EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Generación de preguntas y respuestas con información de Wikipedia aplicadas a través de un chatbot

presentada por

IINF. Juan Jesús Sandoval Villanueva

como requisito para la obtención del grado de

Maestro en Ciencias de la Computación

Director de tesis

Dr. Noé Alejandro Castro Sánchez

Codirector de tesis

Dr. Héctor Jiménez Salazar

Cuernavaca, Morelos, México. Febrero de 2021.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Centro Nacional de Investigación y Desarrollo Tecnológico
Departamento de Ciencias Computacionales

Cuernavaca, Mor., 19/enero/2021

OFICIO No. DCC/016/2021
Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

C. DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del **C. Ing. Juan Jesús Sandoval Villanueva**, con número de control M18CE018, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Generación de preguntas y respuestas con información de wikipedia aplicadas a través de un chatbot"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

Dr. Noé Alejandro Castro Sánchez
Doctor en Ciencias de la Computación
08701806
Director de tesis

Dr. Héctor Jiménez Salazar
Doctor en Ciencias
Co-director de tesis

Dr. Juan Gabriel González Serna
Doctor en Ciencias de la Computación
7820329
Revisor 1

Dr. Máximo López Sánchez
Doctor en Ciencias de la Computación
7498547
Revisor 2

C.c.p. Depto. Servicios Escolares
Expediente / Estudiante
JGGS/lmz

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico
Subdirección Académica

Cuernavaca, Mor., 25/enero/2021
No. de Oficio: SAC/25/2021
Asunto: Autorización de impresión de tesis

JUAN JESÚS SANDOVAL VILLANUEVA
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado ***"Generación de preguntas y respuestas con información de wikipedia aplicadas a través de un chatbot",*** ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
"Excelencia en Educación Tecnológica"
"Educación Tecnológica al Servicio de México"

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO SUBDIRECCIÓN ACADÉMICA

C.c.p. M.E. Guadalupe Garrido Rivera. Jefa del Departamento de Servicios Escolares
Expediente
CMAZ/CHG

Dedicatoria

A Dios por la salud y las bendiciones que me permitieron llegar hasta este momento, a mis padres que hicieron todo lo posible para cumplir esta meta, mi hermana que siempre está a mi lado apoyando y alentándome a ser mejor persona y a mi abuelita que con sus sabios consejos me ayuda a tomar buenas decisiones.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por su Programa Nacional de Posgrados de Calidad (PNPC) por medio del cual me fue otorgada una beca para ser estudiante de tiempo completo y desarrollar este trabajo de investigación. Al Tecnológico Nacional de México (TecNM) y al campus Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET).

A mi director de tesis, el Dr. Noé Alejandro Castro Sánchez, por apoyarme, por todas las reuniones y pláticas acerca del desarrollo del trabajo, por la disposición de ayudarme y animarme con la investigación, por sus consejos e indicaciones, por su respaldo en las iniciativas que se propusieron, por enseñarme cuan interesante es el campo del Procesamiento del Lenguaje Natural y la infinidad de tareas que pueden realizarse, por mostrarme que es posible lograrlo. A mi codirector, por sus aportes al motivarme a experimentar, por sus ideas, por su disponibilidad siempre presente a pesar de la distancia.

A mi comité tutorial conformado por el Dr. Máximo López Sánchez, quien me aconsejó, me dio palabras de aliento, por sus correcciones, por su clase siempre muy motivadora y de gran aporte a la formación científica; y al Dr. Juan Gabriel González Serna, quien me exigió, me corrigió, me hizo esforzarme, por sus clases, por sus consejos. A ellos quienes dedicaron parte de su tiempo a las revisiones de esta investigación y ayudaron a la mejora del mismo.

Resumen

En esta tesis se aborda el problema de la falta de herramientas para la creación automática de cuestionarios, es decir, cuestionarios que se generan por procesos computacionales a partir de un tema en específico. Se sabe que existen herramientas que utilizan bancos de preguntas y respuestas para generar cuestionarios, pero dependen de un proceso previo en el que dicho banco se genera por expertos humanos.

Este tipo de cuestionarios generalmente se aplican en el contexto de educación en línea, siendo los mayores consumidores los profesores, debido a su necesidad de generar cuestionarios para evaluar a sus alumnos. Una de sus limitantes, como se mencionó, es que requieren a un experto que genere dichos bancos, siendo el profesor en quien recae esta actividad. La desventaja es que por cada tema nuevo a evaluar se requiere un nuevo banco.

Esta tesis describe un método de solución utilizado para generar cuestionarios a partir del nombre del tema de interés. La información referente al tema se extrae de Wikipedia, la cual se procesa para generar el cuestionario que incluye respuestas de opción múltiple. La manera de interactuar con este cuestionario es a través de una interfaz chatbot que se realizó con el sistema IBM Watson.

Por último, se generó un cuestionario de prueba con información del área de Ciencias de la Computación para verificar el correcto funcionamiento del método y la generación del cuestionario el cual fue evaluado por 3 expertos del área, midiendo su concordancia a través del coeficiente Kappa de Fleiss.

Abstract

This thesis addresses the problem of the lack of tools for the creation of questionnaires automatically, that is, questionnaires that are generated by computational processes from a specific topic. It is known that there are tools that use question and answer banks to generate questionnaires, but they depend on a previous process in which said bank is generated by human experts.

These types of questionnaires are generally applied in the context of online education, with teachers being the biggest consumers, due to their need to generate questionnaires to evaluate their students. One of their limitations, as mentioned, is that they require an expert to generate these banks, the teacher being the one in whom this activity falls. The disadvantage is that a new bank is required for each new subject to be evaluated.

This thesis describes a solution method used to generate questionnaires from the name of the topic of interest. The information regarding the topic is extracted from Wikipedia, which is processed to generate the questionnaire that includes multiple-choice answers. The way to interact with this questionnaire is through a chatbot interface that was made with the IBM Watson system.

Finally, a test questionnaire was generated with information from the Computer Science area to verify the correct functioning of the method and the generation of the questionnaire, which was evaluated by 3 experts in the area, measuring their agreement through the Fleiss Kappa coefficient.

Índice

Índice de Figuras

Índice de Tablas

Capítulo I
Introducción

Los avances tecnológicos en materia de interacción humano-computadora que se tienen hoy en día, hacen suponer que ya existe la manera en que una computadora pueda realizar preguntas de manera automática sobre un tema en particular y evaluar la respuesta proporcionada, lo que sin lugar a dudas sería de gran interés en el ámbito educativo. Sin embargo, en la actualidad las herramientas para la generación de preguntas y respuestas generadas automáticamente por la computadora no han logrado alcanzar esta funcionalidad, la generación automática de cuestionarios no ha sido atendida ni se ha logrado experimentar a profundidad, dejando así un amplio campo de investigación. El trabajo de Binda (2006) muestra que el problema de estas herramientas radica en que la mayoría depende de un usuario que proporcione las preguntas y respuestas.

Por otro lado, Blanco, Vélez, Marcela, Tobón, & Jairo (2015) establecen el desarrollo de un proceso común que consiste en automatizar las evaluaciones a partir de bancos de preguntas y respuestas, en particular atendiendo la selección de preguntas para generar diferentes evaluaciones, pero siempre dependiendo del banco pregunta-respuesta previamente generada de manera manual.

Considerando una solución a esta problemática, las interfaces conversacionales inteligentes, también conocidas como asistentes conversacionales o chatbots, podrían ser un importante elemento a considerar ya que proponen una arquitectura que facilita la interacción en la modalidad de pregunta-respuesta. Actualmente se enfocan en su mayoría al área de compra-venta ofreciendo respuestas a preguntas de clientes u ofreciendo información que guía en la compra de servicios y productos. Sin embargo, en el ámbito educativo usualmente se aplican para responder a dudas de los usuarios de diferentes departamentos que conforman a las instituciones, como los existentes en servicios escolares, financieros entre otros, es decir, son solo informativos y no orientados hacia la enseñanza.

Lo que se pretende en este trabajo es desarrollar un método para generar preguntas y respuestas de manera automática y hacer las evaluaciones mediante cuestionarios y así aumentar el interés en la generación de instrumentos educativos autónomos con la finalidad de enriquecer las herramientas utilizadas como en plataformas e-learning o cursos MOOC.

1.1 Planteamiento del problema

La evaluación de los alumnos en clases presenciales es un fenómeno complejo que condiciona todo el proceso de enseñanza/aprendizaje. Frecuentemente se utilizan cuestionarios para realizar una evaluación, los cuales tienen una amplia variedad de versiones, como pueden ser de preguntas abiertas, con preguntas de opción múltiple, de enunciados por completar, de selección múltiple, entre otros.

Los cuestionarios de evaluación que son aplicados en clases presenciales, son diseñados y redactados por el profesor que desea implementarlos, lo cual provoca un arduo trabajo, además de llegar a complicarse cuando se trata de diversos temas a evaluar, pues se genera un cuestionario independiente a cada tema.

Por otro lado, para las clases en línea existen herramientas que permiten generar evaluaciones, algunas de las cuales son mediante cuestionarios. Pero, para ello se debe de crear manualmente un banco de preguntas y respuestas suficientemente amplio que permita generar diferentes cuestionarios. Sin embargo, aún se depende de que el profesor sea quien redacte y diseñe las preguntas y respuestas lo que conlleva a tener una fuente de información amplia sobre el tema además de la inversión de tiempo para su elaboración. Lo que esta investigación propone es una herramienta que genere cuestionarios de preguntas de opción múltiple la cual recibirá únicamente como entrada el tema del que se desea generar el cuestionario. Esta herramienta puede ser un complemento en la elaboración manual de cuestionarios e incluso llegar a utilizarse para la preparación de los estudiantes quienes desean conocer su dominio en temas específicos.

1.2 Objetivo principal

Generar preguntas de opción múltiple de manera automática a partir de información estructurada en HTML y en texto plano extraído de Wikipedia para ser implementadas a través de un chatbot.

1.3 Objetivos específicos

1) Diseñar un método para la generación de preguntas de opción múltiple a través de información en formato HTML y texto plano.
2) Diseñar un chatbot para la implementación de las preguntas de opción múltiple.
3) Identificar y aplicar una métrica para evaluar las preguntas de opción múltiple generadas automáticamente.

1.4 Alcances

1) No se dependerá de ningún banco de pregunta-respuesta predefinido.
2) El método para la generación de preguntas de opción múltiple será completamente automático.
3) El chatbot proporcionará un instructivo de llenado del cuestionario, siendo solo un asistente y no afectando la decisión de selección de respuesta correcta.

1.5 Limitaciones

1) No se verificará la veracidad de la información utilizada para generar las preguntas de opción múltiple.
2) No se evaluará las respuestas dadas por el usuario.

Capítulo II
Marco conceptual

En esta sección se mostrarán todos los conceptos que son utilizados en este trabajo y es importante definir para comprender su relación e impacto con la investigación desarrollada.

2.1 Respuestas De Opción Múltiple

El trabajo de Binda (2006) hace énfasis en que una respuesta de opción múltiple (POM) consta de tres secciones que son: el tallo, los distractores y opciones. Dentro del tallo se encuentran las sentencias o preguntas, estos tallos tienen distractores que son n respuestas incorrectas que se parecen a la correcta y las opciones es el conjunto de distractores más aparte la respuesta correcta. Es decir, una pregunta puede tener 3 distractores como opción de respuesta además de la correcta.

De igual modo hace mención a los diferentes formatos para la confección de las POM, es decir, existen las convencionales, emparejamiento, verdadero-falso, opción alternativa, verdadero-falso múltiple, compleja, conjunto de preguntas (mini prueba). Sin embargo, las POM convencionales son las que más fácilmente se podrían automatizar debido a que el formato más aceptado puede ser un enunciado parcial seguido de las opciones o puede ser presentado como una pregunta, con las opciones presentadas como respuestas. Se prefiere el formato en forma de pregunta y una variante interesante -la de la mejor respuesta-, en la cual todas las respuestas son correctas, pero una es claramente verificable: “la mejor respuesta correcta”.

2.2 Chatbot

También conocido como asistente conversacional, es un programa que pretende simular una conversación escrita o por voz, con la intención de por lo menos, temporalmente, hacerle creer a un humano que está hablando con otra persona.

Por otro lado, en la bibliografía puede encontrarse un uso indistinto de los términos “asistente virtual” y “chatbot”, sin embargo, se trata de términos que designan tecnología diferente. Como asistente virtual puede entenderse a un conjunto de programas informáticos capaces de

interactuar con los seres humanos mediante el lenguaje natural, en lugar de una interfaz gráfica/GUI como Windows o una línea de comando al estilo DOS, así mismo, es capaz de simular una conversación inteligente por medio de texto y/o audio, emulando el diálogo que podría mantener el usuario con una persona real. Parece ser igual que un chatbot o una interfaz de conversación, pero a diferencia de estos, los asistentes pueden realizar tareas más complejas como activar luces en una habitación, reproducir música mediante comando de voz, abrir programas automáticamente, entre otras actividades. (Grondona, Mazza, & Dorfman, 2012)

2.3 Procesamiento del lenguaje natural

El Procesamiento del Lenguaje Natural (PLN) es el campo de conocimiento de la Inteligencia Artificial que se ocupa de la investigar la manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales, como el español, el inglés o el chino. (Moreno, 2017)

El PLN es muy importante en este tipo de herramientas, porque, como menciona Fracchia & Roger (2003) deben usar un conocimiento considerable acerca de la estructura del lenguaje mismo a utilizar, incluyendo qué es una palabra, como las palabras son combinadas para formar sentencias, que es lo que significan, y como su significado contribuye a las sentencias. Además de este conocimiento, es necesario el conocimiento del mundo, para poder entablar un diálogo.

Así mismo, en la lectura del estado del arte se encontró que el procesamiento del lenguaje natural es una disciplina que se encuentra en la intersección de varias ciencias, tales como las Ciencias de la Computación, la Inteligencia Artificial y Psicología Cognitiva. Su idea central es la de darle a las máquinas la capacidad de leer y comprender los idiomas que hablamos los humanos. La investigación del Procesamiento del Lenguaje Natural tiene como objetivo responder a la

pregunta de cómo las personas son capaces de comprender el significado de una oración oral / escrita y cómo las personas entienden lo que sucedió, cuándo y dónde sucedió.

2.4 Evaluación automática

En cuanto se refiere a una evaluación automática mediante las tecnologías de información y comunicación, Barberá (2016) menciona que la tecnología contiene bancos de datos que se relacionan entre ellos y que se pueden ofrecer a los alumnos respuestas y correcciones inmediatas. Las pruebas electrónicas tipo test que incluyen respuestas correctas ejemplifican este tipo de aportación. De igual modo las ventajas de la evaluación automática son tan evidentes como sus limitaciones y ambas realidades se manifiestan en las innumerables aplicaciones de este tipo de evaluación en la red.

2.5 Contexto definitorio

Dentro de la propuesta de solución se requiere de extracción de contextos definitorios los cuales definiremos como aquellos fragmentos de un texto especializado que aportan información útil para entender un término en su contexto real, y que pueden ser puntos de inicio para la elaboración de ontologías, glosarios, diccionarios electrónicos, entre otras importantes aplicaciones.

