia. El repositorio consta principalmente de *items*, cada uno con una *label*, una *description* y cualquier número de *aliases*. Los elementos se identifican de forma única, Q, seguido de un número. En la figura 2-4, se muestran los principales términos de un elemento de Wikidata.

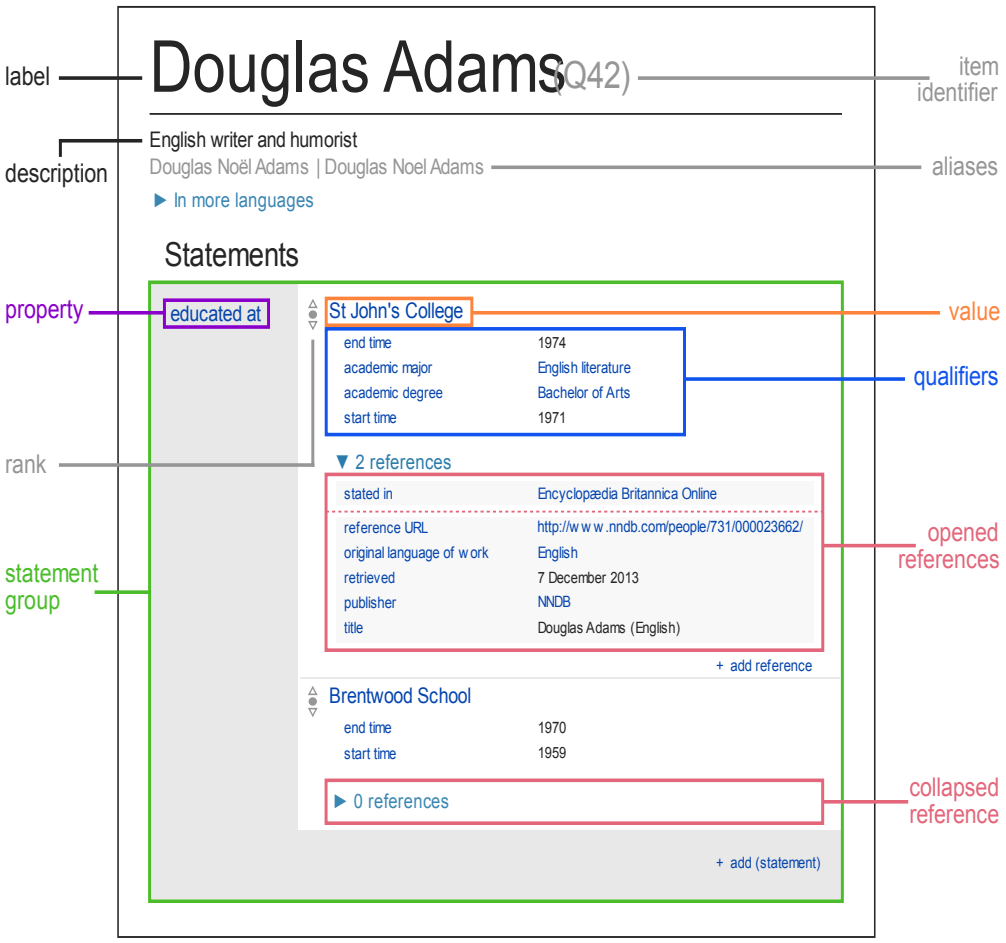

Figura 2-4 Diagrama de un elemento de Wikidata [18].

La plataforma permite varias maneras para realizar consultas, a través de URI desreferenciables que siguen los estándares de datos vinculados, como por ejemplo *https://www.wikidata.org/wiki/Q171969*, o mediante la API de MediaWiki. Para el caso de la API, puede realizarse a través de end-point SPARQL, bots o directamente realizar un volcado a local de todos los datos de WIkidata. En el caso de esta investigación se realiza a través de consultas de end-point SPARQL, para la recopilación de artículos vinculados a las experiencias.

2.5 Stanza

Stanza es una biblioteca de análisis del lenguaje natural en Python, desarrollada por el grupo de procesamiento del lenguaje natural de Stanford (NLP). Stanza admite funcionalidades como tokenización, expansión de token de varias palabras, lematización, parte del discurso (POS), etiquetado de características morfológicas,

análisis de dependencia, reconocimiento de entidades nombradas (NER) y análisis de sentimientos.

El conjunto de herramientas está diseñado para ser paralelo entre más de 70 idiomas, utilizando el formalismo de dependencias universales. Los módulos están construidos sobre la biblioteca PyTorch [19].

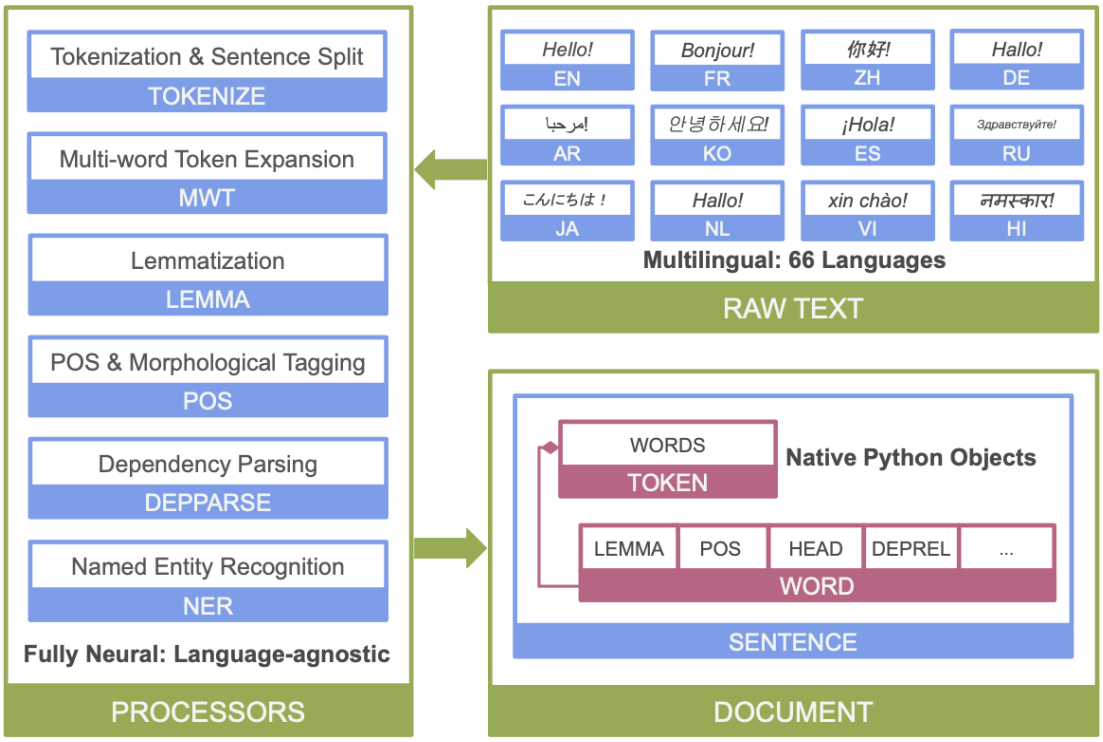

Figura 2-5 Descripción general de la canalización de NPL de la red neuronal de Stanza [19].

La herramienta es emplea en el procesamiento de texto proveniente del corpus y de la experiencia en general. Son utilizadas las herramientas de tokenización, lematización y parte de discurso POS.

2.6 Estudio de trabajos relacionados

Los avances recientes en la Web Semántica, las tecnologías de aprendizaje profundo, el Big Data, unido a la necesidad de automatizar el aprendizaje, da como resultado la creación de proyectos, artículos, y trabajos relacionados con estas temáticas. Dentro del panorama de Big Data, un territorio importante son los datos vinculados o linked data. Según el director del W3C, Tim Berners-Lee, lo definió de una manera muy sencilla: Linked Open Data (LOD) son datos vinculados que se publican

bajo una licencia abierta, lo que no impide su reutilización de forma gratuita. Una tendencia actual es la creación de juegos que hacen uso de la gran cantidad de datos abiertos vinculados (LOD) para crear nuevos artefactos de conocimiento, proporcionar aplicaciones útiles y tratar de mejorar la calidad de LOD.

Varios trabajos en el área de juegos serios han aprovechado el emerger de LOD, para así desarrollar servicios automatizados para generar preguntas, respuestas o cuestionarios completos con el fin de mejorar el aprendizaje en las diferentes áreas de conocimientos. Trabajos [20], se enfocan en recrear experiencias enfocadas en el medio ambiente, aprovechan el potencial de los sistemas de redes sociales, en particular Facebook, para atraer jugadores.

Aplicaciones como Sherlock [21], presenta un sistema semiautomático para generar cuestionarios de preguntas, permitiendo establecer niveles de dificultad a través del cálculo de similitudes de coseno, Distancia Semántica de Datos Vinculados (LDSD).

En el caso de [22], plantean la alternativa para la recopilación de artículos de Wikipedia relevantes para la historia explotando las clasificaciones de Wikicategorias. El trabajo tiene un objetivo muy similar a este, y es la generación automática cuestionarios personalizados de opción múltiple, con alternativas incorrectas a la respuesta correcta, adaptadas al nivel de conocimiento del usuario.

Siguiendo la misma línea de pensamiento, nos encontramos con [23], un proyecto europeo, CrossCult, que se enfoca en fomentar la historia, basándose en el aprendizaje y el entrenamiento [24], utilizando como base los datos de DBpedia y Wikidata. De igual manera, [25], presenta un algoritmo, que ha servido de inspiración para este proyecto en donde el objetivo del trabajo es identificar los distractores apropiados, dada una pregunta y su respuesta correcta. Son utilizadas las ventajas de la Web Semática, combinándose con diferentes plataformas para cubrir y solucionar las limitaciones de estas.

El trabajo [26] resulta muy interesante, es un nuevo enfoque de juego para dispositivos móviles que genera su contenido de forma autónoma, sin intervención humana, utilizando hechos existentes de la versión semántica de Wikipedia. Combina

herramientas como DBpedia, basada en servicios RESTful, Node.js, MongoDB y HTML5. El sistema proporciona a los usuarios preguntas sencillas que comienzan con una de estas cuatro combinaciones ("¿Qué es ...?", "¿Quién es ...?", "¿Cuándo fue ...?" Y "¿Dónde está ...?"). La aplicación final es un cuestionario formado por preguntas y respuestas, autogeneradas, totalmente automático desde el diseño de la pregunta hasta la obtención de las respuestas correctas y sus respectivos distractores.

En [27] se utiliza Linked Open Data para la generación de ítems de evaluación educativa. Generan un cuestionario desde el inicio, definiendo plantillas de Especificación de Interoperabilidad de preguntas y Pruebas de IMS (IMS QTI-XML), en donde se generan los propios distractores para cada combinación en base a consultas DBpedia. Presentan los resultados en TAO3, plataforma semántica de código abierto para la creación y entrega de pruebas y elementos de evaluación.

Capítulo 3 - Obtención de distractores

Este capítulo detalla cada uno de los pasos que componen el algoritmo para cumplir el objetivo planteado: Obtener los distractores dada una pregunta, conociendo la respuesta correcta, en un entorno dirigido a paleontólogos y sitios de relevancia paleontológica. A continuación, se presentará tres secciones, la primera dirigida a la recopilación de artículos relevantes a la experiencia general, dígase paleontólogos y sitios paleontológicos. Un segundo apéndice para la obtención de los distractores, y, por último, la presentación y visualización de resultados, basadas en experiencias generadas durante la investigación.

3.1 Obtención del corpus

Esta sección está enfocada en realizar la recopilación de los textos que puedan llegar a ser relevantes en la paleontología de manera general. El resultado de esta sección posibilitará la creación de un modelo basado en redes neuronales, específicamente con el algoritmo Word2Vect, cuya función es permitir el análisis y cálculo de las similitudes entre vectores de palabras.

Para conseguir una experiencia lo más efectiva posible, hay que pautar una serie de características o establecer un marco. Esto permitirá, que el texto que se procese sea los relacionados con la temática y no se inunde de información innecesaria, que puede provocar ruido en el modelo y gasto de tiempo sin ninguna finalidad.

1. Los textos deben estar relacionados con los dos temas de la experiencia.
2. Como el objetivo es registrar las coocurrencias, deben estar presente entidades del tipo paleontólogo y sitios de importancia paleontológicas, y formar parte del vocabulario del modelo, más detalles en la Sección 3.2.2.
3. Cada texto será comparado con el texto inicial proporcionado por expertos, y deberá contener una semejanza mínima, la cual podrá se configurable según el especialista.

Con el objetivo de complementar el algoritmo, se creó una interfaz para la configuración de las experiencias, Figura 3.1. La herramienta no es considerada como

parte de la finalidad del trabajo, pero si tiene las opciones y recursos básicos para utilizarse como apoyo.

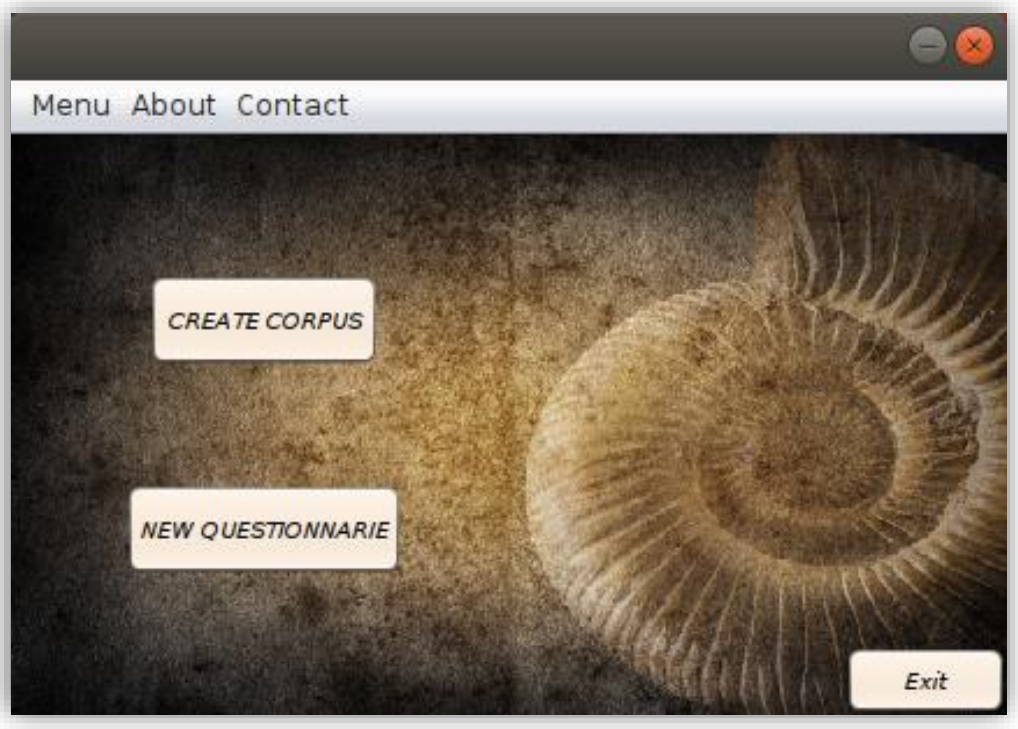

Figura 3-1 Interfaz de usuario gráfica, Menú Inicial.