Los contextos definitorios incluyen un término, una definición y patrones definitorios, como patrones verbales (“se define como”, “constituido por”), o bien elementos estilísticos como la presencia de marcas tipográficas y variaciones en la tipografía textual que ayudan a resaltar la presencia del término o la definición (comillas, cursivas). (Martínez, Martínez, & Aguilar, 2006)

Capítulo III
Estado del arte

En este apartado se muestra los trabajos que tiene mayor similitud con la investigación o sustentos teóricos para la implementación de la solución, los temas relacionados son: instrumentos de evaluación, creación de evaluaciones y generación de preguntas y respuestas. Dichos temas son abordados también con la intensión de agilizar la creación y aplicación de las preguntas, lo que conduce a explorar herramientas de evaluación en línea, plataformas de aprendizaje en línea y chatbots para la asistencia de la educación.

3.1 Una herramienta para introducir las Ciencias de la Computación con la evaluación formativa automática (*A tool for introducing computer science with automatic formative assessment*)

Este artículo menciona que los alumnos de secundaria programan un chatbot para ser inducidos de manera inconsciente a los conceptos de Ciencias de la Computación. Para que los alumnos crearan su chatbot, a un grupo se les dio 15 lecciones en aula y a otro la herramienta Alice para darles a conocer los principios básicos de la herramienta chatbot. Tal experimento resulto en que las lecciones en aula fueron mejores para programar el chatbot. Como extracción para el presente tema de tesis nos ayudó a analizar la herramienta chatbot y ver de alguna forma la implementación de chatbots en la educación, siendo de buen aporte.

El escrito presenta el resultado de dos estudios observacionales que analizan los efectos de Chatbot en el compromiso de los estudiantes. El primer estudio compara el uso de Chatbot con y sin la herramienta de evaluación formativa en el concurso en línea. El segundo estudio compara el uso de Chatbot con y sin evaluación formativa en el aula. Se decidió utilizar estos estudios para analizar el efecto de la evaluación formativa en el compromiso de los estudiantes en Chatbot porque, aunque los contextos de los estudios eran diferentes, los hallazgos eran similares.

En este artículo se documentó la creación de Chatbot, una plataforma de programación de chatbots cuya intención es aumentar la finalización y el compromiso de las tareas de los estudiantes, especialmente en niñas, mientras enseñan conceptos básicos de CS, como una forma de promover el interés hacia carreras relacionadas con CS y como una forma de contribución a la discusión cada vez más importante de cómo presentar a los estudiantes de secundaria los conceptos de CS de una manera atractiva. (Benotti, Martinez, & Schapachnik, 2017)

3.2 Análisis automático de textos en español utilizando NLTK

La información en este artículo se resume en el estudio de la herramienta NTLK para aplicar técnicas del Procesamiento del Lenguaje Natural (PLN) y para constatar las herramientas PLN disponibles en idioma español. Además, que una de las finalidades de este artículo es, analizar automáticamente textos para clasificarlos según su título y descripción de los archivos en PDF. DE igual modo, menciona clasificadores mediante enfoques estadísticos, Naive Bayes y enfoques analíticos. Estos fueron considerados para el análisis del texto de donde se extraerán las preguntas.

El autor consideró un número de herramientas para el Procesamiento del Lenguaje Natural, las cuales fueron evaluadas según las ventajas y desventajas que proporcionarán al desarrollo del proyecto, teniendo en cuenta los objetivos que se deseaban alcanzar. La herramienta utilizada en este caso es la biblioteca NLTK (Natural Language Toolkit) desarrollada en el lenguaje de programación Python.

Por último, lo que se consideró de este trabajo fue que dejó abierto su método para las futuras líneas de trabajo, que consideran la extracción de texto desde archivos con extensión PDF con el fin de obtener una cantidad suficiente de palabras para caracterizar cada una de las instancias a clasificar. Asimismo, Esta herramienta ha permitido un incremento considerable de las aplicaciones prácticas del Procesamiento del Lenguaje Natural durante la última década. Esto ha

creado nuevos puestos de trabajo en grandes compañías y pequeños startups de nueva creación que han decidido tomar la oportunidad que estas nuevas tecnologías brindan. (Hernández, 2016).

3.3 Aportaciones de la tecnología a la e-Evaluación

La información relevante de este artículo se encuentra en que intenta responder las preguntas como: ¿Realiza la tecnología verdaderamente aportaciones especificas en el campo de la evaluación de los aprendizajes o se trata de una mera copia de la evaluación que se realiza en las aulas presenciales?, ahora, si hubiera estas aportaciones ¿facilitan o entorpecen el proceso de enseñanza y aprendizaje o simplemente se trata de un formato distinto y acorde al contexto virtual de educación?

Para responder a estas preguntas, el autor secciona el artículo en varios temas, que son: valoración de la práctica educativa virtual, influencias de la evaluación, concepto multidimensional sobre la evaluación, aportaciones de las TIC, evaluación enciclopédica, evaluación colaborativa, proceso de evaluación, agenda feedback virtual como derecho y como deber.

Dentro del punto de la valoración de la práctica educativa virtual se encuentra una ventaja grande y es la de poder estar accesible en cualquier tiempo, pero esta misma, es su desventaja al obligar al usuario a tener conexión a internet. Sin embargo, tiene muchas otras parecidas a la práctica del aprendizaje presencial, como lo son plan de trabajo, guías de estudio y calendarios que le orientan en esta secuencia temporal de aprendizaje.

Cuando se habla del concepto multidimensional sobre la evaluación hace énfasis en lo importante que es compartir la complejidad del hecho de evaluar. Puede ser entendida como evaluación del aprendizaje, es decir, la evaluación que nos da como resultado -fruto de la aplicación de la función más normativa y social de la evaluación- la conformidad de si los alumnos

son o no capaces delante de la sociedad de saber y de ser competentes en un determinado ámbito. Otro fenómeno que ocurre, es que la evaluación puede servir como aprendizaje, lo cual conlleva a que comúnmente aceptada la idea de que aprender es conectar el conocimiento nuevo al que accedemos por primera vez con el conocimiento que ya poseemos.

Por último, las aportaciones de las TIC las identifica en 3 puntos:

1. La evaluación automática, en el sentido que la tecnología contiene bancos de datos que se relacionan y se pueden ofrecer a los alumnos respuestas y correcciones inmediatas
2. Evaluación de tipo enciclopédico, que hace referencia al cúmulo de contenidos que se manejan de una fuente más compleja o de diferentes fuentes
3. Evaluación colaborativa, sus métodos pudieran ser los debates virtuales, los foros de conversación y los grupos de trabajo

El punto que ayuda es el número uno, el cual consiste en evaluaciones automáticas, que nos dice que su aporte es bueno debido a que las respuestas correctas se dan al instante y sirve como retroalimentación tanto como para el estudiante como para el profesor. El estudiante comprende sus errores, los corrige y estudia, por lo tanto, el profesor identifica las debilidades del alumno para posteriormente una mejor preparación de la clase. Con esta información, valida lo importante que es la evaluación automática mediante el internet. (Barberá, 2016).

3.4 Arquitectura y operatoria de un sistema de corrección de exámenes automatizado, utilizando grafos dirigidos

El artículo nos habla de un sistema que esta subdivido en capas, describiendo la arquitectura diseñada y la operatoria implementada para analizar las respuestas escritas por alumnos en forma de texto redactado en lenguaje natural, a preguntas abiertas del examen. Menciona que, debido a ser respuestas escritas en texto plano primero se debe tener una corrección

de ortografía, posteriormente un análisis de posibles conceptos obtenidos en la respuesta y por último mediante grafico de nodos obtener la profundidad de la respuesta.

Se ha optado por una arquitectura en tres capas horizontales que modelan niveles crecientes de abstracción y que utilizan interfaces de software definidas para aislar los distintos módulos de efectos no deseados al realizar modificaciones. El patrón de diseño utilizado se denomina MVC y es adecuado para realizar la separación de responsabilidades y lograr una mayor flexibilidad a la hora de realizar modificaciones o mantenimiento del sistema.

La arquitectura presentada implementa la utilización de modernas herramientas de bases de datos, como lo es OrientDB, y expone el análisis realizado y las investigaciones que se llevaron a cabo para su adaptación e implementación. La arquitectura en tres capas permite trabajar con las distintas partes del sistema logrando la transferencia de información de manera ordenada y modularizada con vistas a futuras implementaciones y al crecimiento del sistema, ya que con otro tipo de arquitectura sería muy complejo implementar. Este es una breve descripción del método y herramientas utilizadas que fueron tomadas en cuenta para la solución propuesta en esta investigación (Menvielle, y otros, Arquitectura y Operatoria, 2018).

3.5 Asistentes virtuales de clase en la educación universitaria

Para comenzar un poco a hablar del impacto que causan los agentes virtuales en la educación universitaria, se debe ubicar los términos que utiliza, cuando se refiere a agentes virtuales, explica que son un conjunto de programas informáticos capaces de interactuar con los seres humanos en su propio lenguaje, además, esta investigación hace énfasis en los asistentes virtuales de clase, que son un agente virtual especializado en uno o más campos del conocimiento, con rutinas específicas para realizar tutorías, administrar exámenes, entre otras actividades.

Hablando del experimento realizado, especifica que se utilizó el asistente virtual “Ariel”, el cual está compuesto por un “cerebro artificial” capaz de comprender el lenguaje español, por un módulo específico para las tareas propias de un asistente de clases como podría ser explicación de conceptos, revisión, exámenes, entre otros. Su estructura de conocimiento son antologías con los contenidos relacionados a TICs. El cerebro artificial utilizado es el BG200K de BotGenes.

Para poner a prueba el asistente virtual “Ariel” se invitaron a alumnos de Administración de Recursos Informáticos para utilizarlo como elemento adicional en la preparación del examen parcial para complementar las clases presenciales. Además, incluye una evaluación de tipo Verdadero o Falso para cada tema.

Las características muestrales fueron:

- 36 alumnos.
- Varones 75%.
- Edad promedio de 26 años.
- 8% son graduados de otras carreras.
- 97% de los casos no habían tenido experiencia previa de interacción con un Agente Virtual.
- El 58% pertenece a la carrera de Administración y/o contador, el 42% restante corresponde a Sistemas de Información.

Dentro de los resultados se puede observar que en promedio los alumnos utilizaron el asistente virtual 4.4 horas, habiendo 3 casos donde el uso superó las 8 horas, y representando un 20.6% del total del tiempo dedicado a la preparación de la materia. Dentro de este rango de uso la percepción de eficacia y eficiencia, en una escala de 1 a 5, siendo 1 la peor y 5 la mejor, los alumnos calificaron con un promedio de 4.11 a la utilidad del asistente virtual para comprender

mejor los conceptos ya aprendidos (eficacia), mientras que obtuvo un 3.94 promedio la facilidad de estudiar con mayor rapidez gracias a la disponibilidad del asistente (eficiencia).

Por último, la conclusión obtenida a partir de los datos de correlación entre el uso del asistente virtual y la expectativa sobre la calificación a obtener pareciera indicar una influencia positiva en la seguridad con que el alumno encara el examen. Por lo tanto, se da como buen aporte las pruebas realizadas con este asistente virtual, el cual deja en claro que tienen potencial en el proceso de enseñanza-aprendizaje, que es el área el cual se desea implementar esta propuesta. (Grondona, Mazza, & Dorfman, 2012).

3.6 Auto-evaluación a través de Internet: variables metacognitivas y rendimiento académico

El objetivo principal de este artículo es comprobar en qué mediada la auto-evaluación interactiva mejora el rendimiento académico y la percepción de aprendizaje de los estudiantes, así como también, verificar si los sistemas de auto-evaluación benefician a los estudiantes con escasos niveles de motivación.

Parte del artículo es explicar que es la metacognición y como interfiere en la auto-evaluación. La metacognición se refiere al conocimiento que tiene una persona acerca de su propia cognición y del control que ejerce sobre la misma, además, menciona que estas van de la mano ya que el objetivo de la auto-evaluacón es que el alumno aprenda a controlar su proceso de aprendizaje, incrementando asimismo su anatomía y motivación intrínseca.

La selección de este artículo para el estado del arte, radica en que, dentro de sus materiales y métodos donde elaboraron una amplia gama de tareas de auto-evaluaciones interactivas, se pueden encontrar ejercicio de diversos tipos, como son los de opción múltiple, respuestas cortas, frases incompletas y crucigramas.

Los alumnos que participaron fueron de la asignatura de Análisis de datos I de la Licenciatura de Psicología de la Universidad del País Vasco. Además, se utilizó la metodología de investigación no experimental, agregando que para el uso del recurso de auto-evaluación se midieron tres variables: frecuencia de entrada al sistema, tiempo total de dedicación (en minutos) y puntuación media. Por último, se evaluó la motivación hacia la materia de estudio teniendo en cuenta el número de trabajos voluntarios realizados a lo largo del curso.

Como resultado se obtiene que la satisfacción de los estudiantes con la herramienta de auto-evaluación fue de 46% respecto al total de estudiantes matriculados, esto quiere decir que es bastante elevada tomando en cuenta que fue una actividad voluntaria, por lo tanto, los alumnos están dispuestos a usar este tipo de herramientas para mejorar sus calificaciones. También fue demostrado que la frecuencia de utilización de las pruebas de auto-evaluación se relacionaba con el rendimiento académico, debido a que los resultados indican que la frecuencia de la utilización de los ejercicios de auto-evaluacion correlaciona positivamente con el rendimiento académico. (Erostarbe & Albonigamayor, 2007).

3.7 Chatbots y Agentes conversacionales: un análisis bibliométrico (*Chatbots and Conversational Agents: A Bibliometric Analysis*)

En este artículo trata de hacer énfasis en cómo ha ido aumentando la investigación en el área de chatbots conocidos también como agentes conversacionales, realmente su principal función es dar a conocer cuanta información se encuentra publicada en relación a estos temas, para empezar a hablar un poco de esto, se dice, el análisis bibliométrico es fundamenta y un buen método para explorar los patrones y las tendencias futuras de un tema de investigación, menciona que varios investigadores usaron este método para explorar las tendencias de investigación en

diferentes áreas, como son: le innovación en TI, gestión de proyectos e inteligencia empresarial y análisis.

Este estudio utilizó tres herramientas para el análisis bibliométrico. Son: (1) Base de datos de literatura – Web of Science (WoS) y ProQuest: para recopilar información de publicaciones y citas informe, (2) CiteSpace: para analizar y agrupar datos y (3) Bibliometrix: para determinar patrones de concurrencia.

Base de datos de literatura, en esta herramienta se utiliza la web de ciencia para recopilar datos de publicación y citas de base de datos de literatura académica de alta calidad, cuando se utiliza ProQuest es para recopilar datos en publicaciones como: periódicos, blogs, sitios web entre otros. Para esta búsqueda se utilizaron dos términos "chatbot" y "Agente conversacional", además, debido a que el chatbot y los agentes se comunican mediante chat con los usuarios, se agrega el término "Procesamiento natural del lenguaje", asimismo debido a la relación con la Inteligencia Artificial se suma el término "aprendizaje profundo".

CiteSpace: fue desarrollado en 2004 y el software ha sido actualizado constantemente. CiteSpace puede visualizar las redes de co-citación y agrupar los artículos relacionados par aun área de investigación específica y así los investigadores entiendan con mayor eficiencia el desarrollo de esa área, en el estudio de este artículo los datos a analizar se exportaron de WoS a CiteSpace, ya que se pretende investigar la tendencia de los chatbots en el mundo académico.

Bibliometrix: para evitar sesgos de una sola herramienta de medición se utilizó bibliometrix para generar coocurrencias de palabras clave. En contraste con CiteSpace, bibliometrix es una nueva herramienta para bibliometría análisis. Es un paquete desarrollado en base a la R software y puede construir matrices de datos para la co-citación, acoplamiento, análisis de colaboración científica y co-word análisis. En esta herramienta se combinó los términos "Chatbot" y "agente de conversación" como un conjunto de datos únicos para analizar.

En la primera herramienta, se observa que el resultado de todas las palabras clave buscadas muestran un aumento tendencia en los últimos años. En WoS, se han encontrado 583 resultados sobre “chatbot” y “agente de conversación”, mientras que, en ProQuest, se encuentra en 4246. El autor solo se enfoca en los resultados del WoS, por lo tanto, los resultados muestran que hay una caída después de 2010 y un aumento después de 2015 tanto en publicaciones como en citas. Sin embargo, para el término de “Procesamiento natural del lenguaje” es mayor, esto se debe a que los chatbots es solo una de las aplicaciones de NLP y este tiene mayor amplitud de implementación.

3.8 Consideraciones sobre el Examen de Preguntas de Opción Múltiple (*Multiple choice*)

En este trabajo se pretende dar solución a la problemática con un cuestionario de opción múltiple. De igual forma menciona que las Preguntas de Opción Múltiple (POM) deben comenzar identificando la información o las habilidades más importantes que se desean evaluar, cuidando de mantener una estrecha correlación entre los objetivos del aprendizaje y el contenido del examen.

Fundamente que para crear un examen de opción múltiple debe incluir tres niveles

a) Reconocimiento de la información específica.
b) Comprensión y aplicación del conocimiento.
c) Resolución de problemas (aplicación del conocimiento y habilidades en la resolución de problemas).

Otro aspecto rescatado de este artículo es la guía para la redacción de los exámenes de preguntas de opción múltiple, donde menciona los siguientes términos utilizados para cada componente del examen:

a) Ítem: test construido por un “tallo” y varias respuestas.

b) Tallo: puede ser una sentencia o una pregunta.

c) Opciones: respuestas.

d) Distractores: respuestas incorrectas.

Así como también los tipos de formatos para la confección de POM, dejando los más conocidos en esta lista:

1) POM convencionales.

2) Emparejamiento.

3) Verdadero – falso.

4) Opción alternativa.

5) Verdadero – falso múltiple.

6) POM compleja.

7) Conjunto de preguntas (mini prueba).

Con esta información, se decide que la más viable para la automatización es la convencional, debido a que es el formato mayormente aceptado por los usuarios, su redacción es de la siguiente forma, puede ser un enunciado parcial seguido de las opciones o puede ser presentado como una pregunta, con las opciones presentadas como respuestas. (Binda, 2006).

3.9 Interfaces conversacionales: avances y desafíos (*Conversational interfaces: advances and challenges*)

En este documento, se discute sobre los temas de investigación involucrados en el desarrollo de interfaces conversacionales, describe el trabajo reciente realizado en esta área en el Laboratorio de Ciencias de la Computación del MIT y menciona algunos de los desafíos de investigación no satisfechos, incluida la necesidad de trabajar en dominios reales, generación de lenguaje hablado y portabilidad entre dominios e idiomas.

Este artículo nos habla de una metodología cuantitativa, primero mostrando las interfaces conversacionales más recientes, posteriormente mostrando los avances que se tienen en el MIT para analizar los resultados y compararlos con los trabajos a futuro y la aplicación en dominios reales.

En este documento, se trató de resumir algunos de los desafíos de investigación importantes que deben abordarse antes de que las tecnologías del lenguaje hablado puedan ser utilizadas de manera productiva. En otras palabras, este artículo dio a conocer el avance de los chatbots y las herramientas que pueden ser útiles en la metodología de solución propuesta para esta investigación. (Zue & Glass, 2000)

3.10 Detección de conceptos y relaciones para evaluación de respuestas

Aquí se muestra un método para analizar las respuestas escrita por alumnos en forma de texto redactado en lenguaje natural, a preguntas de un examen, con el fin de contrastar su grado de coincidencia con alternativas de respuestas suministrada por un docente. Como resultados se presentó un método alternativo que permitirá calificar exámenes formados por preguntas que se responderán como ensayo escrito en forma de texto libre, por alumnos de nivel universitario. Se concluye que, se continuará con el análisis de textos desde esta perspectiva ampliando el aprendizaje sobre la materia y otras técnicas y metodologías existentes.

El estudio cuantitativo de este proyecto muestra los resultados de la representación de conceptos mediante grafos, redes semánticas, modelado del conocimiento y mecanismos de búsqueda.

Finalmente, en este artículo se ha presentado un método alternativo que permitirá calificar exámenes formados por preguntas que se responderán como ensayo escrito en forma de texto libre, por alumnos de nivel universitario. La metodología se apoya en la teoría de grafos para construir

el dominio del trabajo, que se corresponde con el área de conocimiento de la cátedra Paradigmas de Programación dictada en la Universidad Tecnológica Nacional Facultad Regional Córdoba, y en el estudio de las gramáticas y lenguajes formales. Se consideró este método debido a que a pesar de tener una propuesta de solución se espera escuchar demás propuestas. (Menvielle, Groppo, Marciszack, & Analía Guzmán, 2016)

3.11 Diferentes métricas de medición para evaluar un Sistema de chatbot (*Different measurements metrics to evaluate a chatbot system*)

El artículo hace énfasis en el concurso anual llamado Premio Loebner, el cual usa distintas métricas para la evaluación del chatbot, específicamente en la habilidad de poder engañar al jurado evaluador en una sesión de chat restringido. Así como también, describe métodos para entrenar y adaptar un chatbot al idioma de un usuario en específico a través de un cuerpo de entrenamiento específico.

Si bien ya es conocido el "juego de imitación" de Alan Turing, el Premio Loebner se basó en este, por lo tanto, una de las evaluaciones consistía en que se utilizaban 10 agentes conversacionales de los cuales 6 eran programas de computadoras, 10 jueces tendrían conversaciones con ellos durante quince minutos por último cada juez tendría que dar un orden en el cual se definía desde el menos aparente a un humano hasta el más humano, esto establecido mediante la conversación que obtuvieron.

En estas evaluaciones se utiliza chatbots diseñados con la arquitectura ALICE/AIML la cual consta de datos llamados objetos AIML, que se componen de unidades denominadas temas y categorías. El tema es un elemento de nivel superior, opcionalmente tiene un atributo de nombre y un conjunto de categorías relacionados con ese tema, Cada categoría es una regla para hacer

coincidir una entrada y convertirla en una salida, y consiste en un patrón, que coincide con la entrada del usuario.

Se utilizaron diferentes métricas, la primera es la eficiencia de diálogo, la cual se midió en términos de coincidencia atómica, coincidencia de la primera palabra, coincidencia más significativa y no coincidencia. La métrica que evalúa la calidad del diálogo, consistió en dar la conversación entre el chatbot y el usuario a una persona experta en el idioma de conversación y evaluar si las respuestas contienen "razonabilidad" con las categorías de respuesta razonable, respuesta extraña pero comprensible o sin sentido. Cabe mencionar que, de las dos métricas mencionadas anteriormente, la de evaluación del diálogo puede ser muy subjetiva debido a que cada evaluador experto puede dar diferentes resultados en una sola prueba.

Debido al tipo de métricas que aplica en las diferentes evaluaciones queda claro que evaluar los chatbots solo con el conocimiento de un experto no es suficiente y las métricas deben tener mayor fundamento.

3.12 Evaluación de la calidad de las preguntas de selección múltiple utilizadas en los exámenes de Certificación y Recertificación en Cardiología en el año 2009

En este escrito se muestra como fueron evaluados los exámenes de certificación en cardiología, con preguntas de tipo opción múltiple, por lo tanto, es útil para verificar si el método de solución es realmente funcional o cuales serían los aspectos que le faltan.

La forma en la cual realizó la evaluación fue con 2 observadores independientes (un médico y una pedagoga) revisaron dos exámenes (A y B) de 100 preguntas cada uno, implementados durante 2009 en la SAC en el marco de los procesos de Certificación y Recertificación de Especialistas. La evaluación se realizó utilizando el Índice de Calidad de

Galofré, que es una sistematización de las recomendaciones publicadas sobre la construcción de preguntas de selección múltiple. Los 10 criterios que se tomaron en cuenta son:

1. Presencia de viñeta
2. Enunciado completo
3. Evita uso de negaciones
4. Concordancia gramatical entre enunciado y opciones
5. Distractores verosímiles
6. Extensión similar entre las opciones
7. Evitar las opciones “ninguna” y “todas las anteriores”
8. Opciones ordenadas
9. Opciones homogéneas
10. Nivel de aplicación de conocimiento o superior

Los resultados muestran que: de las 200 preguntas evaluadas, el 30% tiene muy buena calidad técnica (puntajes 4 y 5); cerca del 40% son preguntas aceptables que se deberían mejorar (puntaje 3) y el 30% son inaceptables (puntajes 1 y 2). El Índice Calidad del examen A fue de 2,15 y el del examen B resultó de 3,31. Galofré informa que aplicó el Índice de Calidad a distintos exámenes, de ciencias básicas y clínicas, utilizados en Chile y en otros países y encontró que el Índice de Calidad de esos exámenes variaba entre 1,6 y 4,6. Los exámenes de Cardiología analizados se ubican dentro del amplio rango mencionado por Galofré y muy por debajo del índice de calidad (4, 8) de un examen de selección de residentes implementado en un hospital de comunidad.

En otras palabras, la evaluación realizada permitió identificar un porcentaje elevado de preguntas deficientes; si bien uno de los exámenes presentó una proporción mayor de preguntas de calidad aceptable, se advirtió que el margen de oportunidad para el mejoramiento era amplio

en ambas pruebas. De esta forma se pretende evaluar el método y esperar unos mejores resultados. (Galli, y otros, 2011)

3.13 Extracción automática de contextos definitorios

Para el módulo b) del método de solución es necesario la extracción de contextos definitorios, de modo que se investigó en artículos relacionados su método de extracción o si existe algo similar. El reciente avance en el desarrollo de nuevas tecnologías para el trabajo terminológico ha aportado diversas herramientas para tratar de resolver este problema. Una de estas herramientas son los corpus de textos especializados en los cuales se pueden extraer automáticamente términos y definiciones. Los contextos definitorios incluyen un término, una definición y patrones definitorios, como patrones verbales (“se define como”, “constituido por”), o bien elementos estilísticos como la presencia de marcas tipográficas y variaciones en la tipografía textual que ayudan a resaltar la presencia del término o la definición (comillas, cursivas).

Este artículo es solo informativo, por lo cual hace mención a los trabajos realizados por el Grupo de Ingeniería Lingüística. Estos trabajos nos dicen que existen varios enfoques metodológicos para la extracción conceptual en textos especializados, uno de los proyectos propone desarrollar un sistema completo y coherente de estructura modular, basado en información lingüística, que sea aplicable a diversos corpus textuales especializados en lengua española con el fin de extraer automáticamente términos y definiciones. Igualmente, el proyecto tiene la finalidad de conformar un Corpus de Contextos Definitorios, esto es, un repositorio electrónico para los términos, definiciones y aquellos patrones definitorios que suelen coocurrir en los contextos definitorios. (Martínez, Martínez, & Aguilar, 2006)

3.14 Índice de calidad para evaluar preguntas de opción múltiple

Este documento habla sobre la construcción de un índice de calidad que se basan en algunos ya existentes como el trabajo de Josefowicz. Para este caso se utilizaron factores para evaluar la calidad técnica de cada pregunta y estos son: presencia de viñeta, enunciado completo, evitar uso de negaciones, concordancia gramatical entre enunciado y opciones, distractores verosímiles, extensión similar entre las opciones, evitar ninguna y todas las anteriores, opciones ordenadas, opciones homogéneas y aplicación de conocimiento o superior.

Los factores mencionados se aplican a cada pregunta de opción múltiple y el resultado expresa si cuenta con la calidad de un buen cuestionario. También se hace mención que este tipo de índice de calidad sirve como feedback en algunos talleres de elaboración de preguntas siendo un eficaz mecanismo para comunicar los errores de la pregunta y enfatizar cuáles son las características que necesariamente debe tener una buena pregunta, lo que no impide mencionar, además, las otras características señaladas en la literatura.

De igual modo si se desea generar programas informáticos que genere preguntas de opción múltiple, al momento del modelado de las preguntas este índice de calidad puede servir para su correcta estructura.

3.15 Necesidades de aprendizaje sobre la elaboración de instrumentos evaluativos escritos

La investigación realizada en este artículo se implementó en un campo de estudios de medicina, específicamente con docentes del área de enfermería con la finalidad de generar conocimientos sobre la elaboración de instrumentos evaluativos escritos.

Para la metodología que se utilizó se necesitó de 20 profesores con categoría de docentes que tienen la responsabilidad de generar instrumentos de evaluación que se aplican a nivel nacional

a estos se les aplico un cuestionario con la finalidad de indagar sobre conocimientos en la elaboración de instrumentos de evaluación escritos.

Los resultados obtenidos en su evaluación fueron los siguientes: 55% de los profesores fueron evaluados como bien en las preguntas de tipo verdadero o falso, 50% fueron evaluados como regular en la construcción de preguntas para relacionar columnas, 65% se considera de mayor complejidad en las preguntas de opción múltiple y el 55% considera que son las de tipo ensayo o desarrollo.

La conclusión de los autores de esta investigación refiere que las preguntas de verdadero o falso, son las más sencillas de elaborar, pero se hace énfasis en que no se puede evitar el análisis de una respuesta a partir del azar.

3.16 Preparación de preguntas de opciones múltiples para medir el aprendizaje de los estudiantes

Este artículo menciona la importancia de las POM y algunas ventajas que se tienen frente a la forma de evaluar con ensayos. Pero, para empezar a evaluar hay que saber sobre los tipos de aprendizaje, todos los contenidos de aprendizaje pueden ser clasificados en tres categorías esenciales: conocimientos, destrezas y habilidades. Los autores opinan que las POM tienen un papel importante en medir los conocimientos, un papel limitado en medir destrezas y un papel creciente o naciente en medir algunos aspectos de las habilidades.

Además, sustenta que desde hace ya algunos años se está debatiendo sobre con que herramienta de evaluación es mejor, si con ensayos o POM, a lo que el artículo refiere es que los psicólogos y pedagogos -incluyendo a expertos en los exámenes- han discutido las ventajas y desventajas de los formatos de las POM y del ensayo para medir los conocimientos. A principios del siglo XX, el ensayo era el formato dominante de preguntas, pero con la aparición del Stanford

Achievement Test en los Estados Unidos en 1923, el formato de las POM parecía establecerse sólidamente como el formato de preferencia en los exámenes para medir el aprendizaje. Los argumentos formulados en aquel entonces siguen favoreciendo los exámenes de las POM hoy en día.

Para aplicar las POM hay diferentes formatos, a lo que el artículo reseñó siete formas de las POM que han sido utilizados para medir el aprendizaje de los estudiantes. De estos siete formatos, uno no es recomendado. Tres formatos son los que más típicamente se usan para evaluar el aprendizaje: el de las POM convencionales, el del emparejamiento y el de verdadero-falso. Los otros tres formatos de las POM probablemente le serán menos familiares, pero en realidad son bastante buenos para medir conocimientos, destrezas y algunos tipos de pensamiento de alto nivel.