Para llevar a cabo este paso, el especialista deberá proporcionar un texto general en donde se informe sobre la temática, pero lo más importante es que se mencionen entidades del mismo tipo que las respuestas correctas que se esperan. Esto es uno de los puntos claves en el trabajo, porque será el punto de partida para la recopilación del corpus. En la Figura 3.2 se presenta el texto creado para esta experiencia.

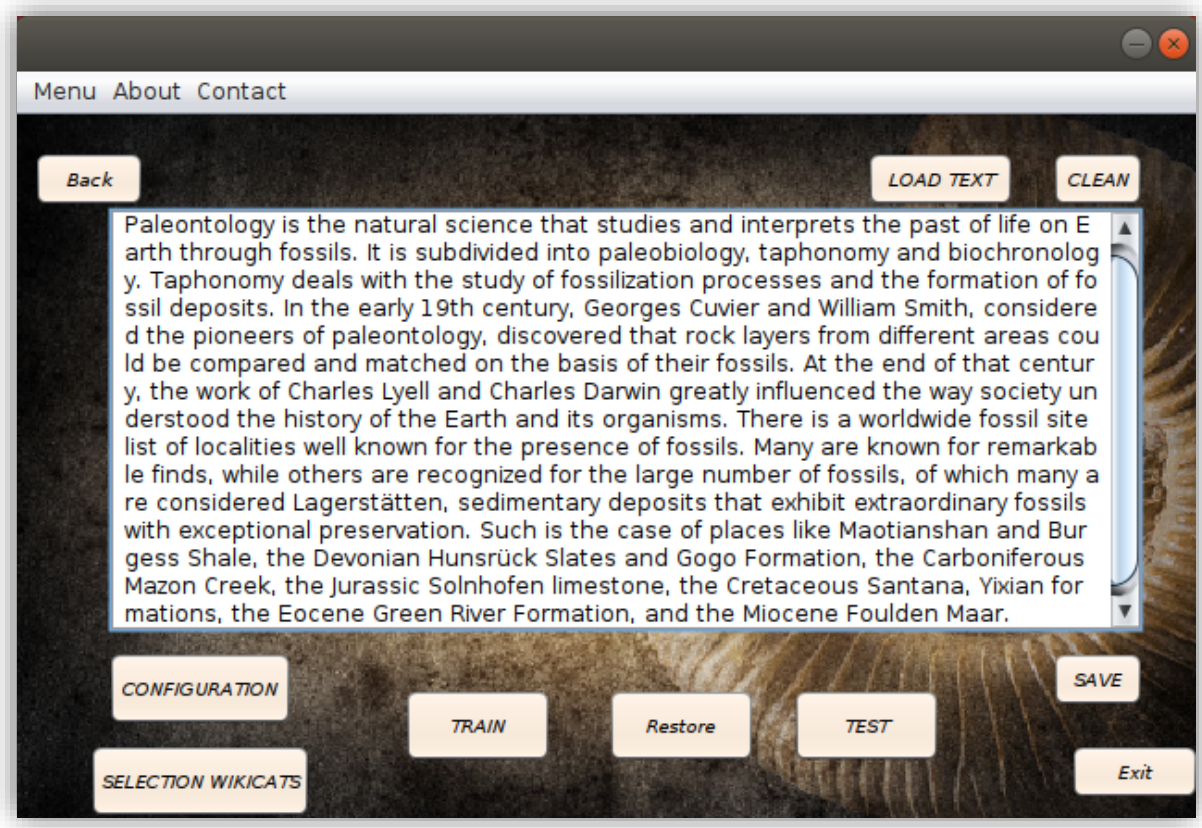

Figura 3-2 Texto introducido por el experto.

Como se puede observar, se desarrolla el concepto de paleontología, menciona algunos de los paleontólogos más relevantes como Georges Cuvier, Charles Lyell, William Smith. Se expone la definición de Lagerstätten, en español depósitos, pero especialmente para fósiles, mencionando algunos de los más relevantes como Maotianshan, Burgess Shale, Devonian Hunsrück Slates, Jurassic Solnhofen limestone, entre otros.

3.1.1 Identificación de títulos de páginas de Wikipedia.

El siguiente paso es el procesamiento del texto inicial. Esto se logra utilizando la herramienta DBpedia SpotLight (DB-SL), donde la aplicación identificará las superficies semánticas, una palabra o conjunto de estas, que considera relevante en el contexto del texto, y lo asocia a entidades DBpedia.

La consulta se utiliza un valor de confidence de 0.4, puntuación para la desambiguación/vinculación, entre mayor es el número menos superficies semánticas

detecta en el texto. Con lo cual, se vuelve más exigente al analizar el texto, con 0.4 se mostró muy buenos resultados, así como es el valor más utilizado en la bibliografía. Otro detalle para tener en cuenta con esta plataforma, es que solo admite hasta 5000 caracteres por consulta, devolviendo un 502 si lo sobrepasa. Por lo que, de suceder esto se deberá dividir el texto para que siempre se cumpla que sea menor a dicho limite. A continuación, se muestra la consulta a DB-SL:

```
curl https://api.dbpedia-spotlight.org/en/annotate --data-urlencode "text=%s" --data
"confidence=%s" -H "Accept: application/json" >> input.json"'%(paragraph, confidence )
```

Tabla 3-1 Modelo de consulta a DBpedia, opción annotate.

La respuesta a este paso será un conjunto de entidades, tal como con la estructura que se representa en la Tabla 2-1 de Georges Cuvier. Como se puede apreciar en dicha tabla existe una variedad de parámetros, de los cuales solo se tendrán en cuenta: @URI, @support, @types y @surfaceform. Un detalle importante en las respuestas de las consultas en DB-SL, en la mayoría de los casos, el parámetro @types, muy importante en la aplicación, suele estar vacío o incompleto, con lo cual por cada entidad que se identifique con esta plataforma, se procede a utilizar la @URI para consultar en DBpedia los @types, y otro parámetro que no devuelve SpotLight, que son los @subjects.

SELECT ?object ?type **WHERE** { dbr:Georges_Cuvier rdfs:label ?object. dbr:Georges_Cuvier rdf:type ?type.}	**SELECT** ?object ?type ?subject **WHERE** { dbr:Georges_Cuvier rdfs:label ?object. dbr:Georges_Cuvier dct:subject ?subject.}

Tabla 3-2 Consultas SparQL de @types y @subjects a DBpedia.

El objetivo principal de esta sección es encontrar las entidades que estén relacionadas con los tipos de respuestas que se tienen previstas para la aplicación. Con lo cual, se realiza un estudio en las clasificaciones en DBpedia, arrojando la información que, al ser una plataforma cooperativa, donde intervienen miles de autores y no siempre las clasificaciones son exactas, sucede en ocasiones que son muy genéricas y a veces erróneas. Dentro de @types se encuentra Wikicats, un sistema de vanguardia para categorizar artículos [28], debido a su precisión se procede utilizarlo para el proceso de recopilación. Se confecciona un grupo de Wikicats resultado de la agrupación de esta

categoría por parte de las entidades que son extraídas del texto inicial, y mostrados al experto para su selección Figura 3.3.

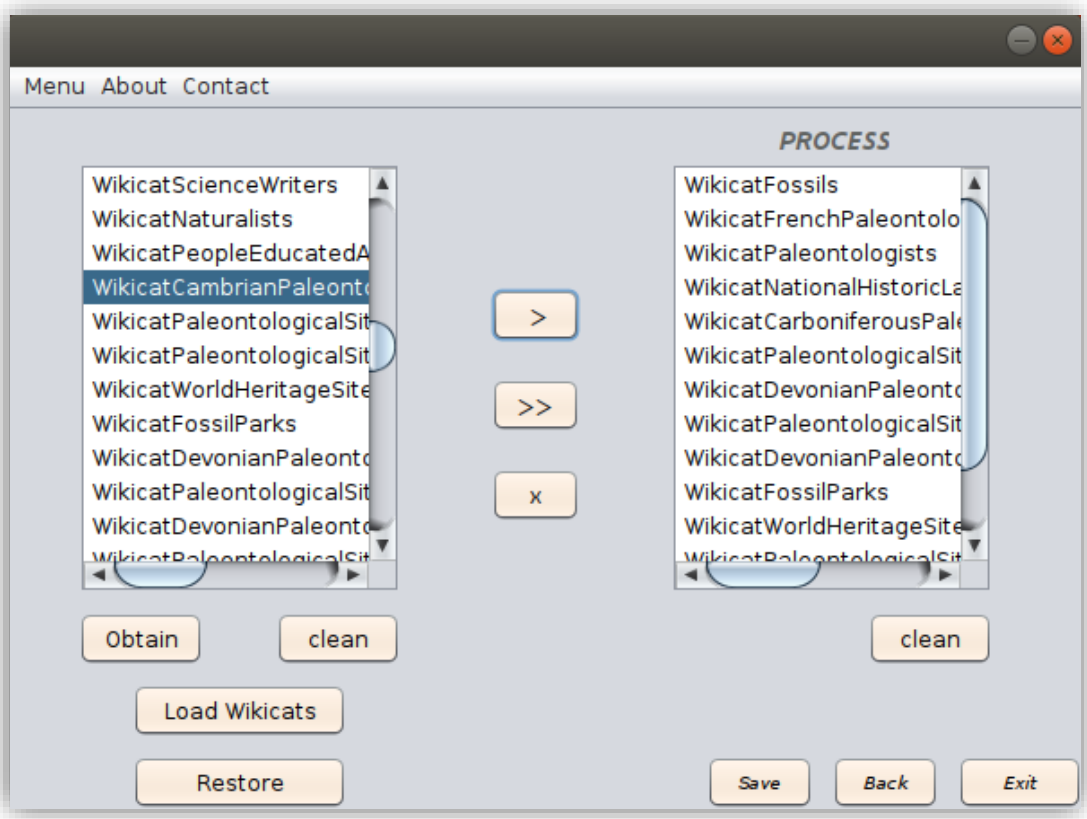

Figura 3-3 Selección de Categorías para la recopilación del corpus.

La importancia de seleccionar correctamente los Wikicats radica en que, dado el caso que se elija uno incorrecto, dará la posibilidad de entrada a muchas entidades innecesarias. La consecuencia real estaría en gasto de tiempo de procesamiento y menos en el ruido que se adicione al proceso de recopilación. Otro factor importante, que puede ser configurado por el experto, es la precisión en generar el texto, que se traduce en definir la cantidad de Wikicats que debe tener un texto para que se considere relevante a la experiencia, Figura 3-4.

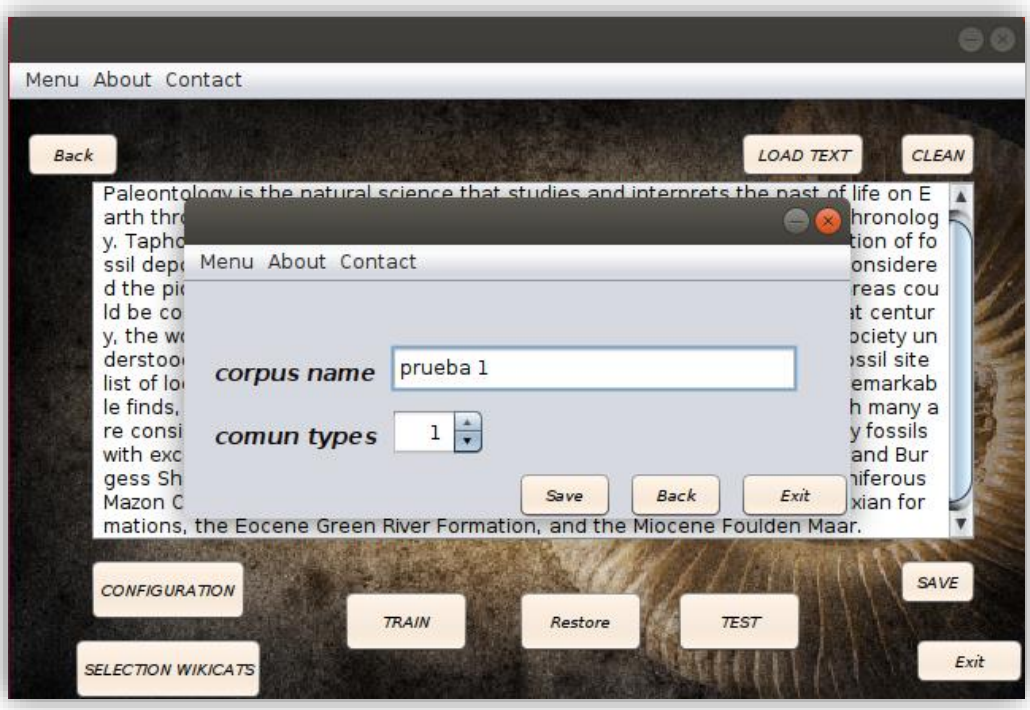

Figura 3-4 Selección del umbral de similitud.

Con la selección de Wikicats realizada por el experto, el próximo paso es obtener los títulos de los artículos de Wikipedia. Esto se logra, por medio de consultas a Wikidata, que se recogen en la Tabla 3-3, donde utilizando las categorías seleccionadas por el especialista, se obtiene listas de títulos de Wikipedia. En la misma tabla se aprecia alrededor 20 títulos obtenidos como respuesta al ejemplo que se muestra.