Por la sugerencia de este y demás autores, se decide investigar un poco más en las POM convencionales donde menciona que este formato es el más popular y más aceptado generalmente. Es interesante notar que tenemos muy poca investigación sobre su valor. Sabemos que las POM convencionales pueden variar en formas básicas. La POM puede ser un enunciado parcial seguido de las opciones o puede ser presentada como una pregunta, con las opciones presentadas. Preferimos el formato con forma de pregunta. Una variación interesante es el formato de la mejor respuesta en el cual todas las opciones son correctas, pero hay una que es clara y verificablemente la mejor respuesta correcta. (Haladyna, Haladyna, & Soto, 2002).

3.17 Sistema de evaluaciones en línea como herramienta para los niveles de educación media superior

El presente artículo trata sobre la información recabada y el software desarrollado como resultado de las principales necesidades que debe de cubrir una herramienta informática para la elaboración de exámenes a nivel bachillerato en México. Como resultado se obtuvo el sistema

Evalua-t que es una herramienta fácil de utilizar por los profesores del nivel medio superior para elaborar exámenes y evaluar los conocimientos que los alumnos adquieren en el aula.

Esta investigación propone como método de solución el desarrollo de un Sistema Web denominado “Evalua-t” para la elaboración de exámenes y evaluación a alumnos del nivel medio superior. La evaluación podrá tener las características de un tipo estandarizado que garantice la fiabilidad y validez de su medición, generando como consecuencia información transparente sobre los resultados obtenidos.

Este sistema no busca que los profesores cambien su forma de evaluar, sino que principalmente sean ellos quienes elaboren los exámenes para sus alumnos. Utilizar un sistema como este ahorra tiempo al profesor, quien puede dedicarlo a calificar los exámenes y a mejorar el conocimiento del estudiante gracias a los resultados que este sistema da de manera inmediata. Este artículo se tomó como referencia de que hay herramientas de automatización de evaluaciones, pero que sin embargo sigue dependiendo de un profesor para ser operado. (Centeno Brambila & Lira Obando, 2015)

3.18 Tendencias en la evaluación del aprendizaje en cursos en línea masivos y abiertos

Este articulo tiene doble propósito, por una parte, el análisis de los procesos de evaluación que se desarrollan en 87 cursos de diferentes plataformas de MOOC (internacionales, de universidades españolas y otras) y, por otra, la reflexión sobre esos procesos basada en el estudio de MOOC desarrollados en la Universidad de Granada (España). Como resultado de la investigación se presentaron varios puntos entre ellos, aumentar la variedad de herramientas de evaluación para una mayor adaptación a diferentes formas de aprendizaje.

El objetivo fundamental de este trabajo es el análisis de los procesos de evaluación que se desarrollan en MOOC y la reflexión sobre esos procesos basada en un estudio de caso llevado a cabo en la Universidad de Granada (España). Se establecen para ello dos fases diferenciadas: una primera fase consistente en un análisis exploratorio de las características de la evaluación en una colección de MOOC y una segunda fase en la que se realiza el estudio de un caso centrado en las percepciones manifestadas por sus participantes.

Este estudio exploratorio analiza la evaluación del aprendizaje en 87 MOOC y la satisfacción ante la evaluación de los participantes en cursos de AbiertaUGR. Como conclusión se obtuvo que la evaluación del aprendizaje de los participantes merece una atención especial debido a su impacto en la construcción del conocimiento. Por lo tanto, se notó que el objetivo principal del artículo es analizar el qué, quién, cuándo y cómo de la evaluación del aprendizaje en los cursos (Arrufat, Sánchez, & Santiuste, 2015)

3.19 Uso de la evaluación electrónica para mejorar el aprendizaje de los estudiantes y la evidencia de los resultados del aprendizaje (*Using e-Assessment to enhance student learning and evidence learning outcomes*)

Este articulo nos muestra que se han estado implementando cambios significativos en los enfoques de la enseñanza y el aprendizaje en la educación superior con la llegada de los MOOC, los enfoques invertidos en el aula, la introducción de espacios de aprendizaje informal, la ramificación del aprendizaje y la expectativa de modos de entrega más flexibles. No hemos visto tanta actividad en el área de evaluación para alinear estos cambios en la enseñanza con la forma en que usamos la evaluación.

Mediante el método cualitativo, compara como es que la tecnología se ha estado invirtiendo en aplicar diferentes métodos para la enseñanza, pero se está olvidando la eficiencia en las evaluaciones.

Este articulo alienta a comprender con mayor detalle cómo se puede utilizar la tecnología para promover evaluaciones auténticas y significativas y no solo hacer que las evaluaciones sean más eficientes, sino que también, cómo implementar la evaluación electrónica junto con otras estrategias. Más tareas abiertas y comentarios detallados para los estudiantes son dos áreas de continua necesidad en la evaluación electrónica. (Crisp, Guárdia, & Hillier, 2016).

3.20 Validez y confiabilidad en la construcción de reactivos utilizados en pruebas de opción múltiple (POM)

En este documento se hace mención a la validez y la confiabilidad de usar reactivos de opción múltiple tanto como en pruebas escritas o digital, de igual modo ver la eficiencia y la forma práctica de evaluar los conocimientos que los alumnos adquieren en el proceso de enseñanza-aprendizaje. Contiene una sección encargada de explicar la validación y ponderación de las preguntas de opción múltiple, donde hace mención a varios objetivos de una validación correcta.

Por otra parte, define como confiabilidad al grado de confianza que se tiene en la información que brinda un instrumento, aludiendo a la estabilidad, equivalencia, permanencia y reiteración de los datos que ofrece el instrumento. La validez de un instrumento

3.21 Análisis por categoría de artículos con mayor aporte

A continuación, se realiza una comparación de los artículos de acuerdo a la información que aportaron, cada tabla consta de diferentes columnas para una mejor comparación en cada sección diferente. Los artículos son divididos en dos secciones: en la sección 3.21.1 se comparan aquellos que tratan el tema de las evaluaciones en línea, en la sección 3.21.2 los relacionados con el tema

de chatbots y finalmente, en la sección 3.21.3, aquellos que estudian el tema de las preguntas de opción múltiple.

3.21.1 Artículos de evaluaciones en línea

Tabla 1. Comparación de evaluaciones en línea.

Nombre	Objetivo	Herramientas de evaluación en línea	Resultados
Aportaciones de la tecnología a la e-Evaluación	Exponer la importancia de la comunicación hacia los alumnos sobre el resultado de sus evaluaciones mediante herramientas tecnológicas	Evaluación automática Evaluación enciclopédica Evaluación colaborativa	Como resultado de la comparación teórica, se obtiene la importancia que aporta la tecnología sobre la retroalimentación de las evaluaciones
Asistentes virtuales de clase en la educación universitaria	Análisis y evaluación a un asistente virtual en específico para medir el impacto sobre el aprendizaje de los alumnos	Asistente virtual de clase "Ariel"	Las estadísticas con las que fueron analizadas los resultados de los alumnos demuestran que el uso de asistentes mejora su desempeño en sus evaluaciones
Auto-evaluación a través de Internet: variables metacognitivas y rendimiento académico	Evaluar si los alumnos se sentían motivados a usar auto-evaluaciones en internet, de ser así, confirmar si obtenían mejores calificaciones por usarlos y por último relacionar un conjunto de variable metacognitivas con el uso de estas herramientas.	Suite Hot potatoes.	Se expresaron en tres apartados, primero detallan la satisfacción del alumno con la herramienta. Segundo, verifican la relación de ejercicios de auto-evaluacion y el rendimiento académico. Tercero, se verifica si las variables metacognitivas se relacionan con el rendimiento académico.
Sistemas de evaluaciones en línea como herramienta para los niveles de educación media superior	Propone el desarrollo de un sistema web que elabore exámenes con pruebas estandarizadas de tipo clase o de uso en el aula para alumnos de nivel medio superior	Sistema Web denominado "Evalua-t" para la elaboración de exámenes y evaluación	Se evaluó el grado de satisfacción al utilizar Evalua-t para la elaboración de exámenes y su aplicación a los alumnos en el nivel medio superior, cubriendo un 95 % las expectativas gracias a la facilidad de la creación de exámenes y a las diferentes herramientas y opciones que ofrece al crear un examen nuevo.
Tendencias en la evaluación del aprendizaje en cursos en línea masivos y abiertos	Analizar el proceso de cursos MOOC de diferentes plataformas con base en un estudio de caso llevado a cabo en la Universidad de Granada (España)	87 cursos de diferentes plataformas de MOOC	En las plataformas analizadas se observa un diseño pedagógico poco colaborativo, centrado en conocimientos y contenidos, frente a un porcentaje bajo centrado en la participación o el aprendizaje colaborativo

3.21.2 Artículos de chatbots

Tabla 2. Comparación de chatbots.

Nombre	Objetivo	Herramientas de trabajo	Idioma
Chatbots and Conversational Agents: A Bibliometric Analysis	Examinar las investigaciones sobre chatbots usando el análisis bibliométrico cuantitativo. La contribución de esta investigación es ayudar a los investigadores a identificar las brechas de investigación para las futuras investigaciones de chatbots	Base de datos de literatura: Web of science, ProQuest, CiteSpace y Bibliometrix	Ingles
Conversational interfaces: advances and challenges	Se analizan los problemas que involucran el desarrollo de dichas interfaces, se describen los trabajos recientes en el MIT y los problemas no resueltos en este mismo, así como también la necesidad de trabajar en dominios reales.	Ninguno útil	Inglés, francés, alemán e italiano
Different measurements metrics to evaluate a chatbot system	Mediante las métricas de eficiencia de diálogo de "caja de cristal" y métricas de calidad de diálogo de "caja negra" y comentarios de satisfacción del usuario, generar un modelo de evaluación que se adapte a la aplicación y a las necesidades del usuario.	Prototipo KGA	Inglés y Afrikaans

3.21.3 Artículos de preguntas de opción múltiple

Tabla 3. Comparación de Preguntas de Opción Múltiple.

Nombre	Objetivo	Herramientas extraídas	Resultados
Consideraciones sobre el Examen de Preguntas de Opciones Múltiples (Multiple choice)	Identifica patrones para la generación de formatos de preguntas de opción múltiple.	Formatos para la confección de preguntas de opción múltiple Guía para la redacción de exámenes de preguntas de opción múltiple	Se obtuvo diferentes formatos para la redacción de preguntas de opción múltiple identificando sus principales ventajas y desventajas.
Índice de calidad para evaluar preguntas de opción múltiple	Generar un índice que indique con una nota o puntaje la calidad relativa en la redacción de las preguntas de opción múltiple	Modelo plantilla de evaluación de preguntas de opción múltiple	Se redacto una plantilla a base de otras encontradas en la literatura con la finalidad de producir un índice que puede ser aplicado en forma independiente por un evaluador, solamente usando el ítem aisladamente en cuanto a su construcción.
Preparación de preguntas de opciones múltiples para medir el aprendizaje de los estudiantes	Comparar la anatomía de las preguntas de opción múltiple (POM) y los formatos alternativos de ensayo, así como también el valor de los exámenes basados en POM al medir los diferentes tipos de aprendizaje.	Anatomía básica de los formatos de las preguntas de opción múltiple	Se concluyó que las POM tienen varios formatos nuevos, interesantes y útiles a disposición y un conjunto de pautas para poder escribirlas eficazmente. Esto con base en la extracción de la anatomía básica de las POM.

Capítulo IV
Método de solución

De acuerdo al problema expuesto, se diseñó el siguiente método de solución (figura 1) que consiste de cinco grandes módulos.

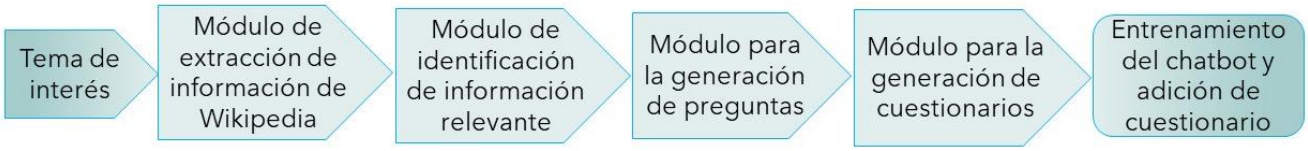

Figura 1. Método de solución.

El módulo de extracción de información de Wikipedia, parte de un tema de interés que es ingresado por el usuario. El módulo realiza la búsqueda de dicho tema en Wikipedia y descarga la información encontrada. A continuación, dicha información ingresa al módulo de identificación de información relevante, que procesa el texto para identificar información de la cual es posible generar preguntas y extraer la respuesta correcta. Esta información se analiza posteriormente por el módulo para la generación de preguntas, para acoplarla a plantillas predefinidas que ajustarán el texto a un formato de pregunta. El siguiente módulo genera respuestas distractoras que son añadidas a las preguntas generadas anteriormente, e incluyen las respuestas correctas para devolver un cuestionario de opción múltiple. Por último, la información generada se le transfiere al chatbot, que se entrenó previamente para fungir como asistente de evaluaciones. En forma sintética los anteriores detalles se pueden observar en la Figura 2.

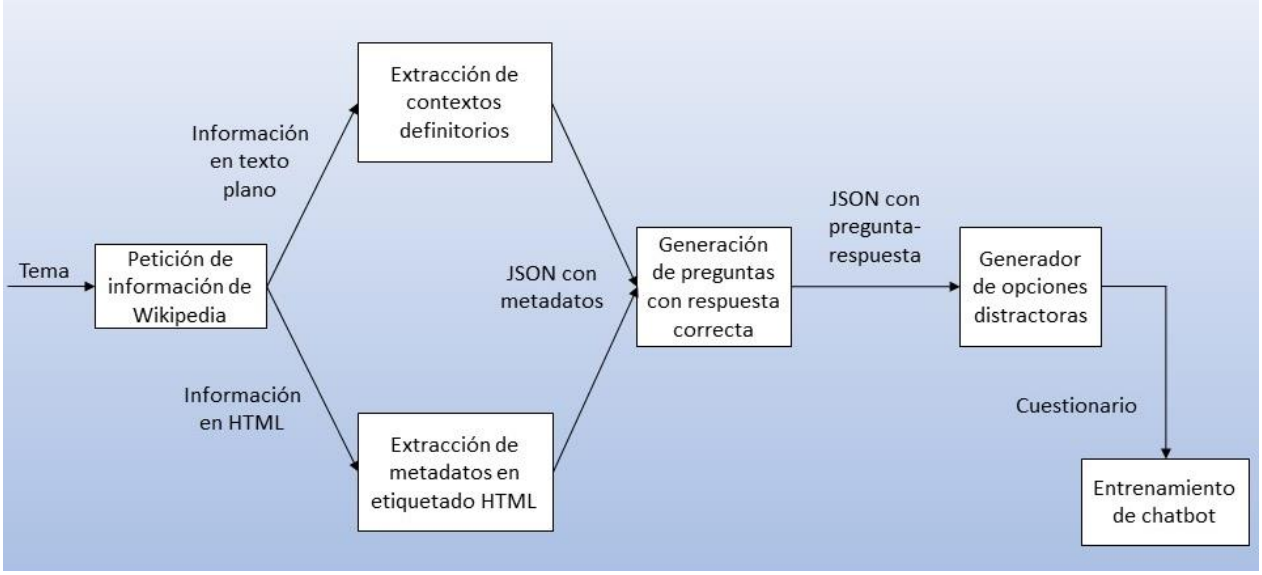

Figura 2. Método de solución extendido.