SELECT * **WHERE {** **SERVICE** wikibase:mwapi { bd:serviceParam wikibase:api "Search". bd:serviceParam wikibase:endpoint "en.wikipedia.org". bd:serviceParam mwapi:srsearch " PaleontologicalSitesOfEurope ". ?title wikibase:apiOutput mwapi:title. }}
Prebreza, Tequendama, Portugal, Fossil collecting, Aetokremnos, Li Mae Long, Gradina on Bosut, Cuddie Springs, Argiles d'Octeville, Naracoorte Caves National Park, Physiology, Siberia, 2020 in paleontology, 2001 in paleontology, 2018 in mammal paleontology, 2016 in paleontology, 2019 in paleontology, Scorpion, Bear, Crkvine (Stubline), Appalachia (landmass), Lady Gaga

Tabla 3-3 Consulta Wikidata para obtener títulos de páginas de Wikipedia relacionadas a la categoría.

3.1.2 Filtrado de las entidades.

Llegados a este punto, se cuenta con más de 70 000 títulos de artículos de Wikipedia, de los cuales más de la mitad no está relacionados con la experiencia. Se puede apreciar en la Tabla 3-3, como existen artículos en donde es muy evidentes que no existe ninguna relación, por ejemplo, Bear, Scorpion, Lady Gaga, Physiology. Por lo que es necesario realizar una etapa de filtrado, antes de incorporar los textos al corpus. A continuación, se resumen los pasos para seleccionar que títulos son relevantes y cuáles serán desechados.

1- Cada título es consultado en DBpedia.

2- Se extrae la propiedad dbo: abstract que existe para cada entidad DBpedia. El proceso deriva a consultar el texto del resumen en DB-SL, siguiendo los pasos de epígrafe anterior. El resultado será un conjunto de Wikicats. Cada entidad consultada deberá tener la cantidad de Wikicats común que seleccionó el experto, con los seleccionados en la etapa. Todos lo que cumplan o superen ese valor, se considera que están relacionados con la temática.

3- Existen casos como Juan Luis Arsuaga, paleontólogo español, profesor de la Universidad Complutense de Madrid, en el que su @abstract se encuentra en español, lo que provoca que la respuesta a la consulta sea nula. Para esos casos y en los que directamente no exista tal propiedad, se procesan los propios @types de la entidad, seleccionado los Wikicats y realizando la misma comparación. El parámetro de selección, como serán muchos menos que los recogidos de un texto, (paso número 2), bastará con tener solo un Wikicat en común con los de la sección 3.1.1, para que ser relevante.

El resultado de los tres pasos anteriores permite estar en posesión de un conjunto de entidades, que estarán en total correspondencia con la experiencia, aunque puedan existir casos en los que no. La presencia de textos que no estén ligados a la temática no genera un mal mayor al corpus, pero si traerá más consumo de tiempo de procesamiento.

Es importante tener en cuenta casos como el de William Smith, geólogo por profesión, pero que incursionó en la paleontología, al ser identificado por la herramienta

DB-SL se asocia a la entidad incorrecta, *William_Smyth* un obispo británico, cuando la correcta debería ser William_Smith_(geologist). Para prever estos errores, cuando se procese unidades de textos, cada superficie semántica que se identifique será consultada en DB-SL, específicamente en la opción */candidates, explicada en el epígrafe 2.2,* y de ellas se selecciona la entidad DBpedia, que más Wikicats en común tenga con los seleccionados por el especialista en la sección 3.1.1.

Con dicho conjunto de entidades, se procede a extraer los textos de Wikipedia. Para esto, se emplea consultas a MediaWiki API**,** que devuelven los textos de cada página o artículo de Wikipedia consultado, filtrando tablas, imágenes y ruido en general.

3.1.3 Procesado del corpus

Este paso va dirigido al tratamiento de los textos seleccionados como relevantes para la experiencia. En todos los casos en los que se vaya a utilizar Natural Language Processing, es muy común utilizar herramientas para el tratamiento previo, solo dependerá de la aplicación la robustez de este paso. A continuación, se listan los pasos seguidos:

1- Se elimina los caracteres y signos de puntuación.

2- Se elimina las palabras que carecen de un significado por sí solas, conocidas por stopwords, y que son fundamentalmente artículos, preposiciones, conjunciones y pronombres.

3- Lematización de las palabras utilizando la herramienta Stanza, sección 2.5.

4- Sustitución de las superficies semánticas por sus correspondientes entidades DBpedia. Este paso es de los más importante, porque proporciona un vocabulario común, y permite el aumento de la coocurrencia en esos casos en los que una misma entidad DBpedia es asociada a diferentes superficies semánticas.

3.2 Entrenamiento de la red neuronal Word2Vect

Esta sección, está dedicada a un breve resumen sobre cómo se entrenó la red neuronal Word2Vect. Para acceder a Word2Vect, se utiliza la librería gensim, específicamente el módulo models. word2vec, para Word2vec embeddings.

Este módulo implementa la familia de algoritmos word2vec, utilizando rutinas C altamente optimizadas, transmisión de datos e interfaces Pythonic. Los algoritmos word2vec incluyen modelos Skip-Gram y CBOW, utilizando softmax jerárquico o muestreo negativo [29]. Para este caso, será utilizado el modelo de Skip-Gram, debido a que se tiene una pregunta y se conoce la respuesta correcta, y lo que se busca es calcular las probabilidades del resto de palabras del vocabulario, que se acerquen a la de entrada.

Los pasos que se realizan son los siguientes:

1- Se divide en oraciones, donde realmente la división es por texto de entidad, una oración corresponderá a un texto completo.

2- Se definen los parámetros para el entrenamiento:

- num_features = 5000
- min_word_count = 1
- num_workers = multiprocessing.cpu_count()
- context_size = 7
- downsampling = 1e-3
- seed = 1

Debido a los experimentos realizados, se concluyó que la cantidad de corpus no es exageradamente grande. Con lo cual, para fomentar una mejor experiencia, se decide tomar a 1, el número mínimo de repeticiones para que se incluya en el vocabulario. Esto provoca la disminución, por una parte, de que se anulen preguntas porque la entidad correcta no se encuentra en el vocabulario, y por otra parte, que entidades distractoras se desechen porque no se encuentran en el modelo.

3- Se construye el modelo con los parámetros anteriores:

model = w2v.Word2Vec(sg=1, seed=seed, workers=num_workers, size=num_features, min_count=min_word_count, window=context_size, sample=downsampling).

El parámetro sg, si es 1 el algoritmo de entrenamiento es Skip-Gram, y 0 para CBOW.

4- Se crea el vocabulario del modelo, model.build_vocab(sentences) .

5- Se procede a entrenar la red, model.train(sentences, total_examples=len(sentences), epochs=5).

En este punto se tiene un modelo, con un vocabulario muy amplio, 66334 vocablos lematizados y que permitirá el estudio de los candidatos a distractores, que serán definido en el próximo epígrafe. El tema de Word2Vect y el modelo que se obtuvo, será retomado más adelante.

3.3 Adquisición de distractores

En esta sección se detallará el proceso creado para la obtención de los candidatos a distractores, el cual tiene una infraestructura previamente creada, debido a que el proceso es muy parecido al de recopilación del corpus. En los siguientes subepígrafes, se detalla cada uno de los pasos para la selección, filtrado y agrupación de los distractores.

3.3.1 Obtención de candidatos

Para la recopilación de candidatos, se exploran dos fuentes de entidades, que posteriormente serán filtradas, pero está sección va dirigida a su recopilación. La primera lista de candidatos será denominada lista interna, la cual se conformará a través de una consulta a Wikidata. El texto para esta consulta será: texto de contexto, más pregunta, más respuesta correcta, tal como se representa en la Figura 3-5, con el objetivo de brindar contexto a la plataforma.

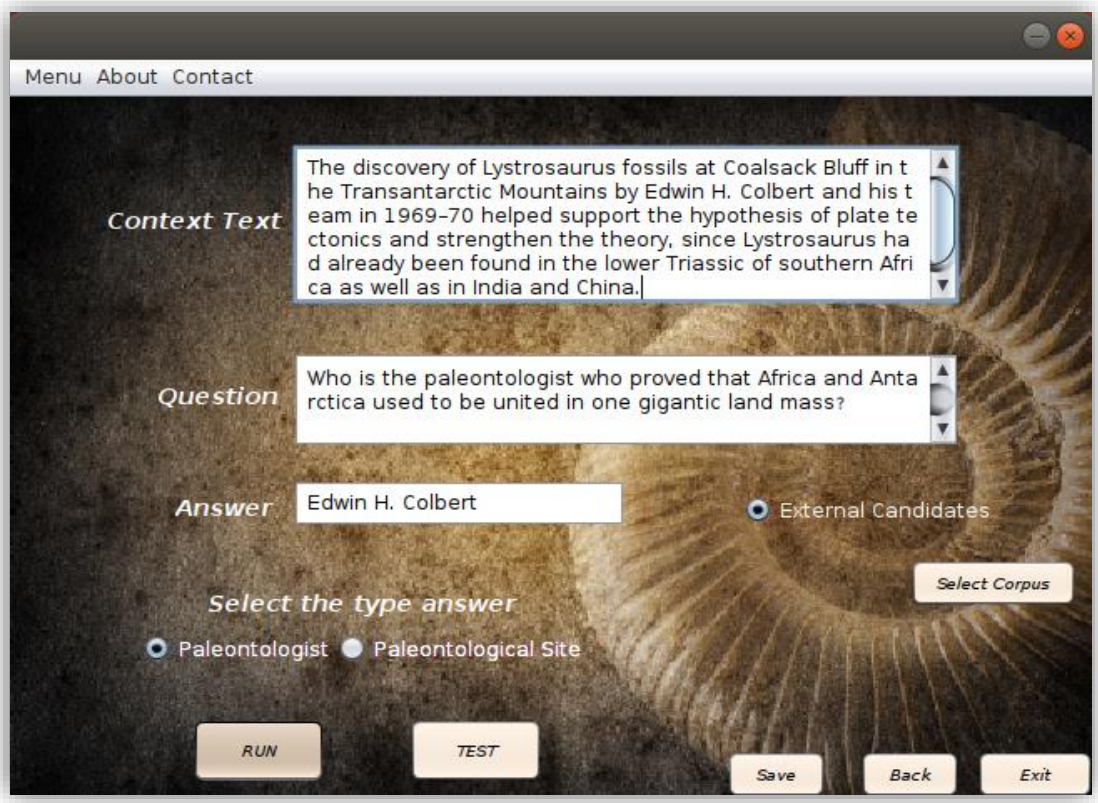

Figura 3-5 Presentación de un ejemplo de un cuestionario.

De la réplica a la consulta, solo se tendrá en cuenta la entidad DBpedia que se asocia a la respuesta correcta, en este caso, https://dbpedia.org/page/Edwin_H._Colbert. El procedimiento que se utiliza a partir de este punto es muy parecido al realizado para la obtención del corpus, por lo que se resumen en una serie de pasos a continuación:

1. De la entidad asociada a la respuesta correcta, se realiza la consulta a DBpedia para obtener sus @types, Wikicats y @subjects, debido a que serán utilizados en la etapa de filtrado
2. La entidad se consulta en Wikidata para obtener títulos de páginas de Wikipedia, tabla 3-3.
3. Cada título, es consultado a DB-SL, para obtener la entidad correcta asociada, almacenado el parámetro @support que será necesario en la agrupación por niveles, Sección 3.3.3.

4. Con esta lista de entidades, se procede a consultar a DBpedia, extrayendo los @types, Wikicats y @subjects.

El resultado de este proceso, son los primeros candidatos distractores, de las cuales muchos no están relacionados con la experiencia, o no son paleontólogos. Con lo cual, es necesario una etapa de filtrado, que garantice que los distractores sean los más equivalentes a la respuesta correcta. De igual manera, para aumentar el nicho de distractores, se propone como fuente externa, candidatos provenientes del corpus para el entrenamiento del modelo, que no estén contenidos en esta primera lista de candidatos internos a distractores. En la siguiente sección se presenta el proceso de depuración sobre los candidatos obtenidos.

3.3.2 Filtrado de los candidatos a distractores

Recapitulando, se cuenta con dos fuentes de entidades, a ambas se les aplicará el mismo proceso de filtrado. La primera fase es verificar que se cuenta con el tipo de entidad que corresponde. Para esto, en un inicio el experto indica, que tipo de candidatos espera, Figura 3-5, que debe corresponder con la respuesta correcta. El procedimiento para la comprobación es realizar consultas a DBpedia sobre propiedades únicas de personas o sitios, los cuales son los únicos tipos definidos o esperados para la aplicación.

En los casos de paleontólogos se verifica que la entidad cuenta con una de estas propiedades: dbp: birthDate, dbp: deathDate, dbp: birthPlace o dbp: deathPlace. Estas características son únicas para personas reales, lugar y fechas de nacimiento y muerte. Con la existencia de una de estas propiedades, la entidad pasa la próxima etapa de filtrado. En el caso de sitios de relevancia paleontológica, las propiedades son: dbp: location, dbp: capital, dbp: municipality, dbp: region.

Con esta fase, pueden haber pasado entidades que sean personas o sitios, pero no son ni paleontólogos ni sitios de relevancia paleontológica. Para esto se mantendrán las entidades que tengas al menos un @types en común con los de la respuesta correcta. Importante que se indica un @types, no un Wikicat, esto es porque durante la investigación se encontraron entidades DBpedia, que corresponden con la experiencia y en total concordancia con la entidad correcta, pero se encuentran clasificadas de

manera muy simple. Por lo que, en la siguiente etapa de filtrado, se limita la entrada a aquellas entidades que poseen un @types en común, pero pueden no estar cualitativamente relacionadas con la entidad correcta.

Para obtener una selección más compacta, lógica y teniendo en consideración características cualitativas, no tan evidentes a primera vista, los candidatos deberán tener a menos un @subjects en común con la respuesta correcta. Por último, se elimina todo candidato que no forme parte del vocabulario del modelo de Word2Vect obtenido.