4.1 Desarrollo del método de solución

Para desarrollar el método de solución es necesario la generación de 5 scripts de programación que a continuación se explicarán cada uno a detalle, estos fueron desarrollados en el lenguaje Python con base en los servicios:

- Spacy: Para el procesamiento del lenguaje natural
- Watson Assistant: Administrador del chatbot y proveedor de la inteligencia artificial
- Flask: Framework web, para el desarrollo de la página web
- Wikipedia: Proveedor de la información
- Beautiful Soup: Para la extracción del etiquetado HTML

4.2 Módulo “Extracción de información de Wikipedia”

Se utilizó Wikipedia como fuente de información relacionada al tema de interés. Para este trabajo, se definieron 10 temas relacionadas al área de computación de los cuales es posible encontrar información en Wikipedia. Estos temas son:

- Lenguaje de programación
- Matemáticas discretas
- Teoría de la computación
- Inteligencia artificial
- Ingeniería de software
- Ciencias de la computación
- Sistemas distribuidos
- Base de datos
- Red neuronal artificial
- Diagramas de flujo

Wikipedia proporciona un API ("Wikipedia Documentation — wikipedia 0.9 documentation", 2019) para extraer información, la cual cuenta con 12 métodos que se listan a continuación:

- Search
- Suggest
- Summary
- Page
- Geosearch
- WikipediaPage
- Lenguages
- Set_lang
- Set_rate_limiting
- Random

- Donate
- Exceptions

La información puede ser extraída en diferentes formatos: html, texto plano, en diferentes idiomas y también imágenes. Los métodos que se utilizaron para desarrollar el módulo de interés, son:

WikipediaPage. Contiene datos de una página de Wikipedia. Utiliza métodos de propiedad para filtrar datos del HTML en bruto. Se usan los métodos *html*, *content* y *url*. *Html* devuelve toda la página en dicho formato, *content* devuelve toda la información de la página en texto plano y *url* el enlace de la página.

Summary. Devuelve un pequeño resumen de la palabra que se envíe, siempre y cuando tenga una página dedicada en Wikipedia. Por ejemplo, si se envía la cadena "lenguaje de programación", se obtiene lo siguiente: *"un lenguaje de programación es un lenguaje formal que proporciona una serie de instrucciones que permiten a un programador escribir secuencias de órdenes y algoritmos a modo de controlar el comportamiento físico y lógico de una computadora con el objetivo de que produzca diversas clases de datos"*

Search. Proporciona títulos de páginas de Wikipedia que tienen relación con la palabra enviada, por ejemplo, si se envía la frase "lenguaje de programación", lo que devuelve son los siguientes temas: lenguaje de programación, c (lenguaje de programación), ada (lenguaje de programación), logo (lenguaje de programación), lenguaje de programación orientado a pila, pascal (lenguaje de programación), lenguaje de programación esotérico, scala (lenguaje de programación), oz (lenguaje de programación), abc (lenguaje de programación). Esta función contiene un parámetro llamado *result*, cuyo propósito es determinar la cantidad de resultados que se quiere obtener, en este caso se solicitan 10, pero se desconoce el límite.

Esta API de Wikipedia es muy útil para generar pruebas, además de que no tiene retardos en la respuesta a las solicitudes. Sin embargo, los resúmenes de los 10 temas se almacenaron en archivos de texto plano (txt) con la intención de hacer pruebas locales.

Para generar el archivo txt con los resúmenes de los 10 temas, se combinaron dos métodos, *search* y *summary*. De esta forma se puede extraer información de Wikipedia de diferentes temas y así poder generar un corpus amplio. La manera en la que se realiza la extracción de información es la siguiente:

Se usa el tema "lenguaje de programación" que es ingresado al método *Search*, que devuelve 10 subtemas relacionados a él: lenguaje de programación, c (lenguaje de programación), ada (lenguaje de programación), logo (lenguaje de programación), lenguaje de programación orientado a pila, pascal (lenguaje de programación), lenguaje de programación esotérico, scala (lenguaje de programación), oz (lenguaje de programación), abc (lenguaje de programación).

De los 10 subtemas de Search, cada uno se ingresa al método *Summary*, esto devuelve un resumen por cada uno. Por ejemplo, al enviar "lenguaje de programación esotérico", *Summary* devuelve: *"un lenguaje de programación esotérico o exótico es un lenguaje de programación minimalista, cuya utilidad para la programación de proyectos de gran tamaño es dudosa normalmente debido a su naturaleza ofuscada u otra característica que no es común en otros lenguajes"*.

El resumen es lo que se almacena como información para generar el archivo txt. Por lo tanto, ingresando el primer tema, el archivo txt se conforma con la siguiente información (tabla 4):

Tabla 4. Resumen del tema Lenguaje de Programación.

Subtemas	Resumen
Lenguaje de programación	un lenguaje de programación es un lenguaje formal que proporciona una serie de instrucciones que permiten a un programador escribir secuencias de órdenes y algoritmos a modo de controlar el comportamiento físico y lógico de una computadora con el objetivo de que produzca diversas clases de datos.
C (lenguaje de programación)	c es un lenguaje de programación originalmente desarrollado por dennis ritchie entre 1969 y 1972 en los laboratorios bell, como evolución del anterior lenguaje b, a su vez basado en bcpl.al igual que b, es un lenguaje orientado a la implementación de sistemas operativos, concretamente unix.
Ada (lenguaje de programación)	ada es un lenguaje de programación orientado a objetos y fuertemente tipado de forma estática que fue diseñado por jean ichbiah de cii honeywell bull por encargo del departamento de defensa de los estados unidos.
Logo (lenguajes de programación)	logo es un lenguaje de programación de alto nivel, en parte funcional, en parte estructurado; de muy fácil aprendizaje, razón por la cual suele ser el lenguaje de programación preferido para trabajar con niños y jóvenes.
Pascal (lenguaje de programación)	pascal es un lenguaje creado por el profesor suizo niklaus wirth entre los años 1968 y 1969 y publicado en 1970.
Lenguaje de programación orientado a pila	un lenguaje de programación orientado a pila es un lenguaje que usa un modelo de máquina de pila para pasar los parámetros.
Lenguaje de programación esotérico	un lenguaje de programación esotérico o exótico es un lenguaje de programación minimalista, cuya utilidad para la programación de proyectos de gran tamaño es dudosa normalmente debido a su naturaleza ofuscada u otra característica que no es común en otros lenguajes.
Scala (lenguajes de programación)	scala es un lenguaje de programación multi-paradigma diseñado para expresar patrones comunes de programación en forma concisa, elegante y con tipos seguros.
Oz (lenguajes de programación)	oz es un lenguaje de programación multi-paradigma y lenguaje de programación esotérico.
Abc (lenguajes de programación)	abc es un lenguaje de programación imperativo de propósito general y entorno de programación desarrollado en el centrum wiskunde & informatica de países bajos por leo geurts, lambert meertens y steven pemberton.

Como resultado se tiene un archivo txt codificado en UTF-8, contiendo un total de 3903 palabras y 25650 caracteres correspondientes a 97 resúmenes.

4.3 Módulo “Extracción de información relevante”

Este módulo tiene la función de identificar aquellos fragmentos de texto que contienen información de la cual es posible realizar preguntas y que, además, mencionan la respuesta. Esto se logra analizando la manera en que se encuentra estructurado el texto.

El módulo se conforma por dos métodos, el primero tiene la función de identificar contextos definitorios y el segundo extraer metadatos del etiquetado HTML. A continuación, se describen ambos métodos.

4.3.1 Método para la identificación de Contextos definitorios

Un contexto definitorio (CD) es un texto cuya estructura, según Hipólito (2018), es {T PD D}[PP], donde T es el término por definir, PD el patrón definitorio, D la definición y, opcionalmente un texto que puede precisar la definición, PP. Esta estructura sirve para hacer el reconocimiento de ocurrencia de un CD en un texto con el fin de extraer el término y la definición. Esto se puede observar con la Figura 3 y el siguiente ejemplo:

“Un **lenguaje de programación** **es un** lenguaje formal que proporciona una serie de instrucciones que permiten a un programador escribir secuencias de órdenes y algoritmos a modo de controlar el comportamiento físico y lógico de una computadora con el objetivo de que produzca diversas clases de datos.”

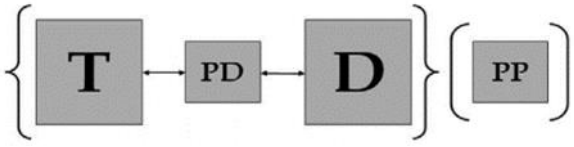

Figura 3. Estructura de un contexto definitorio.

Donde lo subrayado en rojo es T, el azul es PD y lo amarillo es D. También se encuentra una clasificación de CD la cual consiste en: tipográficos, sintácticos, mixtos y complejos.

Existen diversas herramientas y métodos que podrían ser utilizados para identificar los contextos definitorios o para extraer la información de interés que se busca obtener con este tipo de estructuras de texto. En el Anexo se mencionan un par de métodos con los que se experimentó pero que arrojaron resultados no adecuados.

La solución que se implementó en este trabajo, consiste en usar el método *search* de Wikipedia, al cual se le proporciona el tema de interés y devuelve una lista de 10 subtemas, que en este caso servirán como términos, cada uno de los cuales son ingresados al método *summary*, con lo que se obtiene un pequeño resumen de cada término. Dado que ya se conoce el término, se busca identicar el patrón definitorio (PD) y su definición. Para ello, el resumen que devuelve *summary* se procesa por la librería *spacy* ("Spacy Industrial-Strength Natural Language Processing", 2019), la cual convierte el texto en vectores, uno de los vectores que se utiliza son los *tokens*, cuya función es separar las palabras. El texto es convertido a su forma base y se obtiene información respecto a la categoría gramatical de cada palabra (etiquetado POS). A continuación, se proporciona un ejemplo con una sección del tema lenguaje de programación (tabla 5):

Tabla 5. Etiquetado del resumen.

Resumen	Texto
Original	un lenguaje de programación es un lenguaje formal que proporciona una serie de instrucciones que permiten a un programador escribir secuencias de órdenes y algoritmos a modo de controlar el comportamiento físico y lógico de una computadora con el objetivo de que produzca diversas clases de datos.
Etiquetado	un DET uno | lenguaje NOUN lenguaje | de ADP de | programación NOUN programación | es AUX ser | un DET uno | lenguaje NOUN lenguaje | formal ADJ formal | que PRON que | proporciona VERB proporcionar | una DET uno | serie NOUN seriar | de ADP de | instrucciones NOUN instrucción | que PRON que | permiten VERB permitir | a ADP a | un DET uno | programador ADJ

	programador | escribir VERB escribir | secuencias NOUN secuenciar | de ADP de | órdenes NOUN orden | y CONJ y | algoritmos NOUN algoritmo | a ADP a | modo NOUN modo | de ADP de | controlar VERB controlar | el DET el | comportamiento NOUN comportamiento | físico ADJ físico | y CONJ y | lógico ADJ lógico | de ADP de | una DET uno | computadora NOUN computador | con ADP con | el DET el | objetivo NOUN objetivar | de ADP de | que SCONJ que | produzca VERB producir | diversas DET diverso | clases NOUN clase | de ADP de | datos NOUN dato | . PUNCT . |

En la tabla 5 en la fila que lleva por título *Etiquetado*, el texto es procesado por Spacy donde se puede observar que cada palabra analizada está separada por el símbolo "|". Cada fragmento de texto ubicado entre las barras verticales, corresponde a la forma palabra del texto, el etiquetado POS y el token.

Los pasos que se siguieron para obtener los contextos definitorios, se describen a continuación:

1) Por medio del etiquetado POS se identifica el verbo "ser" el cual corresponde a un auxiliar (AUX), y que generalmente es precedido por un determinante (DET). De esta forma, el PD puede tener la forma de AUX+DET.
2) Obtenido el término y el patrón definitorio, se requiere ubicar el patrón dentro del texto e identificar su contexto izquierdo para compararlo y verificar que sea el término que se busca. Si con el PD "ser" no se encuentra ninguna definición, se procede a buscar el PD "," dentro de los vectores que devuelve *spacy*; si tampoco se encuentra, se busca el PD de conjunción el cual está asociado a sinónimos del término y delante de él se encuentre la definición.

En la tabla 6 se muestran los conceptos que fueron identificados manualmente en cada tema (los términos abreviados que se observan en el renglón de títulos son: IS para Ingeniería de Software, CC para Ciencias de la computación y BD para Base de datos). Existen varios patrones

definitorios, el método implementado identifica los contextos definitorios con patrones como la coma, el verbo “ser” y sinónimos, para verificar cuantos patrones puede encontrar en cada tema, se hizo una verificación manual, además de otros que pudiesen sobresalir como: dos puntos o las viñetas. En cambio, la tabla 7 muestra la cantidad de conceptos que encontró automáticamente el script y con qué patrón verbal lo hizo, con esto se observa que se logra obtener más conceptos y que estos tienen mayor congruencia que el resultado de las dos herramientas anteriormente probadas, por lo tanto, se decide usar este script para el modelo.

Tabla 6. CD encontrados manualmente.

Tema	Lenguaje de programación	Matemáticas discretas	Teoría de la computación	Inteligencia artificial	IS	CC	Sistemas distribuidos	BD	Red neuronal artificial	Diagramas de flujo
CD con verbo ser	9	6	6	4	1	3	5	8	7	3
CD con sinónimos	1	1	2	0	0	2	1	1	0	4
CD con ","	0	0	2	2	5	2	2	0	0	1
CD con otro PD	0	3	0	4	4	3	2	1	1	1
Total de CD	10	7	10	6	6	7	8	9	7	8

Los términos abreviados que se observan en el renglón de títulos son: IS para Ingeniería de Software, CC para Ciencias de la computación y BD para Base de datos.

Tabla 7. CD encontrados por script.

Tema	Lenguaje de programación	Matemáticas discretas	Teoría de la computación	Inteligencia artificial	IS	CC	Sistemas distribuidos	BD	Red neuronal artificial	Diagramas de flujo
CD con verbo ser	9	4	6	4	2	4	5	7	4	3
CD con sinónimos	1	0	1	0	0	1	1	1	0	4
CD con ","	0	0	2	2	3	0	2	0	0	0
CD con otro PD	0	0	0	0	0	0	0	0	0	0
Total de CD	10	4	9	6	5	5	8	8	4	7

4.3.2 Extracción de metadatos en etiquetado HTML

Se le denominó metadato a la información que es útil para la generación de cuestionarios, es decir, aquella información que se desglosa a partir de un título, subtítulo o un enlistado de puntos, esto con la finalidad de poder generara una pregunta a partir del título y la respuesta será la descripción que tenga en consecuencia. Las herramientas que se utilizaron para la generación de este modelo fueron dos, Beautifulsoup4 para el web scraping el cual consiste en la extracción de información de documentos HTML y Wikipedia quien proporciona la información en formato HTML como se explicó en el módulo “Extracción de información de Wikipedia”.