Figura 3-6 Candidatos obtenidos en el proceso para el cuestionario número 2 de ejemplo.

Tras finalizar todos estos pasos, las listas internas y externas son las entidades para distraer a los usuarios en las preguntas, Figura 3-6. Pero surge una problemática, y es con qué criterios seleccionar los candidatos, ya que se prevé respuestas con 4 opciones, y una ya es la correcta. Para dar solución a esto, es para lo que se entrenó el modelo Word2Vect, el cual permitirá definir la semejanza entre los candidatos y la respuesta correcta. Los detalles de este proceso serán explicados en la siguiente sección.

3.3.3 Agrupación por niveles de dificultad

El especialista tendrá la opción de definir hasta tres niveles como máximo. El reto es establecer un criterio por el cual escoger y dividir por dificultad. Existe un parámetro, @support, que está vinculado al número de inlinks de Wikipedia, y refleja la popularidad de dicha página para los usuarios que utilizan la plataforma. Pero esto no es suficiente, para establecer semejanza con la respuesta correcta, porque solo indica notoriedad o fama. El criterio inicial para establecer los niveles estará relacionado con los cálculos de similitud que proporciona el modelo obtenido con Word2Vect. Luego, cuando se cuente con las listas de niveles, estas serán organizadas por el parámetro @support, permitiéndole al usuario definir la complejidad dentro de cada nivel, logrando mayor dominio para crear los cuestionarios.

El modelo obtenido con Word2Vect cuenta con un vocabulario muy amplio, que en su mayoría está compuesto por texto sobre paleontólogos o lugares relevantes ligados a descubrimientos de fósiles o temas de paleontología en general. Al ser un vocabulario muy amplio, se puede perder importancia el contexto de la pregunta, por lo que, en vez de obtener la similitud en base a todo el corpus, se hará en base a las dimensiones de la pregunta.

Para conseguir homogeneidad entre el vocabulario del modelo y las preguntas, será utilizado la herramienta Stanza. para aplicar la lematización y el algoritmo POS. El objetivo será las palabras relevantes de las preguntas, permitan establecer un mejor contexto descartando así preposiciones, pronombres, artículos. Por lo tanto, en este caso solo se utilizarán sustantivos, verbos y adjetivos. El módulo Part-of-Speech & Morphological Features (POS), etiqueta POS específicas de bancos de árboles (XPOS, empleada en este caso). Por ejemplo, "Who is considered the winner of the Bones Wars?.", en la Figura 3-7 se muestra cómo se etiqueta el fragmento de texto y accede la parte gramatical de cada palabra, y en [30] se detalla lo que significa cada etiqueta:

Word	Lemma	POS
is	be	VBZ
considered	consider	VBN
winner	winner	NN
bone	bone	NN
wars	war	NNS

Figura 3-7 Ejemplo para la lematización y aplicación de los POS TAG.

Conseguido esto, se contará con un vector pregunta, y n cantidad de vectores distractores, donde n es la cantidad de distractores, que pasaron la etapa de filtrado y que se encuentran en el vocabulario del modelo. En los próximos subepígrafes, se describe el algoritmo para el cálculo de la similitud o distancia entre la respuesta correcta y sus distractores.

3.3.3.1 Wmdistance

Calcular la similitud de oraciones requiere construir un modelo gramatical de la oración, comprender estructuras equivalentes (por ejemplo, "caminó a la tienda ayer" y "ayer, caminó a la tienda"), encontrar similitud no solo en los pronombres y verbos sino también en los nombres propios, búsqueda de co-ocurrencias/relaciones estadísticas en muchos ejemplos textuales reales, etc.

Word mover´s distance fue introducido en el artículo [31]. Está inspirado en la "Distancia del Movimiento de la Tierra" y emplea un solucionador del problema del transporte. Mide la similitud entre dos documentos de texto como la distancia mínima que las palabras incrustadas o words embeddings de un documento tienen que tener para llegar a las palabras incrustadas de otro documento [32]. WMD, como un caso especial de Earth Mover's Distance, es la distancia entre dos documentos de texto x, y ∈ X, que tiene en cuenta las alineaciones entre palabras. Sea |x|, |y| ser el número de palabras distintas en x y *y*. Sea $f_x \in R^{|x|}$ y $f_y \in R^{|y|}$ la definición para los vectores de frecuencia de cada palabra normalizadas en los documentos x y *y*, respectivamente (de modo que $f_x^T 1 = f_y^T 1 = 1$) [33]. Por lo que la distancia entre ambos documentos se define como:

$$WMD_{(x,y)} \coloneqq \min \langle C, F \rangle \ , \ F \epsilon \mathbb{R}_{+}^{|x| \ x \ |y|} \ s.t., \ F1 = \ f_x \ , F^T 1 \ ) \ f_y \qquad \text{Ec.1}$$

Donde F es la matriz de flujo de transporte con $F_{i,j}$, que denota la cantidad de flujo que viaja desde la i-ésima palabra en x en x_i hasta la j-ésima palabra y en y_j. C es el costo con $C_{i,j} = dist(v_{xi}, v_{yj})$, siendo la distancia entre dos palabras medidas en el espacio de incrustación de Word2Vec. Construido sobre Word2Vec, WMD es particularmente útil y preciso para medir la distancia entre documentos con palabras semánticamente cercanas, pero sintácticamente diferentes, como se ilustra en la Figura 3-8.

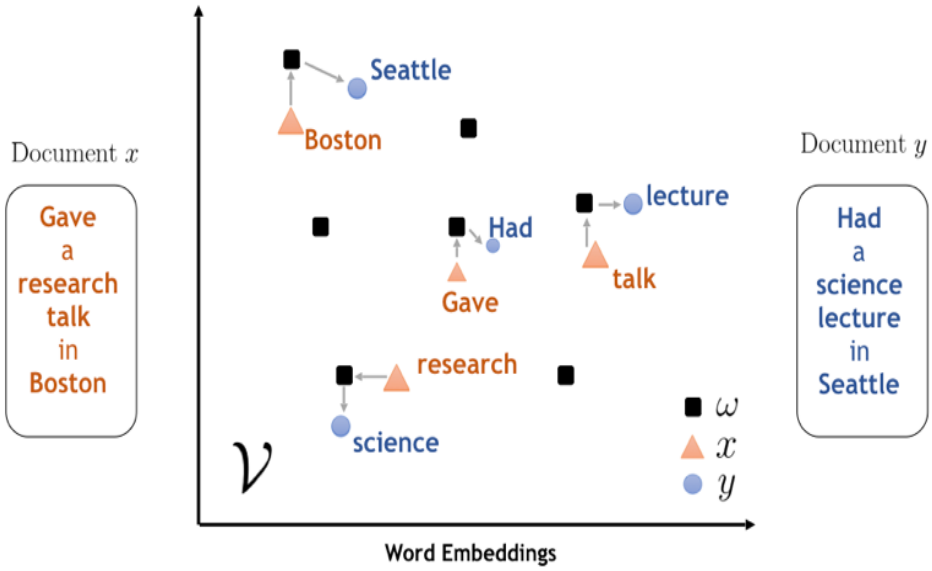

Figura 3-8 Ejemplo de representación de las distancias de dos documentos [33].

La funcionalidad WMD de Gensim, que consiste en el wmdistance método para el cálculo de distancias y la WmdSimilarity clase para consultas de similitud basadas en corpus. El utilizado para este caso es wmdistance [29], aplicado entre dos vectores, la respuesta y los candidatos, respecto al vector pregunta. En la Figura 3-8 se muestra un ejemplo, ajeno a la temática, pero permite una mayor compresión del significado del método. En donde se observa como dos oraciones, que se refieren a temáticas similares, devuelve un valor menor de distancia, que dos que no están relacionadas en nada.

```
import gensim.downloader as api
model = api.load('word2vec-google-news-300')
sentence_obama = 'Obama speaks to the media in Illinois'
sentence_president = 'The president greets the press in Chicago'
distance = model.wmdistance(sentence_obama, sentence_president)
print('distance = %.4f' % distance)

distance = 1.0175

sentence_orange = preprocess('Oranges are my favorite fruit')
distance = model.wmdistance(sentence_obama, sentence_orange)
print('distance = %.4f' % distance)

distance = 1.3663
```

Figura 3-9 Ejemplo de aplicación del método wmdistance entre oraciones [34].

El método es aplicado tanto a la respuesta correcta como a los distractores, y se almacena, para cada candidato, la diferencia absoluta entre el valor de la respuesta correcta, calculado previamente, y el valor wmd del candidato. El resultado será una lista ordenada de distractores, de acuerdo con dichos resultados en orden ascendente. El nivel de dificultad estará dado por, con mayor dificultad, los distractores son los más similares y viceversa, encontrando un intermedio, que dependerá de la cantidad de niveles que solicite el experto, manteniendo una equidad en número entre distractores por niveles. Dentro de cada nivel, se organizará por el parámetro @support, donde el mayor valor equivale a mayor popularidad, y viceversa. En las Figuras 3-9 y 3-10, se reflejan los resultados para dos y tres niveles respectivamente.

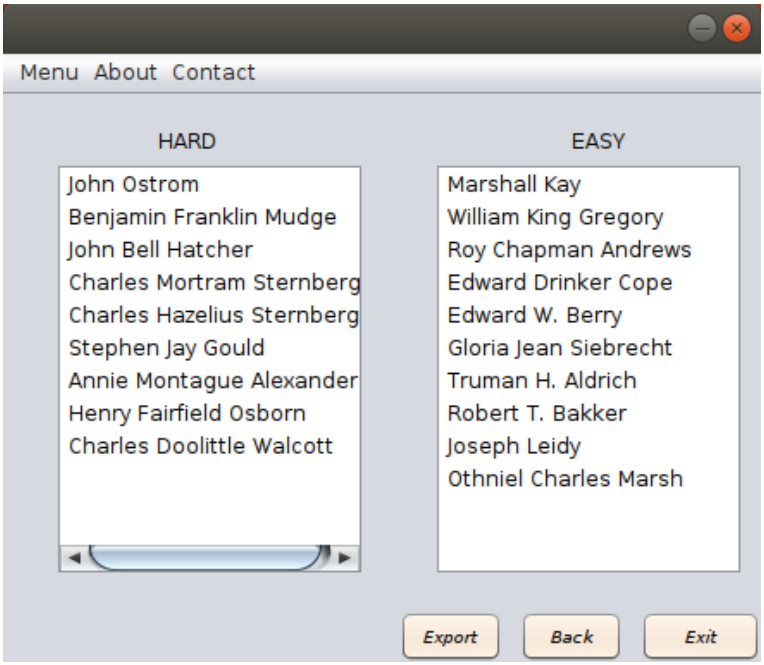

Figura 3-10 Resultados para un cuestionario de dos niveles.

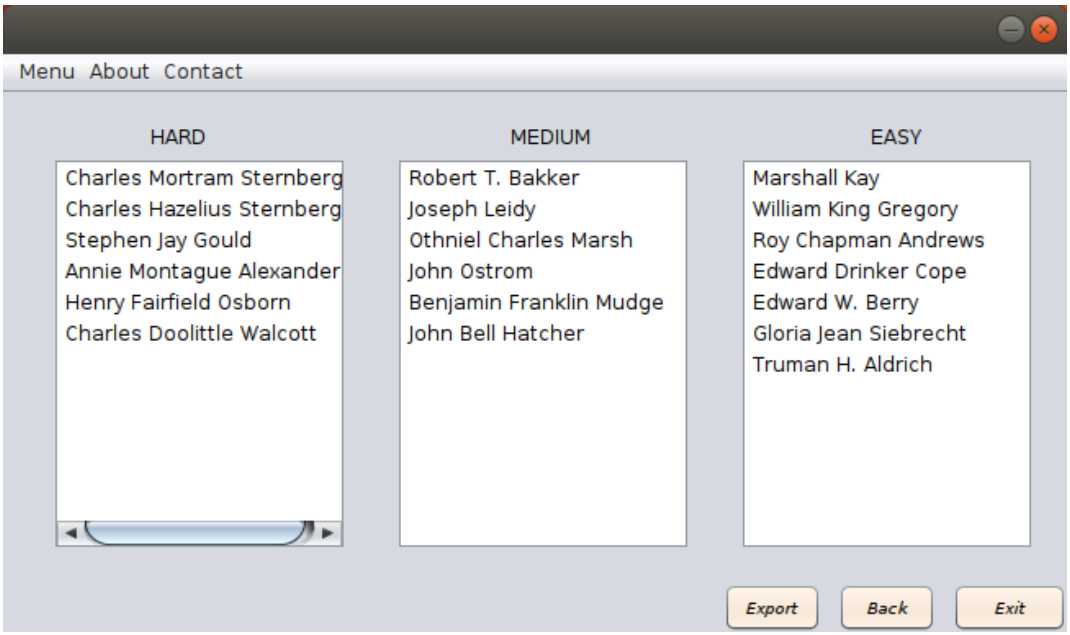

Figura 3-11 Resultados para un cuestionario de tres niveles.

3.3.4 Sugerencia de nuevos cuestionarios

Aprovechando que los resultados son un conjunto amplio de posibilidades, se agrega la opción de sugerir al experto nuevas preguntas al cuestionario. El principio es explorar las propiedades de la entidad definida como respuesta correcta. Luego de un

estudio de las propiedades que pueden llegar a presentar las entidades, se concluyó que, en cuanto a los dos tipos de respuestas, son los paleontólogos los únicos con posibilidad de explotar esta variante. Las propiedades a las que se concluyeron son, dbp: award, dbp: author y dbp: almamater.

El algoritmo determinará cuál de estas propiedades está presente en la entidad correcta, y de existir, recorrerá cada uno de los distractores, y almacenará los resultados a las consultas de estas propiedades en aquellos que la presenten. Esta opción estará disponible siempre y cuando, la respuesta correcta tenga la propiedad y el total de estos distractores a las propuestas de pregunta sean igual o superior a tres. Las Figuras 3-12 y 3-13 muestran dos de los tres casos a tipos de preguntas que se proponen. La propiedad autor no estaba presente los paleontólogos utilizados en los cuestionarios.

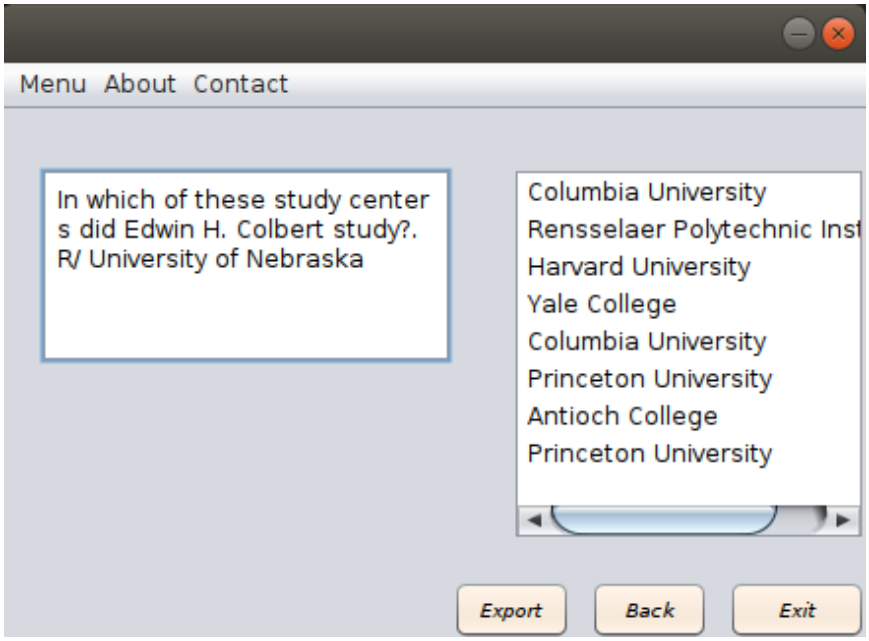

Figura 3-12 Opción de nuevos cuestionarios para complementar la aplicación, ejemplo para Edwin H. Colbert.

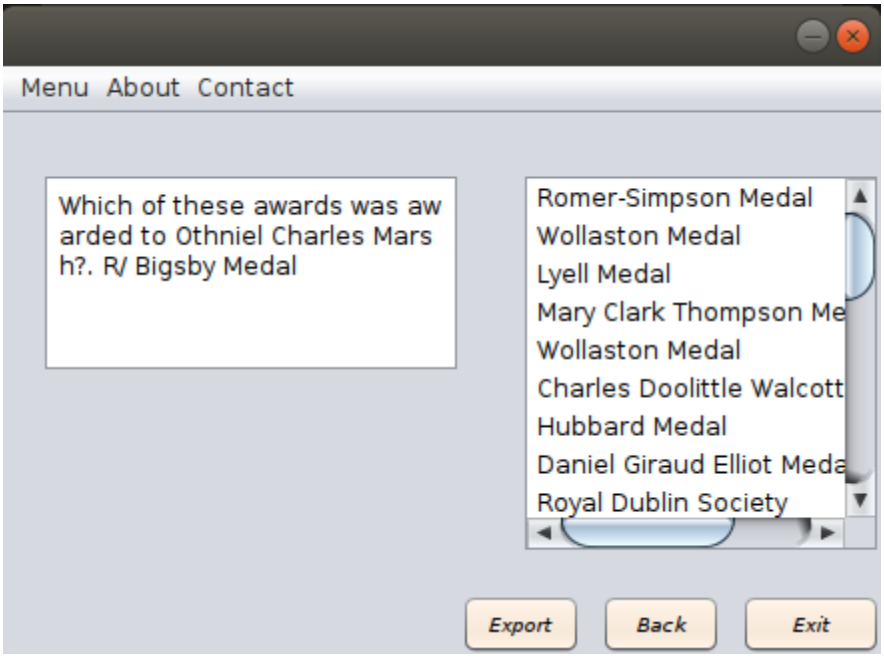

Figura 3-13 Opción de nuevos cuestionarios para complementar la aplicación, ejemplo para Othniel Charles March.

3.4 Presentación de ejemplos y análisis de resultados

El algoritmo fue aplicado a 10 preguntas referidas a paleontólogos y otras 10 a sitios o yacimientos paleontológicos. Lo primero que se aplicó fue la recopilación del corpus, con el texto y la experiencia generada que se muestra en la Figura 3-2. Con el resultado, se creó un modelo lo bastante abarcador y cuyo vocabulario permitió elaborar los cuestionarios que se encuentran el *Apéndice A. Cuestionarios*. A continuación, se presenta las tablas que resumen, las preguntas, cantidad de distractores finales por fuente de entrada y total.

Número	Preguntas	Candidatos Internos	Candidatos Externos	Total
1	Who is known as the founding father of paleontology? R/ Georges Cuvier	29	57	86
2	Who is considered the winner of the bone wars? R/ Othniel Charles Marsh	14	90	104

3	Who is credited as the discoverer of the tyrannosaur? R/ Barnum Brown	2	79	81
4	What is the name of the paleontologist who discovered the first fossil teeth? R/ Gideon Mantel	9	58	67
5	Who is the paleontologist who proved that Africa and Antarctica used to be united in one gigantic land mass? R/ Edwin H. Colbert	7	85	92
6	Spanish paleontologist who is known for being the initiator of the study of the Pleistocene sites in the Sierra de Atapuerca? R/ Emiliano Aguirre	0	57	57
7	Paleontologist known as the king of collectors and for discovering two dinosaur genera Torosaurus and Triceratops? R/ John Bell Hatcher	6	75	81
8	Female paleontologist who recovered and named the dinosaur Podokesaurus holyokensis and 1909 became the first woman elected to membership in the Paleontological Society? R/ Mignon Talbot	0	88	88
9	Name of the Paleontologist did not discover it, but he did find the most complete skeleton of Sarcosuchus or also known as SuperCoc.? R/ Paul Sereno	4	110	114

10	What is the name of the paleontologist that one of his most relevant findings was the first recognized bird with teeth? R/ Benjamin Franklin Mudge	3	75	78

Tabla 3-4 Estadística para los cuestionarios sobre paleontólogos.

La Tabla 3-4 resume para cada pregunta los resultados finales, en cuestión de recopilación de distractores. Es muy fácil identificar como el proceso de recopilación para los candidatos internos propició menos resultados que los externos. Esto puede ser atribuido a las carencias que puede existir en el algoritmo de Wikidata para relacionar temas o para dar más cantidad de artículos. De igual manera, el proceso de obtención para los distractores externos es resultado de muchas consultas a Wikidata contra una sola por parte de los internos. Otro factor relevante, es que los candidatos externos tienen mayor probabilidad de aparecer en el vocabulario del modelo que los internos.

En los cuestionarios 6 y 8 de la Tabla 3-4, no existen distractores internos, por lo que se concluye, que, a pesar de ser configurable, se recomienda siempre utilizar la opción del corpus como origen de candidatos. Esta situación fue analizada a profundidad por etapas definidas en la sección 3.3 y para encontrar el punto donde el algoritmo de candidatos internos retornara 0 distractores. Para ambos casos, el paso 3.3.1 devolvió 270 y 52 candidatos respectivamente. En la etapa de filtrado, 3.3.2, el caso de Mignon Talbot tipificada en DBpedia como paleontóloga y geóloga, solo se encontró un distractor que cumplía todas las condiciones, la geóloga Ida Helen Ogilvie, específicamente la única que tenía un @subjects en común *dbc:American_geologists*[35] . Pero, al no encontrase dentro del vocabulario del modelo entrenado, etapa 3.3.3, no es considerada como distractor. Se concluye que la razón determinante en este caso fue que Wikidata solo detectó 52 candidatos, un número muy pobre de artículos y nada valiosos para la experiencia de este cuestionario. Retomando la pregunta seis, sucede algo similar, hasta el punto de filtrado se poseía nueve candidatos, pero ninguno de ellos formaba parte del vocabulario. En este caso, sí que se considera que el proceso de entrenamiento del modelo, sección 3.2, falló.

Número	Preguntas	Candidatos Internos	Candidatos Externos	Total
1	Which archaeological site has the privilege of being the first place in the world where the existence of Upper Paleolithic Rock Art was identified? R/ Cave of Altamira	15	7	22
2	What is the fossil site where Materpiscis was found, a placoderm preserved with an embryonic juvenile still attached by its umbilical cord? R/ Gogo Formation	1	25	26
3	What is the name of the most extensive cave system in the world? R/ Mammoth Cave National Park	11	16	27
4	What is the fossil deposit that is recognized as the first place where the first massive strandings of Cambrian scyphozoans (jellyfish) appeared? R/ Blackberry Hill	0	19	19
5	What is the geological formation where the first confirmed dinosaur eggs were found? R/ Djadochta Formation	1	21	22
6	What is the name of the area or region where the oldest fossils on Earth were found and is said to have found Halszkaraptor, the first non-avian dinosaur that could move both on land and in water? R/ Pilbara Craton	0	3	3
7	What is the name of the fossil site where the dinosaur species Archaeopteryx,	0	6	6

	known as the perfect "missing link" between dinosaurs and birds, was found? R/ Solnhofen Limestone			
8	What is the name of the place where the skeleton of a child was discovered, which is considered the most complete skeleton of a prehistoric human ever found.? R/ Lake Turkana	4	1	5
9	what is the name of the fossil site where all the human species that lived in Europe were found.? R/ Archaeological site of Atapuerca	11	5	16
10	What is the name of the fossil site where Sinosauropteryx was found, which revealed the unmistakable impression of primitive hair-like feathers? R/ Yixian Formation	5	25	30

Tabla 3-5 Estadística para el cuestionario sobre sitios destacados para la paleontología.

Para el caso de los sitios paleontológicos, Tabla 3-5, la recopilación resultó ser menos cantidad en comparación con los cuestionarios de paleontólogos. Quizás pueda atribuirse a deficiencias en el corpus de entrenamiento, ranzón por la cual en algunos casos los distractores internos son mayores en cantidad de los externos.

En cuanto a las definiciones de los niveles de los jugadores, fue utilizado el método de wmdistance, y calculada la diferencia absoluta en cuanto al valor de la respuesta correcta. Es complicado, sin ser expertos en paleontología, definir cuál sería el primer, segundo, tercer o cuarto mejor distractor. Pero existe un caso, la pregunta número 2, de la Tabla 3-4, la cual hace referencia a quien es el ganador de la "Guerra de Huesos". Este suceso se reconoce como el período de mayor nivel de especulación y descubrimientos de fósiles durante la Gilded Age (Edad Dorada) de la historia de los Estados Unidos. Estuvo marcada por una gran rivalidad entre Edward Drinker Cope (de

la Academia de Ciencias Naturales de Filadelfia) y Othniel Charles Marsh (del Museo Peabody de Historia Natural de Yale).

Como bien se indica en la definición del cuestionario número 2, Tabla 3-4, la respuesta correcta es Othniel Charles Marsh, pero dado el contexto se entiende que el distractor más obvio sería Edward Drinker Cope. Con lo cual, resulta interesante verificar la posición que ocupa dicho paleontólogo en los resultados organizados de menor a mayor valor de similitud.

	Entidad	Types	Subjects	Support	wmdistance	wmdistance diff
0	Charles_Doolittle_Walcott	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:United_...	193.0	2.320309	0.000089
1	Benjamin_Franklin_Mudge	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:Kansas_...	32.0	2.318243	0.002155
2	Charles_Hazelius_Sternberg	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:Amateur...	69.0	2.342561	0.022162
3	Edward_Drinker_Cope	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:America...	1036.0	2.345767	0.025369
4	Samuel_Wendell_Williston	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:Univers...	377.0	2.291775	0.028623
5	John_Ostrom	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:Beloit_...	125.0	2.378919	0.058521
6	James_Dwight_Dana	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:Wollast...	164.0	2.381192	0.060793
7	Joseph_Henry	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:People_...	277.0	2.383738	0.063339
8	Tilly_Edinger	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:20th-ce...	20.0	2.493589	0.173191
9	John_Bell_Hatcher	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:Grinnel...	69.0	2.590711	0.270312

Figura 3-14 Primeros 10 distractores para el cuestionario número dos sobre paleontólogos.

En la Figura 3-14 se listan los primeros 10 distractores, organizados por sus valores en la columna *wmdistance diff*, que representan las respuestas con menor distancia a la correcta, es decir, mayor similitud, en las dimensiones de las palabras incluidas en las preguntas. Como queda señalada en dicha imagen, el distractor del cual se esperaba que ocupase la primera posición, observando los resultados, se aprecia que se sitúa el cuarto lugar.

Debido a que el resultado no es el esperado en un inicio, se procede a verificar las primeras ubicaciones, arrojando que existe un nexo con la respuesta correcta. Por ejemplo, Charles Doolittle Walcott [36], fue paleontólogo que administró de Smithsonian Institution, centro donde se encuentran muchos de los descubrimientos de Marsh. Los casos de Benjamin Franklin Mudge [37] y Charles Hazelius Sternberg [38], paleontólogos que vivieron en la época de March, y que además, trabajaron para él recolectando fósiles. A pesar de que no se cumple lo esperado, estos distractores poseen sentido, al

ser paleontólogos como profesión, y tener una relación comprobada con la respuesta correcta.

Para el resto de los cuestionarios los criterios de valoración son menos evidentes. Por ejemplo, el cuestionario número 1 de la Tabla 3-5: "¿Qué yacimiento arqueológico tiene el privilegio de ser el primer lugar del mundo donde se identificó la existencia de arte rupestre del Paleolítico Superior?". La respuesta correcta es la Cueva de Altamira, ubicada en el municipio español de Santilla del Mar en Cantabria, y conserva la evidencia de arte rupestre del paleolítico más importante de la prehistoria. Le corresponde la distinción de ser el primer lugar del mundo en identificarse la existencia del arte rupestre en la etapa del Paleolítico Superior.

Se analizan los primeros distractores, definidos por el algoritmo de cálculo de distancia, como los más similares a la respuesta correcta. Se aprecia en la Figura 3-15 las primeras 10 opciones obtenidas, donde se verificó que todas son cuevas, con lo cual, están en concordancia con la respuesta correcta.