Para describir mejor este modelo se utiliza el ejemplo del tema “Lenguaje de programación” que como primer paso es extraer su página HTML de Wikipedia una vez obtenida, se almacena el texto de las etiquetas <h> y <ul> esto con base a su dependencia, es decir, si <h1> tiene etiquetas <h2> anidadas y dentro de estas se tienen <ul> esta jerarquía será respetada al momento de guardarse en formato JSON. En la figura 4 se puede observar la estructura del HTML que representa la página web de Wikipedia sobre el tema, posteriormente se genera el JSON con los títulos y listas encontradas, el cual se explica más adelante.

Figura 4. Sección del HTML de la página de Wikipedia.

Siguiendo el ejemplo, parte del JSON queda como se muestra en la Figura 5. Donde "Nombre" es el título y que en este caso sería "Clasificación de los lenguajes de programación", respetando su jerarquía le sigue un subtítulo el cual este mismo contiene una lista. De este modo los metadatos se almacenan según su relevancia para posteriormente ser procesados y generar preguntas con mayor congruencia, las cuales se explica en los siguientes apartados.

```
{"Nombre": "Clasificacion de los lenguajes de
    programacion",
"Subtitulos": [{"NombreSub": "Clasificacion
    historica o por generaciones",
"lista": [{"punto": "Primera Generacion"},
{"punto": "Segunda generacion"},
{"punto": "Tercera Generacion"},
{"punto": "Cuarta generacion"},
{"punto": "Quinta generacion"}]}]}
```

Figura 5. Sección del archivo JSON.

4.4 Módulo “Generación de preguntas”

Para poder generar un cuestionario de preguntas con respuestas de opción múltiple se consideran los siguientes elementos:

✓ Tallo: pueden ser sentencias o una pregunta

✓ Opciones: respuestas

✓ Distractores: respuestas incorrectas

También se debe dejar en claro que para haber respuestas de opción múltiple el tallo debe estar bien redactado y con las siguientes condiciones:

➢ Solamente debe contener información relevante y necesaria, siendo su enunciado conciso y claro.

➢ Se debe evitar el uso de palabras o expresiones confusas que lleven a errores de interpretación.

➢ Su contenido no debe ser repetido en las preguntas.

En cambio, para las respuestas se debe tener la siguiente estructura:

❖ Debe tener una sola respuesta correcta.

❖ Cuando haya más de una opción correcta, se requerirá en el tallo seleccionar “la mejor respuesta”.

❖ Deben evitarse las preguntas negativas como “cuál de las siguientes no es.” o “todas las siguientes excepto”.

Con las recomendaciones anteriores, la plantilla para generar preguntas a partir de los CD’s es la siguiente: “¿Cuál de las siguientes es la definición correcta de « T » ?”, donde “T” es el término del CD del cual se está cuestionando, para su respuesta correcta es su definición, la cual se alternara dentro de 4 opciones conformadas por los incisos A), B), C), y D), donde 3 son definiciones de otros términos que fungen como distractoras y solo uno es el correspondiente al término del tallo, debido a que las distractoras son definiciones del mismo tema esto evita la obviedad de la respuesta correcta.

En el caso de las preguntas generadas a partir de información extraída del HTML se usó la plantilla que se muestra en la Figura 6, en la cual, << t >> corresponde al título y <<st>> corresponde a cada uno de los puntos de las listas o subtítulos. Por otra parte, la clasificación de mixto (Figura 6) se encarga de generar tallos de los cuales no lograron identificarse los puntos o subtítulos, esto pretende disminuir los errores de incongruencia. De esta forma se generan tallos mayormente concisos y claros, lo cual permiten realizar búsquedas de la respuesta correcta con mayor facilidad.

Como se observa en la Figura 6 se realizan preguntas de un tema en específico, por ejemplo: en el tema Lenguaje de programación se genera el tallo: De acuerdo a la Clasificación de los lenguajes de programación ¿Qué es la Clasificación por propósito?

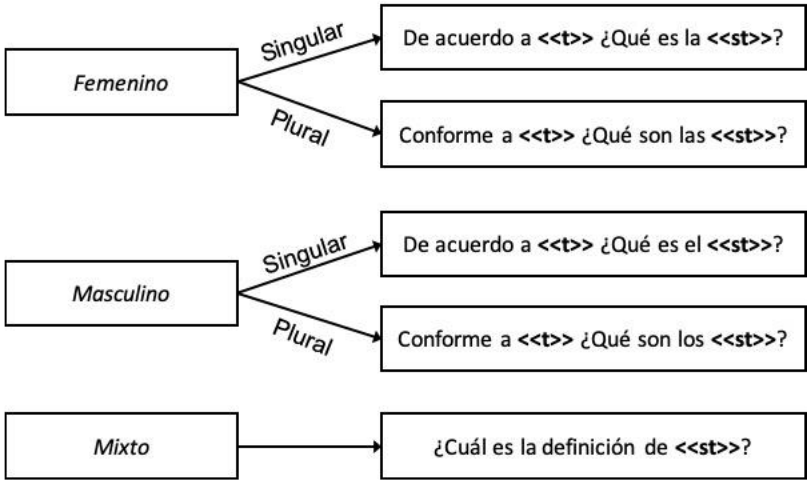

Figura 6. Plantilla para la generación de preguntas.

Tomando en cuenta este formato, para extraer la respuesta correcta se realiza una búsqueda con el término: Clasificación de los lenguajes de programación dentro de las etiquetas (<h>). Una vez identificado se convierte en nodo padre y dentro de él se encuentran sus nodos hijos.

Posteriormente, el nuevo nodo hijo Clasificación por propósito, debe encontrarse dentro de las etiquetas (ol, ul ó h). Una vez identificado, este se convierte en nodo padre para buscar dentro de sus nodos hijos la primera etiqueta << p >>, que pasa a ser la respuesta correcta.

Previo a haber generado las preguntas y su respectiva respuesta correcta, se generan las respuestas incorrectas (distractores). Esto se realiza de forma aleatoria tomando tres respuestas de otras preguntas. Es necesario mencionar que, pueden generarse varias preguntas a partir de un mismo título debido a que tienen varios subtítulos o listas y estas tienden a parecerse entre sí, con esto se evitará la obviedad de la respuesta correcta.

4.5 Módulo “Entrenamiento del chatbot”

Esta sección es la encargada de llevar el flujo de conversación y ayudar al usuario cuando desee saltarse una pregunta. La herramienta utilizada aquí es de IBM llamada *Watson Assistant* la cual proporciona un chatbot que se puede entrenar para cualquier tipo de conversación. Su estructura consiste de *Intents* (intenciones), *Entities* (entidades) y *Dialog* (diálogos).

Las intenciones consisten en propósitos u objetivos que se expresan en la entrada de un usuario, tales como responder a una pregunta o acceder a iniciar el cuestionario. Al reconocer la intención expresada en una entrada de usuario, el servicio Watson Assistant puede elegir el flujo de diálogo correcto para responder a la misma.

En cambio, las entidades representan información en la entrada de usuario que es relevante para la finalidad del usuario. Si las intenciones representan verbos (la acción que un usuario quiere llevar a cabo), las entidades representan nombres (el objeto o el contexto de una acción). Por ejemplo, si la intención es para obtener una previsión meteorológica, se necesitan las entidades relevantes correspondientes a ubicación y a fecha para que la aplicación pueda devolver una previsión precisa.

El diálogo utiliza las intenciones que se identifican en la entrada del usuario, además del contexto de la aplicación, para interactuar con el usuario y finalmente proporcionar una respuesta útil. El diálogo compara intenciones (lo que dicen los usuarios) con respuestas (lo que responde el bot). La respuesta puede ser la respuesta a una pregunta como ¿Dónde puedo poner gasolina? o la ejecución de un mandato, como encender la radio. La intención y la entidad podrían ser suficiente información para determinar la respuesta correcta. Otra posibilidad es que el diálogo solicite más información para responder correctamente.

Por lo tanto, las intenciones quedan como se muestran en la tabla 8, mostrando una breve descripción de lo que cada intención identifica en la respuesta del usuario.

Tabla 8. Intenciones de chatbot entrenado.

Intenciones	Descripción	No. ejemplos
#AceptarExamen	Dice que si hará examen	6
#Contador	Identifica cuantas faltan y cuantas quedan	8
#Despedida	Se despide	14
#EspPreguntas	Cuestiona sobre lo que le van a preguntar	5
#General_Connect_to_Agent	Solicita un agente humano	34
#General_Greetings	Saluda al bot	20
#General_Human	Pregunta si habla con un humano o un bot	12
#inicio	El bot inicia la conversación	3
#NegarseExamen	Se niega a hacer el examen	13
#pausarExamen	Pregunta si puede pausar el examen	7
#PregConfusas	El alumno no compre las preguntas	8
#PregIdenti	Pregunta sobre que eres	5
#RespCorrecta	Responde mal a las preguntas	12
#SaltarPregunta	Pide otra pregunta	6
#Saludo	Saludo de inicio	5
#SeguirExamen	Sigue el examen después de dudarlo	5

Las entidades se muestran en la Tabla 9. En este caso, en los de tipo sinónimo se agregaron las diferentes variables de cómo se pueden referir a la entidad; en cambio, en patrones se definieron día, noche y tarde por la forma en que puede ser usado en un saludo.

Tabla 9. Entidades de chatbot entrenado.

Entidad	Valores	Tipo
@cuestionario	cuestionario	Sinónimo
@FalComprencion	comprendo	Sinónimo
@hora	Dia, noche, tarde	Patrones
@profesor	doctor	Sinónimo
@Soy	Evaluador	Sinónimo

La estructura de los diálogos se diseñó como se observa en la Tabla 10. La finalidad de los diálogos es dar asistencia cuando se desea saltar pregunta, cancelar cuestionario, iniciar cuestionario e instrucciones de como contestar.

Tabla 10. Nodos de diálogos de chatbot entrenado.

Diálogo	Entidad	No de respuestas
Seguir	#SeguirExamen	1
Contar	#Contador	1
Pausa	#pausarExamen	1
Despedida	#Despedida	1
AceptarEval	#AceptarExamen	1
Inicio	#inicio || #General_Greetings || #Saludo	1
SaltoPregunta	#SaltarPregunta	1
PregIdentificacion	#PregIdenti || #General_Human_or_bot	1
EspecificaEvaluacion	#EspPreguntas	1
PregConfusa	#PregConfusas	1
NegaExamen	#NegarseExamen	1
ExigeeHumano	#General_Connect_to_Agent	1
FueraTema	Anything_else	1

Capítulo V Evaluación de módulos y resultado de pruebas

5.1 Evaluación de preguntas y respuestas de opción múltiple

Para evaluar los cuestionarios de respuestas de opción múltiple se identificó el trabajo de Rivera Jiménez, Flores Hernández, Alpuche Hernández, & Martínez González (2017) que es un cuestionario que funciona como instrumento de evaluación para la estructura correcta de un cuestionario de opción múltiple, abarcando los factores de comprensión del reactivo, contenido del reactivo, precisión del reactivo y redacción de opciones de respuesta. Todos ellos haciendo un total de 14 preguntas de evaluación.

Los elementos utilizados para evaluar los cuestionarios fueron solamente los referentes a los factores de comprensión del reactivo, precisión del reactivo y la redacción de opciones de respuesta. Se omitió el factor del contenido del reactivo debido a que se basa en verificar que el reactivo sea diseñado a partir del tema que se está evaluando, lo cual en este caso se da por hecho, porque las preguntas y respuestas se generan a partir de la información del tema deseado.

Por lo tanto, el instrumento de evaluación se redujo a un total de 8 preguntas, las cuales son:

- Comprensión del reactivo
  1. ¿La cantidad de texto en el tallo es adecuada para su comprensión?
  2. ¿La pregunta o instrucción se encuentra redactada con claridad?
  3. ¿El reactivo cuenta con una gramática, puntuación y ortografía correctas?
- Precisión del reactivo
  4. ¿Las opciones son independientes entre sí?
- Redacción de opciones de respuesta
  5. ¿Las opciones son similares en cuanto a estructura gramatical, contenido y extensión?
  6. ¿Las opciones evitan dar pistas sobre la respuesta correcta?
  7. ¿Los distractores son plausibles, es decir, no se descartan por inferencia lógica o sentido común?
  8. ¿El reactivo cuenta con tres o cuatro opciones de respuesta?

El instrumento de evaluación se proporcionó a tres evaluadores, dos con grado de Maestro en Ciencias de la computación y uno con estudios de Ingeniería en informática. Posteriormente se les dieron 3 cuestionarios con 12 reactivos cada uno y 1 cuestionario con 6 reactivos, haciendo un total de 42 reactivos. Los reactivos consisten en tallo (pregunta), respuestas distractoras y respuesta correcta.

Se seleccionaron tres reactivos de cada uno de los temas de Ciencias de la computación, por ejemplo, del tema de lenguaje de programación se seleccionaron 3 reactivos creados con el método de CD y 3 con el método HTML. Dando como resultado 18 reactivos evaluados a partir de CD y 24 a partir de HTML. Por lo tanto, el instrumento de evaluación se aplicó para 42 reactivos.

Para cada pregunta del instrumento de evaluación se contestó una escala Likert, que consiste en 5 niveles:

1. Totalmente en desacuerdo
2. En desacuerdo
3. Neutral
4. De acuerdo
5. Totalmente de acuerdo

5.1.1 Evaluador A

Los resultados en el factor de comprensión del reactivo fueron los que se muestran en la figura 7, donde se observa que en promedio evalúa con 4, lo equivalente a decir que si en cada pregunta de evaluación.

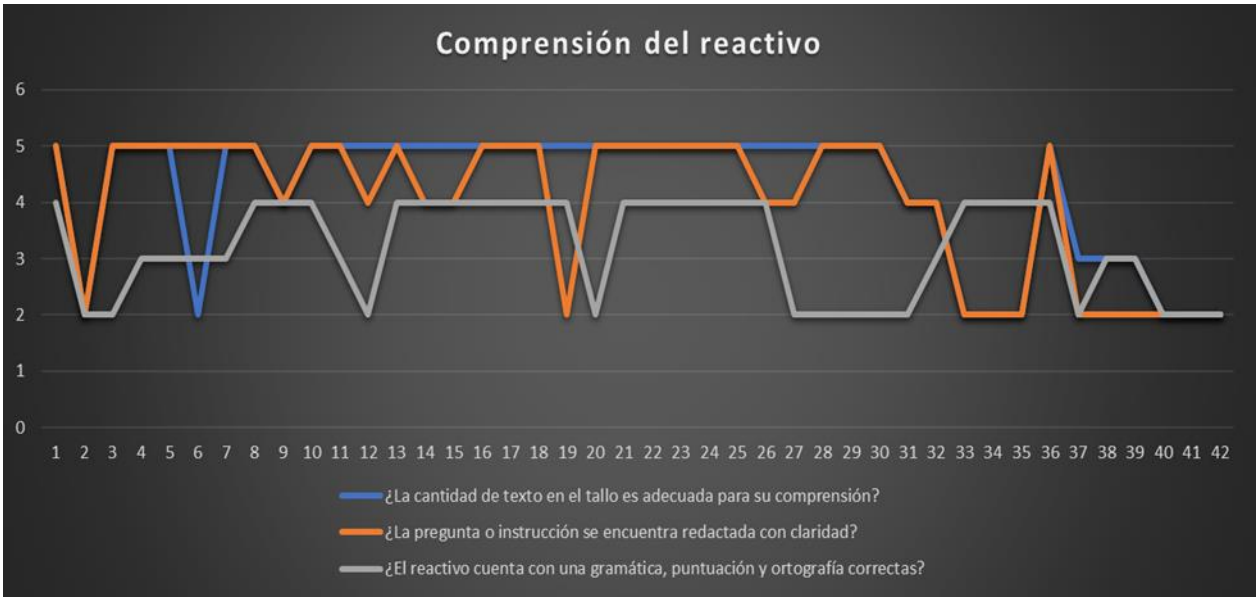

Figura 7. Compresión del reactivo, evaluador A.