	Entidad	Types	Subjects	Support	wmdistance	wmdistance diff
0	Sandia_Cave	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:Archaeo...	13.0	2.589731	0.002170
1	Cave_of_El_Castillo	['http://www.w3.org/2003/01/geo/wgs84_pos#Spat...	['http://dbpedia.org/resource/Category:Caves_o...	29.0	2.576510	0.015390
2	Plovers_Lake	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:South_A...	8.0	2.613161	0.021260
3	Makapansgat	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:Prehist...	69.0	2.525601	0.066299
4	Cave_of_La_Pasiega	['http://www.w3.org/2003/01/geo/wgs84_pos#Spat...	['http://dbpedia.org/resource/Category:Stone_a...	24.0	2.658619	0.066718
5	Wookey_Hole_Caves	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:Grade_I...	93.0	2.521859	0.070042
6	Kromdraai_fossil_site	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:South_A...	17.0	2.710149	0.118248
7	Rising_Star_Cave	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:South_A...	63.0	2.468514	0.123387
8	Sistema_Ox_Bel_Ha	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:Underwa...	21.0	2.439296	0.152604
9	Gladysvale_Cave	['http://www.w3.org/2002/07/owl#Thing', 'http:...	['http://dbpedia.org/resource/Category:South_A...	20.0	2.421005	0.170895

Figura 3-15 Primeros 10 distractores para el cuestionario número uno sobre sitios de relevancia paleontológica.

A continuación, se profundiza en los tres primeros distractores, para verificar que tiene en común con la Cueva de Altamira. Sandia Cave, se encuentra ubicada en México, e indagando en las propiedades de ambas cuevas, se descubre que tienen en común la categoría Limestone Cave [39]. Dicha categoría indica que son del mismo tipo de cueva natural de piedra caliza, formadas debajo de la superficie de la Tierra, pero no se encuentra una razón por la cual se considera la más similar. En cambio, la segunda cueva El Castillo, el nexo común es que se ubican en la misma comunidad

autonómica, Cantabria, aunque en el caso del distractor forma parte de las cuevas del Paleolítico Inferior. Se cree que puede ser muy buen distractor para un alto nivel de dificultad, lo cual coincide con la relación de mayor similitud más complejidad, al estar ubicadas en un mismo territorio y pertenecer al Paleolítico, pero diferentes subetapas. Por último, el tercer distractor Plovers Lake, situado en Sudáfrica, coincide en el tipo de cueva, Limestone Cave y en ser del Paleolítico, pero en este caso Medio.

El total de distractores por cuestionarios resultaron ser una fuente, amplia para generar diferentes versiones a preguntas, especialmente para la experiencia de paleontólogos, aunque se encontraron algunas deficiencias en el proceso. De igual manera, se considera que puede ser solucionado, si los orígenes del texto inicial provienen de un especialista, que, para este trabajo no fue el caso.

En cuanto a la temática de cálculo de distancias, similitudes entre palabras, corpus, documentos, es bastante inexacta a medida que aumenta el vocabulario, y más dependiendo en el entorno que se aplique. Se pudo comprobar que los resultados de distractores fueron lógicos y dentro del marco esperado, pero con deficiencias puntuales.

Capítulo 4 - Conclusiones

El presente trabajo proporciona un algoritmo completo y probado para recopilar contestaciones incorrectas a preguntas de las cuales se conoce la respuesta. Tiene el fin de automatizar y agilizar la definición masiva de cuestionarios, específicamente como propuesta a la aplicación móvil de Enigma MNCN, proyecto actual en explotación.

La solución parte de textos sin formato alojados en Wikipedia, y utiliza las relaciones entre los datos estructurados LOD, que proporciona plataformas como DBpedia y Wikidata, a través de consultas SPARQL. Para la explotación de las mencionadas relaciones, se emplean técnicas de reconocimiento de entidades (NER), opción que brinda de DB-SL, al igual que la desambiguación entre superficies semánticas. Para el procesado de texto, etiquetado de características morfológicas, análisis de dependencia y lematización, se aprovecha la librería en lenguaje Python, Stanza, que forma parte del poderoso conjunto de herramientas que posee Stanford's NPL Group. Para complementar, con el objetivo preciso de calcular las similitudes entre entidades, Word2vect ofrece los medios necesarios para definir las distancias semánticas entre los vocablos.

El resultado es un algoritmo adaptado específicamente a paleontólogos y sitios de importancia en esta rama de las ciencias naturales, que se deriva a las categorías Personas y Lugares. De igual manera, el algoritmo responde a niveles de conocimientos diseñados por el propio experto, para los usuarios finales de la aplicación. Como un extra, aprovechando las relaciones entre las entidades y los beneficios de trabajar con datos estructurados, son creadas y sugeridas nuevas preguntas, que parten de propiedades conocidas para los tipos de entidades de la experiencia.

A continuación, se resumen las características y opciones que proporciona el algoritmo diseñado:

- Generación de entidades candidatas a distractores, de acuerdo con un texto de entrada.
- Selección de tipos de respuestas.

- Ajuste de hasta tres niveles de dificultad.
- En los casos que se requiera y las condiciones lo permita, sugerencia de nuevos cuestionarios relacionados con la temática, aportando las preguntas, las respuestas y los distractores.
- Configuración a medida del panorama de recopilación de entidades, por medio de Wikicats.
- Selección de cantidad de tipos en común, el cual a mayor valor proporciona más rigor en la selección de artículos para agregar a la experiencia.
- Posibilidad de exportar los distractores en diferentes formatos.

En un principio, estos pasos y métodos pueden ser aplicados a diferentes ramas del conocimiento, y sobre cualquier temática, aunque los resultados alcanzasen a ser mejores o peores de acuerdo con la información existente en la Web Semántica, específicamente en la plataforma DBpedia. Con lo cual, realizando ligeros ajustes, el algoritmo pudiera emplearse en otros tipos de experiencias, ya que posee una gran versatilidad al solo depender, en principio, de un texto concreto. A continuación, se enumeran los cambios al algoritmo diseñado en este trabajo, para ser aplicado sobre otras experiencias:

1. Definir los tipos de respuestas esperadas: personas, lugares, libros, películas, eventos, pinturas, etc. Esto es necesario para la etapa de filtrado, ya que, para establecer correspondencias, entre la respuesta correcta y los distractores, hay que verificar las propiedades comunes y únicas para un tipo de entidad.
2. El texto debe contener los tipos de entidades semejantes a las respuestas correctas que se esperan.
3. Cambios en el algoritmo de definición de preguntas sugeridas, sección 3.3.4, para explorar propiedades específicas de los tipos de respuestas que permitan la generación de preguntas interesantes para el jugador.

Como se puede observar lo cambios son mínimos, ya que el algoritmo se diseñó lo suficientemente abierto para no depender totalmente de la experiencia, solo en cuestiones imprescindibles. Con lo cual, se consideró innecesario aplicar los tres puntos

anteriores, para un marco definido desde el inicio por la aplicación en donde se pretender implementar, que es en el Museo de Ciencias Naturales de Madrid.

Capítulo 5 -

Introduction

General Planning

The use of new technologies is key to improving the user experience in museums, as well as taking advantage of and exploiting opportunities in new forms of interactive cultural activities. Memory institutions need to maintain, if not reinforce, their attractiveness and the interest of their visitors, especially for new generations.

According to the Statutes of the International Council of Museums (ICOM), adopted by the 22nd General Assembly in Vienna, Austria, on August 24, 2007, the definition of a museum is "a permanent, non-profit institution at the service of society and its development, open to the public, which acquires, conserves, research, communicates and exhibits the tangible and intangible heritage of humanity and its environment for the purpose of education, study and enjoyment" [1]. Jette Sandahl, Danish curator who heads the ICOM commission, argues that the current definition "does not speak the language of the 21st century" by ignoring the demands of "cultural democracy". According to the article published by the curator herself [2], a new definition must be designed that fits the museums of the 21st century, that recognizes their existence as diverse and changing societies, and that supports the development of new paradigms.

It is clear that memory institutions have become aware of the complexities of today's world. Having to cope with new visitor preferences, rivaling the cinema, theaters, shopping malls, and an even greater challenge, the internet and all that implies access to information, from an electronic device. The function of conserving and preserving heritage is still the most important thing, the only thing that has changed is the way museum workers and other specialists decide to approach the public [3].

The term "Semantic Web", as defined by the World Wide Web Consortium (W3C), is the vision of the W3C when referring to the Web of linked data. Semantic Web technologies allow people to create data stores on the Web, build vocabularies and

write rules to manage information. Linked data is powered by technologies such as RDF, SPARQL, OWL and SKOS [4].

Another way to understand the meaning of Semantic Web can be defined in [5], abstractly as the alliance between knowledge representation and tools powerful enough to allow reasoning about online data. Strictly, the alliance between languages and semantic web applications for sharing, analyzing, and processing data.

Currently, the Semantic Web and its benefits are exploited for uses of different types in museums, as is the case of the Knowledge Graph of the Prado Museum, built on Semantic Web standards and according to the principles of the Web of Linked Data [6]. The British Museum created a new version of its database based on the Semantic Web [7].

Publishing as open data is growing in recent years by all kinds of organizations such as municipalities, museums and institutions related to cultural heritage. So this project adds to this line of these initiatives and aims to encourage them to join the revolution and digitization of memory institutions.

Motivation

Recent research on interactive digital storytelling, personalization, adaptability, and mixed reality, together with systems that enable mobility, promise not only to make cultural heritage sites more attractive, but also to provide new means for conveying knowledge, interpretation, and analysis [8].

With recent advances in Semantic Web and Deep Learning technologies, a new field of application opens, due to the existence of large amounts of digital information. The case on which this work is based is the National Museum of Natural Sciences (MNCN), one of the oldest Natural History museums in Europe and the most important in Spain. The MNCN is currently operating a project to promote informal learning through a treasure hunt type game. The MNCN Enigma game is a treasure hunt for mobile devices [9], which combines the fun, through mini-games, with the didactic through the reading of writings established in the history of the game.

Objectives

The general objective of this work is to create a semi-automatic procedure, where experts have punctual interactions, using Semantic Web technologies and Deep Learning techniques. With an input text, a question, which should be related to that story, and knowing the correct answer, generate options to possible incorrect answers or as they will be called in the work distractors. The distractors are the group of entities, paleontologists, or places of importance in paleontology according to each question, which will have the function of confusing the players when they must select the correct answer, in questions that present four options.

The term experience in the work will be defined as the framework of application of the work and in which the algorithm tests were performed, multiple choice questions. The questions allowed in our experience will be those in which the answers are paleontologists and places that appear as relevant in this branch of natural sciences.

The specific objectives are the following:

- **Study of the tools used:** to deepen in the characteristics of each of the libraries and platforms related to the Semantic Web and Deep Learning techniques used in this work.
- **Create a corpus tailored to the subject matter:** design an algorithm that allows the collection of Wikipedia articles related to the experience to be developed. The process must incorporate a filtering stage and text processing techniques, with the objectives of creating a homogeneous vocabulary and increasing cooccurrences in the corpus.
- **Train a model with neural networks:** generate a model with a large vocabulary and that matches the experience, using the Word2vcet algorithm, with the final objective of calculating the similarities between entities.
- **Predict distractors according to the experience indicated by the expert:** collect the entities related to each of the correct answers of the questionnaires defined by the specialist. Filter and display each set of options, where each distractor is related to the correct answer.

- **Generate levels of knowledge for users:** establish levels of difficulty considering the similarities between the distractors and the correct answer, and the level of inlinks or popularity of each entity.
- **Evaluate the results obtained:** evaluate the behavior and results, by submitting the complete algorithm to different experiences and questionnaires.

Work plan and structure

To achieve each of the objectives set out above, the first thing is to be clear about what experience you want to work on, and how to approach it. In the case of this project, two types of responses will be addressed, scientists linked to paleontology and places with significance for this branch of science. Therefore, the process of writing the text that will feed the training of the model must be carried out by a specialist, who must be very clear about the characteristics of the questionnaires he/she intends to design. This phase is the most important and will have an impact on the rest of the stages.

In the following link, https://drive.google.com/drive/folders/1RiaWRKMcrBwR6QmiMWSiV4BFZ1117_aE?usp=sharing, you will find the programming of the application used as support for this work, where its use is allowed locally. In this shared folder in Drive, if you wish to look at the questionnaires, you will find them with their results by stages. And if you want to run it online, you can also find the version in Google Colab. Everything is under open-source license, both the libraries and the development of the application in general. Similarly, there is a Readme.txt file, which details the structure of the folder.

This work is organized in four chapters, the first one, which is the current one, is directed to make an introduction that collects the general approach, motivations and objectives outlined in the project. The second chapter describes each of the technologies and tools that are used, presenting a summary of works related to the subject of the project. Chapter three consists of the implementation of the proposed objectives, going in depth in each of the stages of the process, culminating with the analysis of the results obtained. Finally, the conclusions, where a balance of the work is made in a comprehensive manner.

Capítulo 6 -

Chapter - Conclusions

The present work provides a complete and tested algorithm to collect incorrect answers to questions for which the answer is known. It aims to automate and speed up the massive definition of questionnaires, specifically as a proposal to the Enigma MNCN mobile application, a project currently in operation.

The solution starts from raw texts hosted in Wikipedia, and uses the relationships between LOD structured data, provided by platforms such as DBpedia and Wikidata, through SPARQL queries. For the exploitation of the mentioned relationships, entity recognition techniques (NER) are used, an option provided by DB-SL, as well as disambiguation between semantic surfaces. For text processing, morphological feature labeling, dependency analysis and lemmatization, we take advantage of the Python language library, Stanza, which is part of Stanford's powerful NPL Group toolkit. To complement this, with the precise objective of calculating similarities between entities, Word2vect provides the means to define semantic distances between words.

The result is an algorithm specifically adapted to paleontologists and sites of importance in this branch of the natural sciences, which is derived to the categories People and Places. Similarly, the algorithm responds to levels of knowledge designed by the expert himself, for the end users of the application. As an extra, taking advantage of the relationships between entities and the benefits of working with structured data, new questions are created and suggested, starting from known properties for the types of entities of the experience.

The following is a summary of the features and options provided by the designed algorithm:

- Generation of candidate entities for distractors, according to an input text.
- Selection of response types.
- Adjustment of up to three levels of difficulty.