Para el factor de precisión del reactivo se muestra sus resultados en un promedio de 4 tal y como se muestra en la figura 8.

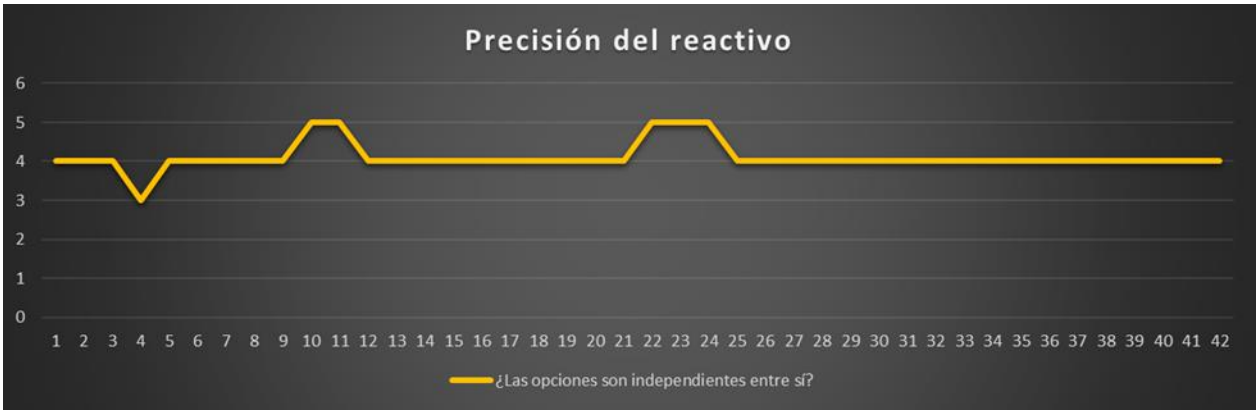

Figura 8. Precisión del reactivo, evaluador A.

En el factor de la redacción de opciones de respuestas al igual que los otros factores mantiene su promedio en un 4, figura 9. Esto nos da un promedio en general de 4 lo cual indica que el evaluador en su mayoría aprueba los reactivos, sin embargo, si se analizan las gráficas, se puede observar que en los reactivos de HTML sus evaluaciones tienden a bajar.

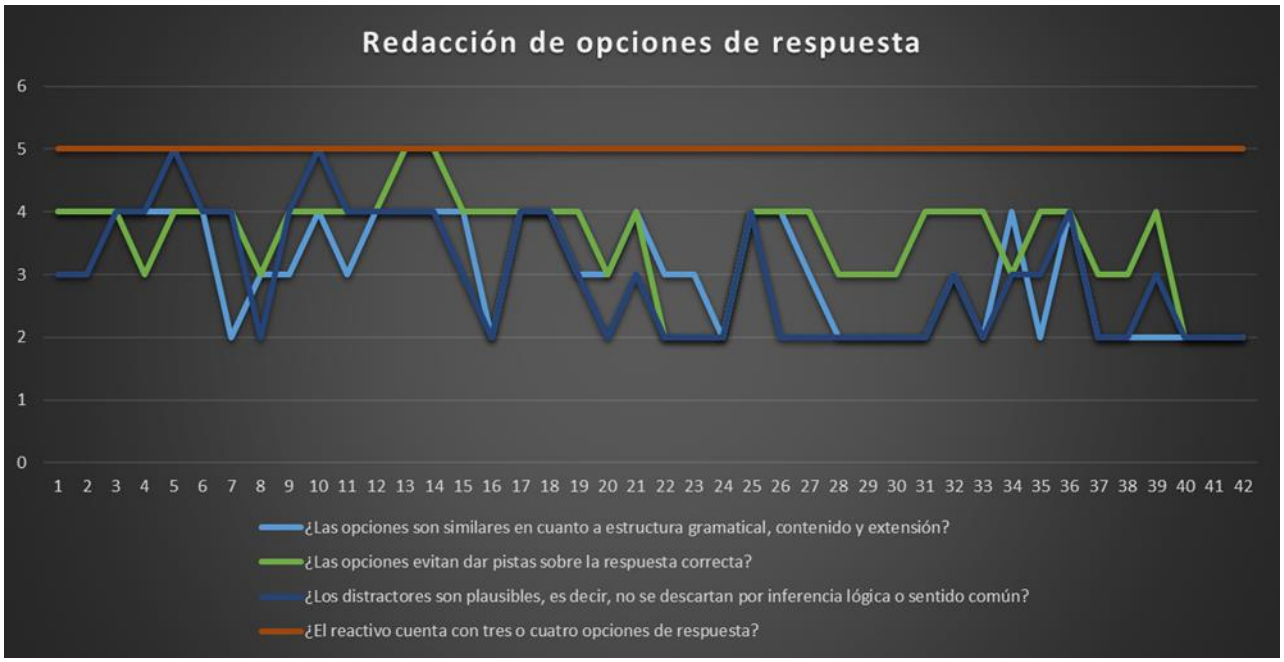

Figura 9. Redacción de opciones de respuesta, evaluador A.

5.1.2 Evaluador B

Con el evaluador B se obtuvo un promedio de 5, pero a pesar de tener una buena aceptación por parte del evaluador se nota una caída (figura 10) en las últimas preguntas exactamente a partir de la 30, estas últimas fueron generadas con el método HTML el cual nos dice que no está siendo lo suficientemente efectivo, no por lo menos para este evaluador.

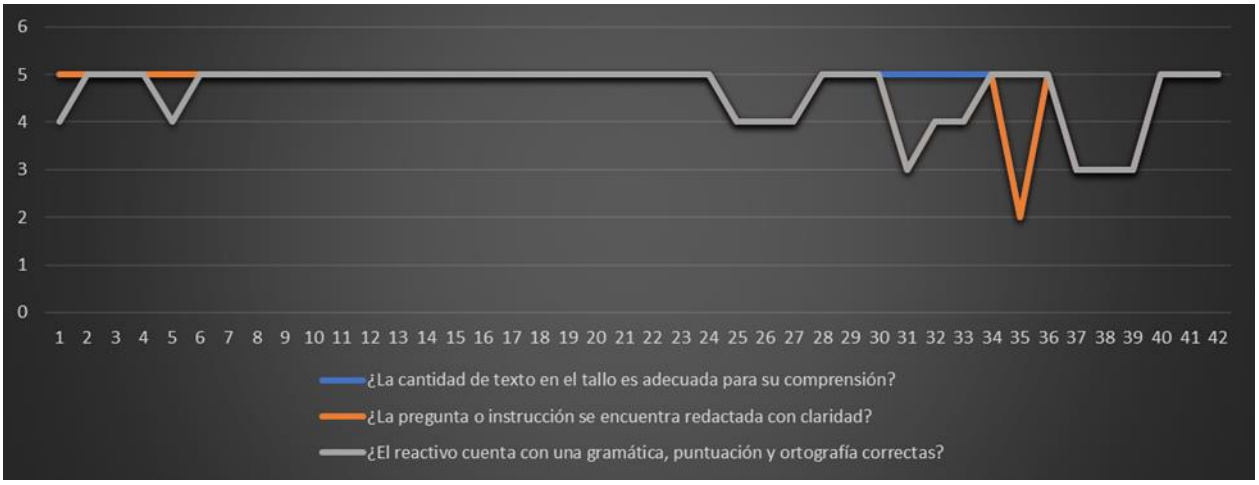

Figura 10. Compresión del reactivo, evaluador B.

En el caso de la figura 11 se muestra una aceptación notable y constante, esto hace referencia a que las preguntas y respuestas de contextos definitorios como las de HTML contienen precisión en el reactivo. El factor de la figura 12 se muestra una inconformidad con las últimas preguntas, redactadas con el método HTML, pero aun así logra tener un promedio de 4.

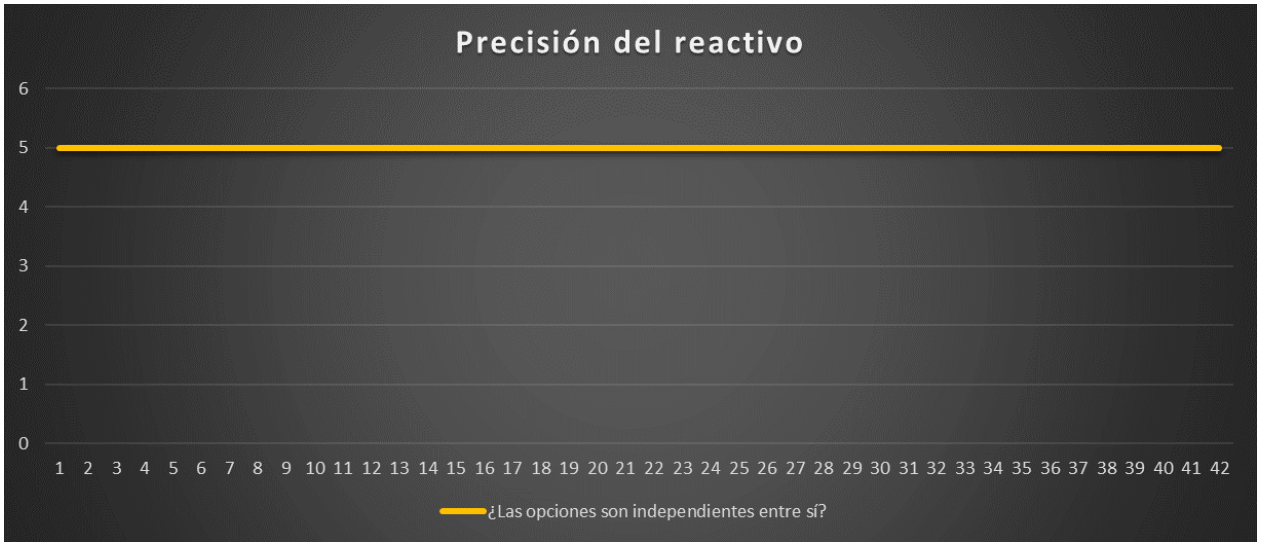

Figura 11. Precisión del reactivo, evaluador B.

Figura 12. Redacción de opciones de respuesta, evaluador B.

5.1.3 Evaluador C

La figura 13 muestra evaluaciones bajas en las últimas preguntas tal y como el evaluador B, pero en promedio sigue manteniendo una evaluación de 4.

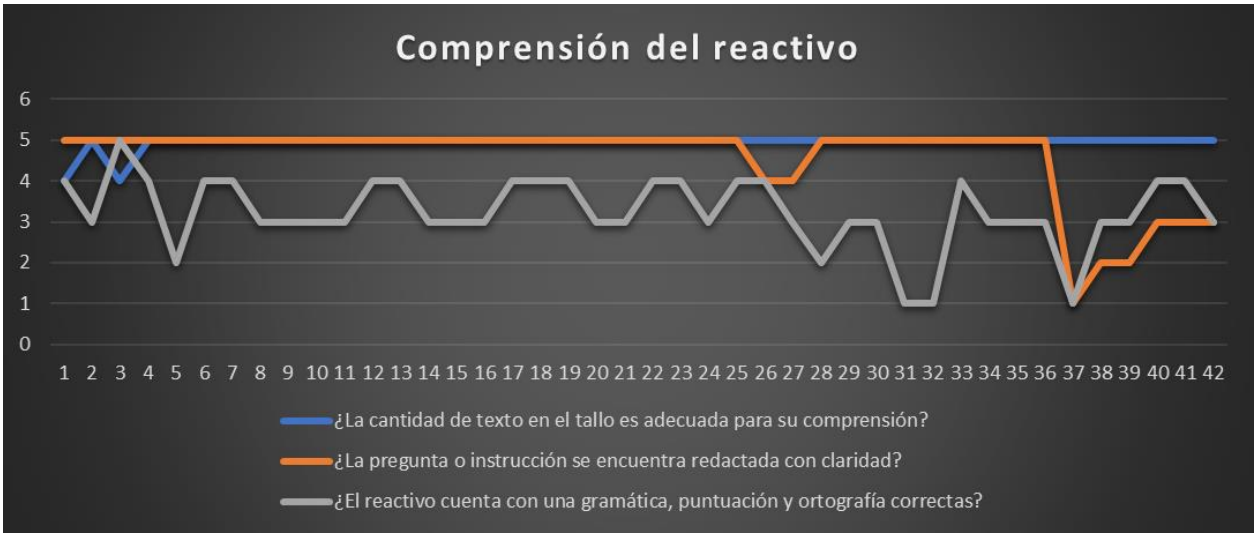

Figura 13. Comprensión del reactivo, evaluador C.

Aquí se puede observar (figura 14) que a comparación de los evaluadores anteriores este evaluó de manera neutra, dejando en promedio una evaluación de 3.

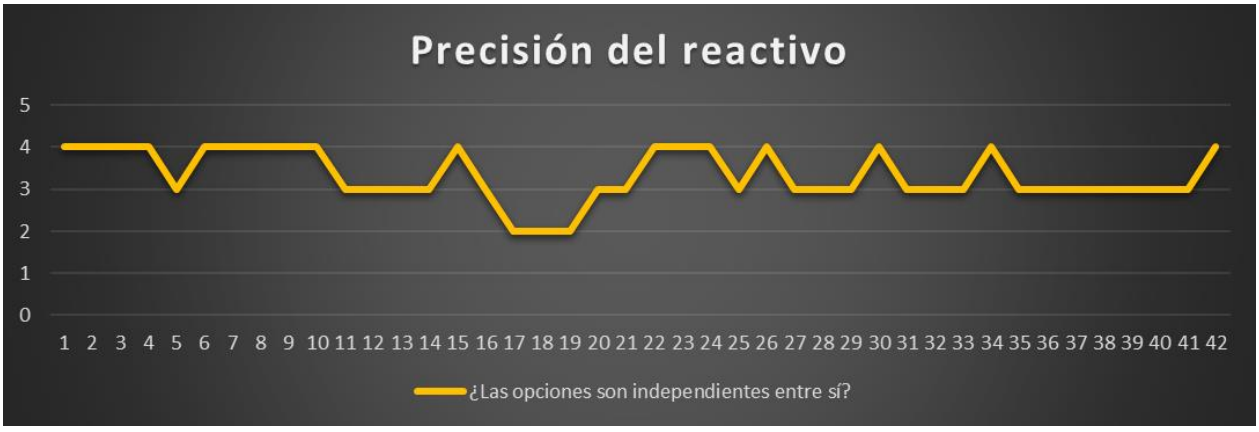

Figura 14. Precisión del reactivo, evaluador C.