- Where required and conditions permit, suggestion of new questionnaires related to the subject matter, providing the questions, answers and distractors.
- Tailor-made configuration of the entity collection landscape, by means of Wikicats.
- Selection of the number of types in common, the higher the value, the more rigorous the selection of items to add to the experience.
- Possibility of exporting the distractors in different formats.

In principle, these steps and methods can be applied to different branches of knowledge, and on any subject, although the results could be better or worse according to the existing information in the Semantic Web, specifically in the DBpedia platform. Thus, by making slight adjustments, the algorithm could be used in other types of experiences, since it has a great versatility as it only depends, in principle, on a specific text. The following is a list of changes to the algorithm designed in this work, so that it can be applied to other experiences:

1. Define the types of expected responses: people, places, books, movies, events, paintings, etc. This is necessary for the filtering stage, since, to establish correspondences between the correct answer and the distractors, it is necessary to verify the common and unique properties for a type of entity.
2. The text should contain the types of entities like the expected correct answers.
3. Changes in the algorithm for defining suggested questions, section 3.3.4, to explore specific properties of the types of answers that allow the generation of interesting questions for the player.

As can be seen, the changes are minimal, since the algorithm was designed to be sufficiently open so as not to depend totally on experience, only on essential questions. Therefore, it was considered unnecessary to apply the three previous points, for a framework defined from the beginning, by the application where it is intended to be implemented, which is in the Museum of Natural Sciences of Madrid.

BIBLIOGRAFÍA

[1] «Museum Definition», *ICOM*. https://icom.museum/en/resources/standards-guidelines/museum-definition/ (accedido jun. 14, 2021).

[2] «J. Sandahl - The Museum Definition as the Backbone of ICOM». https://click.endnote.com/viewer?doi=10.1080%2F13500775.2019.1638019&token=WzMyNDUwODQsIjEwLjEwODAvMTM1MDA3NzUuMjAxOS4xNjM4MDE5Il0.jdwYK44jBgSCeoTQh7eEvW06Iyk (accedido jun. 14, 2021).

[3] A. Gheorghilaș, «THE CHALLENGES OF THE 21ST-CENTURY MUSEUM: DEALING WITH SOPHISTICATED VISITORS IN A SOPHISTICATED WORLD», p. 13, 2017.

[4] «Semantic Web - W3C». https://www.w3.org/standards/semanticweb/ (accedido jun. 14, 2021).

[5] S. Varlan, «ADVANTAGES OF SEMANTIC WEB TECHNOLOGIES IN THE KNOWLEDGE BASED SOCIETY», *Sci. Ann. Alexandru Ioan Cuza Univ. Iasi Econ. Sci. Ser.*, ene. 2010.

[6] «Estándares semánticos y datos enlazados - Museo Nacional del Prado». https://www.museodelprado.es/grafo-de-conocimiento/estandares-semanticos-y-datos-enlazados (accedido jun. 14, 2021).

[7] «Semantic tools in ResearchSpace», *ResearchSpace*. http://researchspace.org/semantic-tools/ (accedido jun. 14, 2021).

[8] M. Majd y R. Safabakhsh, «Impact of machine learning on improvement of user experience in museums», en *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, Shiraz, oct. 2017, pp. 195-200. doi: 10.1109/AISP.2017.8324080.

[9] I. Camps-Ortueta, P. A. González-Calero, M. A. Quiroga, y P. P. Gómez-Martín, «Measuring Preferences in Game Mechanics: Towards Personalized Chocolate-Covered Broccoli», en *Entertainment Computing and Serious Games*, Cham, 2019, pp. 15-27. doi: 10.1007/978-3-030-34644-7_2.

[10] «DBpedia versión 2016-10 | DBpedia». https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10 (accedido mar. 26, 2021).

[11] E. Blomqvist *et al.*, Eds., *Semantic Systems. In the Era of Knowledge Graphs: 16th International Conference on Semantic Systems, SEMANTiCS 2020, Amsterdam, The Netherlands, September 7–10, 2020, Proceedings*, vol. 12378. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-59833-4.

[12] «About DBpedia», *DBpedia Organization*. https://www.dbpedia.org/about/ (accedido mar. 27, 2021).

[13] J. Lehmann *et al.*, «DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia», *Semantic Web*, vol. 6, n.º 2, pp. 167-195, 2015, doi: 10.3233/SW-140134.

[14] T. Mikolov, K. Chen, G. Corrado, y J. Dean, «Efficient Estimation of Word Representations in Vector Space», *ArXiv13013781 Cs*, sep. 2013, Accedido: abr. 01, 2021. [En línea]. Disponible en: http://arxiv.org/abs/1301.3781

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, y J. Dean, «Distributed Representations of Words and Phrases and their Compositionality», *ArXiv13104546 Cs Stat*, oct. 2013, Accedido: abr. 01, 2021. [En línea]. Disponible en: http://arxiv.org/abs/1310.4546

[16] T. Mikolov, Q. V. Le, y I. Sutskever, «Exploiting Similarities among Languages for Machine Translation», *ArXiv13094168 Cs*, sep. 2013, Accedido: abr. 01, 2021. [En línea]. Disponible en: http://arxiv.org/abs/1309.4168

[17] «Word Embedding with Skip-Gram Word2Vec», mar. 31, 2019. https://maelfabien.github.io/machinelearning/NLP_3/ (accedido jun. 13, 2021).

[18] «Wikidata:Introduction - Wikidata». https://www.wikidata.org/wiki/Wikidata:Introduction/en (accedido jun. 13, 2021).

[19] «Overview», *Stanza*. https://stanfordnlp.github.io/stanza/ (accedido may 03, 2021).

[20] A. Scharl, M. Sabou, y M. Föls, «Climate quiz: a web application for eliciting and validating knowledge from social networks», en *Proceedings of the 18th Brazilian symposium on Multimedia and the web*, New York, NY, USA, oct. 2012, pp. 189-192. doi: 10.1145/2382636.2382677.

[21] C. Lin, D. Liu, W. Pang, y Z. Wang, «Sherlock: A Semi-automatic Framework for Quiz Generation Using a Hybrid Semantic Similarity Measure», *Cogn. Comput.*, vol. 7, n.º 6, pp. 667-679, dic. 2015, doi: 10.1007/s12559-015-9347-7.

[22] B. Varela-Brea, M. López-Nores, Y. Blanco-Fernández, J. J. Pazos-Arias, A. Gil-Solla, y M. Ramos-Cabrer, «Deep Guessing: Generating Meaningful Personalized Quizzes on Historical Topics by Introducing Wikicategories in Doc2Vec», en *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, sep. 2018, pp. 43-47. doi: 10.1109/SMAP.2018.8501891.

[23] M. López-Nores *et al.*, «Technology-Powered Strategies to Rethink the Pedagogy of History and Cultural Heritage through Symmetries and Narratives», *Symmetry*, vol. 11, n.º 3, Art. n.º 3, mar. 2019, doi: 10.3390/sym11030367.

[24] «CrossCult: Empowering reuse of digital cultural heritage in context-aware crosscuts of European history | CROSSCULT Project | H2020 | CORDIS | European Commission». https://cordis.europa.eu/project/id/693150 (accedido jun. 01, 2021).

[25] Y. Blanco-Fernández, A. Gil-Solla, J. J. Pazos-Arias, M. Ramos-Cabrer, A. Daif, y M. López-Nores, «Distracting users as per their knowledge: Combining linked open data and word embeddings to enhance history learning», *Expert Syst. Appl.*, vol. 143, p. 113051, abr. 2020, doi: 10.1016/j.eswa.2019.113051.

[26] B. Iancu, «A Trivia like Mobile Game with Autonomous Content That Uses Wikipedia Based Ontologies», *Inform. Econ.*, vol. 19, n.º 1/2015, pp. 25-33, mar. 2015, doi: 10.12948/issn14531305/19.1.2015.02.

[27] M. Foulonneau, «Generating Educational Assessment Items from Linked Open Data: The Case of DBpedia | EndNote Click». https://click.endnote.com/viewer?doi=10.1007%2F978-3-642-25953-1_2&token=WzMyNDUwODQsIjEwLjEwMDcvOTc4LTMtNjQyLTI1OTUzLTFfMiJd.AGNFpoEyshymk2ILLAtV2Lg8zhI (accedido jun. 12, 2021).

[28] J. Szymański, «Towards Automatic Classification of Wikipedia Content | EndNote Click». https://click.endnote.com/viewer?doi=10.1007%2F978-3-642-15381-5_13&token=WzMyNDUwODQsIjEwLjEwMDcvOTc4LTMtNjQyLTE1MzgxLTVfMTMiXQ.o1QZB1F31vu8b4bzhR9T_C3KFZc (accedido jun. 01, 2021).

[29] «Gensim: topic modelling for humans». https://radimrehurek.com/gensim/models/word2vec.html (accedido may 03, 2021).

[30] «Penn Treebank P.O.S. Tags». https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html (accedido may 04, 2021).

[31] M. J. Kusner, Y. Sun, N. I. Kolkin, y K. Q. Weinberger, «From Word Embeddings To Document Distances», p. 10.

[32] «An Exploration in Earth & Word Movers Distance • Peter Baumgartner», *Peter Baumgartner*, jun. 18, 2017. https://pmbaumgartner.github.io/blog/word-movers-distance-exploration/ (accedido jun. 12, 2021).

[33] «Word Mover's Embedding: Universal Text Embedding from Word2Vec», *IBM Research Blog*, nov. 01, 2018. https://www.ibm.com/blogs/research/2018/11/word-movers-embedding/ (accedido jun. 12, 2021).

[34] «Gensim: topic modelling for humans». https://radimrehurek.com/gensim/auto_examples/tutorials/run_wmd.html (accedido jun. 05, 2021).

[35] «About: American geologists». https://dbpedia.org/page/Category:American_geologists (accedido jun. 19, 2021).

[36] «Charles Doolittle Walcott», *Wikipedia*. jun. 05, 2021. Accedido: jun. 10, 2021. [En línea]. Disponible en: https://en.wikipedia.org/w/index.php?title=Charles_Doolittle_Walcott&oldid=1026965295

[37] «Benjamin Franklin Mudge», *Wikipedia*. abr. 19, 2021. Accedido: jun. 10, 2021. [En línea]. Disponible en: https://en.wikipedia.org/w/index.php?title=Benjamin_Franklin_Mudge&oldid=1018803165

[38] «Charles Hazelius Sternberg», *Wikipedia*. dic. 19, 2020. Accedido: jun. 10, 2021. [En línea]. Disponible en:

https://en.wikipedia.org/w/index.php?title=Charles_Hazelius_Sternberg&oldid=995173439

[39] «About: Limestone caves». https://dbpedia.org/page/Category:Limestone_caves (accedido jun. 19, 2021).

APÉNDICES

Apéndice A - Cuestionarios

Paleontólogos

Cuestionario 1

Text_context: Jean Léopold Nicolas Frédéric, Baron Cuvier, known as Georges Cuvier, was a French naturalist and zoologist, sometimes known as the "founding father of paleontology". Cuvier was a major figure in natural science research in the early 19th century and was instrumental in establishing the fields of comparative anatomy and paleontology through his work in comparing living animals with fossils. Cuvier's work is considered the foundation of vertebrate paleontology, and he expanded Linnean taxonomy by grouping classes into phyla and incorporating both fossils and living species into the classification.

Question: Who is known as the founding father of paleontology?

Answer: Georges Cuvier

Distractors: Loren Eiseley, Richard Owen, Jean-Baptiste Lamarck, Henri Marie Ducrotay de Blainville, Charles Lyell, Édouard Lartet, Gotthelf Fischer von Waldheim, Pierre-Joseph van Beneden, Louis Agassiz, Pierre André Latreille, …

Cuestionario 2

Text_context: The Bone Wars, also known as the Great Dinosaur Rush, was a period of intense and ruthlessly competitive fossil hunting and discovery during the Gilded Age of American history, marked by a heated rivalry between Edward Drinker Cope (of the Academy of Natural Sciences of Philadelphia) and Othniel Charles Marsh (of the Peabody Museum of Natural History at Yale). Marsh managed to find 80 new dinosaur species, while Cope found 56. The efforts of both men led to the discovery and description of more than 136 new dinosaur species, 80 by Marsh's team and the rest attributed to Cope. The products of the Bone Wars resulted in an increase in knowledge

of prehistoric life, and sparked the public's interest in dinosaurs, leading to continued fossil excavation in North America in the decades to follow.

Question: Who is considered the winner of the bone wars?

Answer: Othniel Charles Marsh

Distractors: Charles Doolittle Walcott, Benjamin Franklin Mudge, Charles Hazelius Sternberg, Edward Drinker Cope, Samuel Wendell Williston, John Ostrom, James Dwight Dana, Joseph Henry, Tilly Edinger, John Bell Hatcher, …

Cuestionario 3

Text_context: Tyrannosaurus is a genus of coelurosaurian theropod dinosaur. Fossils are found in a variety of rock formations dating from the Maastrichtian age of the Late Cretaceous period, 68 to 66 million years ago. It was the last known member of the tyrannosaurids and one of the last non-avian dinosaurs to exist before the Cretaceous-Paleogene extinction event. Barnum Brown, assistant curator of the American Museum of Natural History, found the first partial skeleton of T. rex in eastern Wyoming in 1900.

Question: Who is credited as the discoverer of the tyrannosaur?

Answer: Barnum Brown

Distractors: John Bell Hatcher, Al Smith, Joseph Leidy, George Gaylord Simpson, Othniel Charles Marsh, Samuel Wendell Williston, Kenneth Carpenter, Preston Cloud, Edward Drinker Cope, Henry Fairfield Osborn

Cuestionario 4

Text context: Gideon Algernon Mantell MRCS FRS was an English obstetrician, geologist, and paleontologist. His attempts to reconstruct the structure and life of Iguanodon initiated the scientific study of dinosaurs: in 1822 he was responsible for the discovery (and eventual identification) of the first fossil teeth, and later much of the skeleton, of Iguanodon. Mantell's work on the Cretaceous of southern England was also important.

Question: What is the name of the paleontologist who discovered the first fossil teeth?

Answer: Gideon Mantell

Distractors: Richard Owen, Mary Anning, Edward Forbes, Joseph Dalton Hooker, Etheldred Benett, Charles Lyell, Samuel Stutchbury, George Rolleston, James Parkinson, Richard Owen, ….

Cuestionario 5

Text_context: The discovery of Lystrosaurus fossils at Coalsack Bluff in the Transantarctic Mountains by Edwin H. Colbert and his team in 1969–70 helped support the hypothesis of plate tectonics and strengthen the theory, since Lystrosaurus had already been found in the lower Triassic of southern Africa as well as in India and China.

Question: Who is the paleontologist who proved that Africa and Antarctica used to be united in one gigantic land mass?

Answer: Edwin H. Colbert

Distractors: Roy Chapman Andrews, Edward Drinker Cope, William King Gregory, John Ostrom, Barnum Brown, Zdeněk Burian, Marshall Kay, Vladimir Zherikhin, Edward Hitchcock, Robert Broom, …

Cuestionario 6

Text_context: The Atapuerca archaeological site is in the Sierra de Atapuerca in northern Spain. The archaeological importance of the area became increasingly evident during the construction of a railway line. Further extended campaigns and interdisciplinary work has been carried out by several teams, led by Emiliano Aguirre from 1978 to 1990, considered the formal initiator of the research, and later jointly by Eudald Carbonell, José María Bermúdez de Castro and Juan Luis Arsuaga.

Question: Spanish paleontologist who is known for being the initiator of the study of the Pleistocene sites in the Sierra de Atapuerca?

Answer: Emiliano Aguirre

Distractors: Eudald Carbonell, Donald Prothero, Alberto Angela, Trevor H. Worthy, Miquel Crusafont i Pairó, Philip D. Gingerich, Tim Flannery, Natascha Heintz, Mark Norell, …

Cuestionario 7

Text context: In 1888 John Bell Hatcher was sent to Wyoming, to discover the origins of some very peculiar horned fossils that Othniel Charles Marsh had sent. Here he worked for four years and managed to find 33 ceratopsian skulls and the remains of 17 other individuals. He established himself as one of the greatest dinosaur hunters of all time, referred to as the "King of the Collectors" by none other than Marsh himself. In 1889 near Lusk, Wyoming, Hatcher excavated the first fossil remains of Torosaurus.

Question: Paleontologist known as the king of collectors and for discovering two dinosaur genera Torosaurus and Triceratops?

Answer: John Bell Hatcher

Distractors: George Jarvis Brush, John Kerry, Georg Baur, Charles Hazelius Sternberg, Edward Drinker Cope, Othniel Charles Marsh, Donald Prothero, Elmer S. Riggs, John Mason Clarke, Walter W. Granger, …

Cuestionario 8

Text context: Talbot recovered and named the only known fossils of the dinosaur Podokesaurus holyokensis, found near Mount Holyoke College in 1910, and published a scientific description of the specimen in 1911 and thus became the first woman to discover and name a non-bird dinosaur. Podokesaurus is a genus of cellophisoid dinosaur that lived in what is now the eastern United States during the Early Jurassic Period.

Question: Female paleontologist who recovered and named the dinosaur Podokesaurus holyokensis and 1909 became the first woman elected to membership in the Paleontological Society?

Answer: Mignon Talbot

Distractors: Henry Nathaniel Andrews, Meemann Chang, Gerta Keller, John Alroy, Charles Repenning, Hildegarde Howard, Timothy Abbott Conrad, Gloria Jean Siebrecht, Truman H. Aldrich, Amadeus William Grabau, …

Cuestionario 9

Text_context: Sarcosuchus was a giant relative of crocodylians , and it is estimated that fully developed individuals reached 9 to 9.5 m (29.5 to 31.2 ft) in total length and 3.5 to 4.3 metric tons (3.9 to 4.7 short tons) in weight. In 1964, a nearly complete skull was found in Niger by the French CEA, but it was not until 1997 and 2000 that science learned most of its anatomy, when an expedition led by American paleontologist Paul Sereno discovered six new specimens, including one with about half of the skeleton intact and most of the spinal column.

Question: Name of the Paleontologist did not discover it, but he did find the most complete skeleton of Sarcosuchus or also known as SuperCoc?

Answer: Paul Sereno

Distractors: Dong Zhiming, Gianluigi Buffon, Stephen L. Brusatte, Barack Obama, Robert T. Bakker, John Ostrom, Roy Chapman Andrews, Derek Briggs, Charles Mortram Sternberg, Michael Benton, …

Cuestionario 10

Text_context: Ichthyornis has been historically important in shedding light on the evolution of birds. It was the earliest known prehistoric bird relative preserved with teeth. Ichthyornis is believed to have been the Cretaceous ecological equivalent of modern seabirds such as gulls, petrels, and skimmers. An average specimen was the size of a pigeon, 24 centimeters (9.4 inches) long, with a skeletal wingspan (excluding feathers) of about 43 centimeters (17 inches). was discovered in 1870 by Benjamin Franklin Mudge, a professor at Kansas State Agricultural College who recovered the initial fossils from the North Fork of the Solomon River in Kansas, USA.

Question: What is the name of the paleontologist that one of his most relevant findings was the first recognized bird with teeth?

Answer: Benjamin Franklin Mudge

Distractors: Charles Hazelius Sternberg, Samuel Wendell Williston, Benjamin Franklin, Stephen Jay Gould, Richard Swann Lull, Charles Hazelius Sternberg, Charles Doolittle Walcott, Paul Sereno, Charles W. Gilmore, John Ostrom

Sitios Paleontológicos

Cuestionario 1

Text context: Located near Santillana del Mar in Cantabria, Spain, the Altamira cave is a treasure trove of information about life in the Paleolithic period. From rudimentary stone tools to carved bones, this place houses many artifacts that give us an insight into daily life during the Stone Age. Altamira Cave has the privilege of being the first place in the world where the existence of Upper Paleolithic Rock Art was identified. Altamira was also a unique discovery because of the quality, the magnificent preservation, and the freshness of its pigments. The caves are currently protected by UNESCO. After being recognized as a Historic Artistic Monument, in 1985 it received the title of World Heritage Site.

Question: Which archaeological site has the privilege of being the first place in the world where the existence of Upper Paleolithic Rock Art was identified

Answer: Cave of Altamira

Distractors: Sandia Cave, Cave of El Castillo, Plovers Lake, Makapansgat, Cave of La Pasiega, Wookey Hole Caves, Kromdraai fossil site, Rising Star Cave, Sistema Ox Bel Ha, Gladysvale Cave

Cuestionario 2

Text context: The Gogo Formation in the Kimberley region of Western Australia is a Lagerstätte exhibiting exceptional preservation of a Devonian reef community. The fossils of the Gogo Formation show the three-dimensional preservation of soft tissues as fragile as nerves and embryos with umbilical cords. The discovery of Materpiscis, a preserved placoderm with an embryonic juvenile still attached by its umbilical cord, has revealed that at least some placoderms gave birth to live young.

Question: What is the fossil site where Materpiscis was found, a placoderm preserved with an embryonic juvenile still attached by its umbilical cord?

Answer: Gogo Formation

Distractors: Burgess Shale, Sirius Passet, La Brea Tar Pits, Walcott–Rust quarry, London Clay, Kaili Formation, Hunsrück Slate, Mazon Creek fossil beds, Doushantuo Formation, Mangrullo Formation

Cuestionario 3

Text context: Mammoth Cave National Park is an American national park in west-central Kentucky, encompassing portions of Mammoth Cave, the longest cave system known in the world. The purpose of Mammoth Cave National Park is to preserve, protect, interpret, and study the internationally recognized biological and geologic features and processes associated with the longest known cave system in the world, the park's diverse forested, karst landscape, the Green and Nolin rivers, and extensive evidence of human history; and to provide and promote public enjoyment, recreation, and understanding.

Question: What is the name of the most extensive cave system in the world?

Answer: Mammoth Cave National Park

Distractors: Fisher Ridge Cave System, Yosemite National Park, Yellowstone National Park, Carlsbad Caverns National Park, Sistema Ox Bel Ha, Jenolan Caves, Wookey Hole Caves, Poverty Point, Sandia Cave, Mammoth Cave Parkway

Cuestionario 4

Text context: Blackberry Hill is a Konservat-Lagerstätte of Cambrian age located within Marathon County, Wisconsin. It is found in a series of quarries and outcrops that are notable for their large concentration of exceptionally preserved trace fossils in Cambrian tidal flats. One quarry in particular also has the distinction of preserving some of the first land animals. These are preserved as three-dimensional casts, which is unusual for Cambrian animals that are only lightly biomineralized. Additionally, Blackberry Hill is the first occurrence recognized to include Cambrian mass strandings of scyphozoans (jellyfish).

Question: What is the fossil deposit that is recognized as the first place where the first massive strandings of Cambrian scyphozoans (jellyfish) appeared?

Answer: Blackberry Hill

Distractors: Mangrullo Formation, Two Creeks Buried Forest State Natural Area, Ashfall Fossil Beds, Hunsrück Slate, Doushantuo Formation, Gogo Formation, La Brea Tar Pits, Mazon Creek fossil beds, Miguasha National Park, London Clay

Cuestionario 5

Text context: The Djadochta Formation is a geological formation situated in central Asia (Gobi Desert), dating from the Late Cretaceous Period. It preserves an arid habitat of sand dunes, with little freshwater apart from oases and arroyos. In fact, the present-day climate at most Djadochta Formation sites differs little from what it was some 80 mya, except by being somewhat warmer and perhaps a bit less arid then. Most notable fossil discoveries have been the first confirmed dinosaur eggs (a clutch, probably of Oviraptor) and several dinosaurs finds, Protoceratops, Pinacosaurus and Velociraptor being the most prominent. The holotype specimen of Halszkaraptor probably came from the Djadochta Formation at Ukhaa Tolgod in southern Mongolia and was illegally extracted by fossil poachers.

Question: What is the geological formation where the first confirmed dinosaur eggs were found?

Answer: Djadochta Formation

Distractors: Nemegt Formation, Horseshoe Canyon Formation, Bajo de la Carpa Formation, Cerrejón Formation, Anacleto Formation, Cloverly Formation, Hanson Formation, Gobi Desert, Old Red Sandstone, Caturrita Formation

Cuestionario 6

Text context: A team of scientists unveiled the world's oldest fossils: ruins of colonies of ancient bacteria known as stromatolites in 3.7-billion-year-old rocks in Greenland, 200 million years older than the 3.48-billion-year-old fossil stromatolites discovered in the Pilbara Craton in northwestern Australia. The earliest direct evidence of life on Earth may be fossils of permineralized microorganisms in Australian Apex flint rocks.

Question: What is the name of the area or region where the oldest fossils on Earth were found and is said to have found Halszkaraptor, the first non-avian dinosaur that could move both on land and in water?

Answer: Pilbara Craton

Distractors: Gogo Formation, Paleontological Site Chiniquá, Paleontological Site Arroio Cancela

Cuestionario 7

Text context: The local area is famous in geology and palaeontology for Solnhofen limestone. This is a very fine-grained limestone from the Jurassic period which is an exceptionally fine Lagerstätte that preserves detailed fossil specimens. Alois Senefelder used specially prepared blocks of the fine Solnhofen limestone for the process of lithography which he invented in 1798. The quarrying of this lithographic limestone subsequently yielded spectacular finds, including Archaeopteryx, commemorated in the bird's full name Archaeopteryx lithographica. All 13 known specimens have come from the Solnhofen area.

Question: What is the name of the fossil site where the dinosaur species Archaeopteryx, known as the perfect "missing link" between dinosaurs and birds, was found?

Answer: Solnhofen Limestone

Distractors: Gogo Formation, Anacleto Formation, Allen Formation, Mangrullo Formation, Lameta Formation, Messel pit

Cuestionario 8

Text context: Turkana Boy, also called Nariokotome Boy, is the name given to fossil KNM-WT 15000, a nearly complete skeleton of a Homo ergaster (alternatively referred to as African Homo erectus) youth who lived at c. 1.5 to 1.6 million years ago. This specimen is the most complete early human skeleton ever found. It was discovered in 1984 by Kamoya Kimeu on the bank of the Nariokotome River near Lake Turkana in Kenya.

Question: What is the name of the place where the skeleton of a child was discovered, which is considered the most complete skeleton of a prehistoric human ever found.?

Answer: Lake Turkana

Distractors: Omo River, Gilgel Gibe III Dam, Lomekwi, Kariandusi prehistoric site, Omo River

Cuestionario 9

Text context: The Atapuerca site in Burgos (Spain) is more than 400,000 years old, the Sima de los Huesos is the largest human fossil site in history. The first human fossils were found in 1976, but it was hard to imagine the vast amount of paleontological material that lay buried in its sediments. The site is full of human bones attributed to Homo heidelbergensis (considered the ancestor of Homo neanderthalensis), dating back some 300,000 years. It houses 2,000 bones belonging to at least 32 individuals. "All the human species that lived in Europe are represented in Atapuerca which, after more than 40 years of excavations, 99% of the fossils are still buried.

Question: What is the name of the fossil site where all the human species that lived in Europe were found?

Answer: Archaeological site of Atapuerca

Distractors: Sidrón Cave, Vindija Cave, Betal Rock Shelter, Abrigo do Lagar Velho, Burgos, Cave of Altamira, Es-Skhul Cave, Mezmaiskaya cave, Shanidar Cave, Denisova Cave

Cuestionario 10

Text context: Sinosauropteryx was the first in a spectacular series of dinosaur discoveries in the Yixian Formation in China's Liaoning Province. The fossil reveals the unmistakable impression of primitive hair-like feathers; it was the first-time paleontologists had directly identified this feature in a dinosaur. Unexpectedly, an analysis of the remains of Sinosauropteryx showed that it was only related in isolation to another famous feathered dinosaur, Archaeopteryx, leading paleontologists to revise their theories about how and when dinosaurs became birds. This dinosaur lived in the mid-Cretaceous period, approximately 120 million years ago. Sinosauropteryx lived in what is now northeastern China during the Early Cretaceous period.

Question: What is the name of the fossil site where Sinosauropteryx was found, which revealed the unmistakable impression of primitive hair-like feathers?

Answer: Yixian Formation,

Distractors: Tiaojishan Formation, Haifanggou Formation, Huajiying Formation, Dabeigou Formation, Sunjiawan Formation, Tiaojishan Formation, Green River Formation, Messel pit, Hamilton Quarry, Haman Formation