En la redacción de opciones de respuesta (figura 15) se observa al igual que el evaluador anterior una caída de evaluación en las últimas preguntas, de tal modo que se decide buscar una métrica para medir la concordancia de los 3 evaluadores.

Figura 15. Redacción de opciones de respuestas, evaluador C.

5.1.4 Comprobación de concordancia

Esta evaluación muestra la concordancia que tuvieron los evaluadores en sus calificaciones sobre los reactivos. Se utiliza el índice Kappa porque este refleja la concordancia inter-observador y puede ser calculado en tablas de cualquier dimensión, siempre y cuando se contrasten dos observadores, en este caso para la evaluación de concordancia de tres o más observadores se utiliza el coeficiente Kappa de Fleiss, El coeficiente Kappa puede tomar valores entre -1 y +1, mientras más cercano a +1, mayor es el grado de concordancia inter-observador, por el contrario, mientras más cercano a -1, mayor es el grado de discordancia inter-observador.

Para poder sacar el coeficiente de kappa de Fleiss primero fue necesario promediar la evaluación de cada reactivo, lo que resulto en la Tabla 11.

Tabla 11. Promedio de evaluaciones.

Pregunta	Evaluador			Pregunta	Evaluador		
	A	B	C		A	B	C
1	4	5	5	22	4	4	4
2	3	5	5	23	4	4	4
3	4	5	5	24	4	4	4
4	4	5	4	25	4	4	4
5	4	4	4	26	4	4	4
6	4	5	4	27	4	4	4
7	4	5	4	28	4	5	4
8	4	4	4	29	4	5	4
9	4	5	4	30	4	5	4
10	5	5	4	31	3	3	3
11	4	5	4	32	4	4	3
12	4	5	4	33	3	3	4
13	5	5	4	34	3	5	4
14	4	5	4	35	3	4	4
15	4	4	4	36	4	5	4
16	4	5	4	37	3	3	3
17	4	4	4	38	3	3	3
18	4	4	4	39	3	3	3
19	4	4	4	40	3	4	3
20	4	5	4	41	3	4	3
21	4	4	4	42	3	4	4

Visualmente se puede observar la concordancia de evaluación en la figura 16, donde se muestra cómo los evaluadores coinciden en algunos reactivos con la misma calificación, sin embargo, esto no asegura que los evaluadores coincidan en estar de acuerdo o desacuerdo sobre la redacción del reactivo. Para poder medir eficazmente la evaluación del reactivo se consideró que si tiene un promedio menor igual a 3 el evaluador está en desacuerdo y por lo tanto el reactivo no cuenta con las características de evaluación, en cambio, sí se encuentra con un promedio mayor a 4 el evaluador consideró que el reactivo es correcto y por lo tanto cuenta con las características de evaluación propuestas.

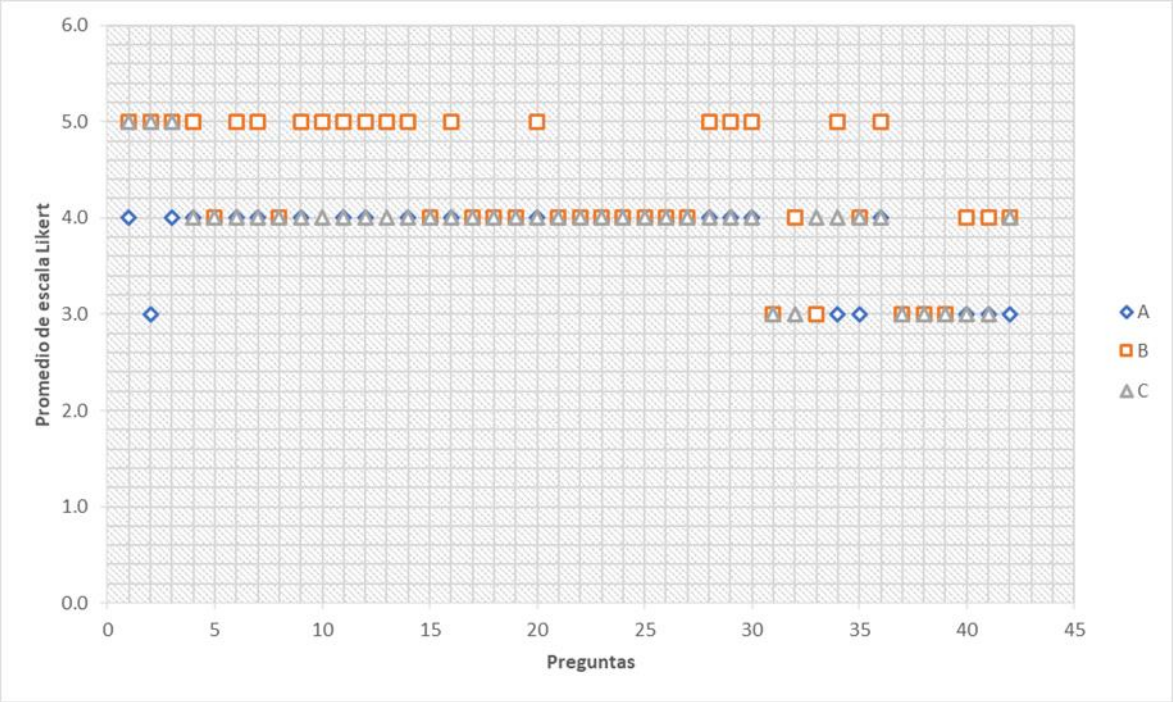

Figura 16. Gráfica del promedio de evaluaciones.

En la tabla 12 se puede observar la columna de total que se refiere al coeficiente de Kappa de Fleiss este resultado expresa el valor de concordancia en los evaluadores al dar las mismas calificaciones a los reactivos dejando un resultado de 0.575, que en la teoría se toma como una concordancia media. En otras palabras, los 3 evaluadores en mayor parte concordaron con sus calificaciones a los reactivos

La columna que indica <=3 hace referencia a la concordancia que tuvieron los 3 evaluadores en calificar como una mala redacción de los reactivos, en cambio la columna faltante indica que hubo una concordancia media al evaluar los reactivos con una buena redacción. Dejando como entendido que los evaluadores en su mayoría de calificación concordaron en estar de acuerdo en que los reactivos están correctamente redactados.

Tabla 12. Coeficiente de Kappa de Fleiss.

Total	< = 3	> 3
0.575	0.574	0.575

5.2 Evaluación del chatbot

El chatbot se probó con 316 mensajes, de los cuales solo se evaluaron 202 debido a que los primeros mensajes ayudaron a identificar algunas intenciones que no se tomaron en cuenta. Los mensajes evaluados y que el chatbot respondió correctamente (tomados como verdaderos positivos o VP), se muestran en la Tabla 13.

Tabla 13. Intenciones detectadas VP.

Intenciones	
RespCorrecta	60
inicio	26
AceptarExamen	18
Despedida	10
SaltarPregunta	13
NegarseExamen	9
Saludo	7
SeguirExamen	7
Contador	2
Total de mensajes	152

Watson cuenta con una intención llamada “Irrelevante”, aquí se consideran todos los mensajes de los cuales el chatbot no logra identificar alguna intención de las entrenadas, por lo tanto, los considera como mensaje fuera del contexto. Estos se clasificarán como VP. En el caso de clasificarlo como Irrelevante y pertenecer a una intención se consideró como falso negativo (o FN) tal y como se muestra en la Tabla 14.

Tabla 14. Intenciones detectadas como irrelevantes.

Usuario	Chatbot	Intención
Comencemos	Irrelevante	inicio
terminat cuestionario	Irrelevante	Despedida
claro	Irrelevante	AceptarExamen
adiós	Irrelevante	Despedida
ssalir	Irrelevante	Despedida
salir	Irrelevante	Despedida

Se consideran falsos positivos (FP) aquellos mensajes identificados con una intención errónea, tal y como se muestra en la Tabla 15.

Tabla 15. Intenciones detectadas erróneamente.

Usuario	Chatbot
de que trata el cuestionario	Despedida
terminar pregunta	Despedida
total de preguntas	EspPregunta
total de preguntas	EspPregunta
total de preguntas	EspPregunta
total de preguntas?	EspPregunta
total de preguntas	EspPregunta
que?	Irrelevante

Los resultados quedan como se muestra en la Tabla 16 y el cálculo de cada métrica explica lo siguiente:

- Precisión que consiste en saber qué mensaje de los que identifica como correctos son realmente correctos.
- La cobertura es la fracción de los mensajes correctos que selecciona el chatbot.
- La exactitud consiste en afirmar que fracción de los mensajes identifica e ignora correctamente.
- Tasa de error consiste en saber que tanto se equivoca en la clasificación de los mensajes como correctos e incorrectos.
- F1 score es la combinación de las medidas de precisión y cobertura para devolver una medida de calidad más general del modelo.

Estas medidas hablan sobre la capacidad que tiene el chatbot de identificar los mensajes que se le envían y poder responder adecuadamente con base en lo que se le entrenó, es decir, los porcentajes expresan que todos los mensajes obtenidos son respondidos ayudando en la asistencia

de poder concluir el cuestionario, seguir contestando a pesar de no haber contestado nada, saltar preguntas o expresar la falta de comprensión de la pregunta generada.

Tabla 16. Métricas de evaluación al chatbot.

Métricas	VP	VN	FN	FP
Total	152	36	6	8
Precisión	95.0%			
Exactitud	93.1%			
Cobertura	96.2%			
Tasa de error	6.9%			
F1 score	95.6%			

Capítulo VI
Conclusiones

6.1 Conclusiones y trabajos futuros

Se presentó el desarrollo de una herramienta que permite generar preguntas de opción múltiple a partir de un tema ingresado por el usuario. La información de dicho tema se busca y se extrae desde Wikipedia. Los fragmentos relevantes de dicha información se identificaron utilizando contextos definitorios y metadatos del etiquetado HTML. La interacción con este sistema se realiza a través de un chatbot implementado con Watson Assistant.

La herramienta se evaluó por secciones debido a la complejidad y los dos diferentes métodos de extracción de información para la generación de preguntas de opción múltiple, así como también la evaluación sobre la creación de los reactivos, en particular los generados a través del contexto definitorio.

La evaluación de los reactivos generados de manera automática se realizó a través de un instrumento que abarca factores sobre comprensión del reactivo, precisión del reactivo y la redacción de opciones de respuesta. Para la evaluación participaron tres evaluadores a los que se les proporcionaron tres cuestionarios con 12 reactivos cada uno y un cuestionario con seis reactivos, haciendo un total de 42 reactivos. La evaluación se realizó utilizando una escala Likert y se utilizó el coeficiente Kappa de Fleiss para medir el nivel de concordancia de los evaluadores, el cual debe situarse en un intervalo de -1 a 1. El valor de concordancia obtenido entre los evaluadores fue de 0.575.

Por otro lado, el chatbot a pesar de ofrecer resultados favorables, se entrenó con un número limitado de intenciones, entidades y nodos de diálogo, debido a que solo da asistencia a evaluaciones estructuradas con indicaciones muy precisas, de manera que no interfiere en la creación del cuestionario o extracción de información. Aunque lo ideal sería que se entrenara con

cada diferente tema del cual se genera el cuestionario para poder tener un poco de información acerca del tema.

Como posibles trabajos futuros se propone la implementación de mejores técnicas para la extracción de información en el método de HTML para la generación de mejores reactivos y aumentar la evaluación en escala Likert con el instrumento de Rivera Jiménez, Flores Hernández, Alpuche Hernández, & Martínez González (2017). Es decir, el método de solución general puede ser aplicable en un futuro si se mejora la sección de HTML.

Anexos

Métodos alternos para la identificación de contextos definitorios

Como trabajos relacionados que abordan la identificación de contextos definitorios, se exploraron dos métodos alternos al propuesto en el método de solución de esta investigación:

1) Tema de tesis desarrollado en la misma institución, titulado Identificación de las relaciones de hiperonimia en contextos definitorios por medio de un autómata determinista (Hipólito, 2018).
2) Herramienta Watson NLU ("Watson Natural Language Understanding", 2019).

Las soluciones y recursos propuestos por ambos métodos, fueron implementados en esta investigación y se experimentó con ellos para determinar su desempeño en términos de precisión y cobertura. Los resultados se mencionan a continuación.

Identificación de las relaciones de hiperonimia en contextos definitorios por medio de un autómata determinista

En este trabajo se analizan los contextos definitorios sintácticos, de modo que el patrón definitorio (PD) es el verbo “ser”, y a partir de éstos verifica si existen relaciones de hiperonimia. El trabajo se realizó sobre textos médicos y por lo tanto se exploró para verificar si era útil con las áreas que aborda esta tesis.

El método propuesto por Hipólito extrae información de Wikipedia, y como salida genera un documento tipo texto plano con los contextos definitorios y sus respectivos hiperónimos. El texto que devuelve el método se conforma por diversos campos que se separan con el símbolo numeral “#”. Sirva el siguiente ejemplo como muestra:

La IA es una de las disciplinas más nuevas junto_con la genética moderna.#IA#disciplinas

La oración inicial, antes del primero signo numeral, indica el texto original. El segundo campo refiere al término que se define y tercero a su hiperónimo.

Se realizaron pruebas con los temas indicados en la primera columna de la tabla 17 y se obtuvieron los resultados que se mencionan en la misma tabla.

Tabla 17. Pruebas con trabajo de (Hipólito, 2018) explorando temas de Ciencias de la Computación.

Tema	Conceptos correctos	Conceptos incorrectos	Conceptos totales
Lenguaje de programación	2	3	5
Lenguaje máquina	1	0	1
Matemáticas discretas	3	0	3
Teoría de la computación	1	1	2
Reconocimiento de patrones	0	1	1
Red neuronal artificial	3	1	4
Cómputo distribuido	0	1	1
Procesamiento del lenguaje natural	0	0	0
Inteligencia artificial	2	5	7

Puede apreciarse el total de conceptos identificados de los cuales el total de correctos es menor a cuatro, dejando como improbable la redacción de preguntas y respuestas de opción múltiple. Los considerados incorrectos se llaman así debido a que no tienen una definición o porque no son funcionales para la generación de preguntas y respuestas.

A continuación, se muestran diversas definiciones que devuelve el método:

la IA es una rama de las ciencias computacionales encargada de estudiar modelos de cómputo capaces de realizar actividades propias de los seres humanos en base a dos de sus características primordiales : el razonamiento y la conducta. [5] <np>#IA#rama

La representación es una cuestión clave a la hora de encontrar soluciones adecuadas a los problemas planteados.#representación#cuestión clave

La IA es una de las disciplinas más nuevas junto_con la genética moderna.#IA#disciplinas

la inteligencia es un programa capaz de ser ejecutado independientemente de la máquina que lo ejecute, computador o cerebro ».#inteligencia#programa capaz

de caminos, la navegación es un subcampo de la IA de el juego que se centra en dar a los PNJ la capacidad de navegar en su entorno, la búsqueda de un camino hacia un objetivo, evitando colisiones con otras entidades o colaborar con ellos.#navegación#subcampo

el tablero es una estructura de datos de tipo matriz que contiene unas casillas las cuales están ocupadas por un jugador o vacías.#tablero#estructura de datos

Una partida es una secuencia de estados por los que pasa un tablero.#partida#secuencia de estados

En la tabla 18 se agrupan los campos por columnas.

Tabla 18. Pruebas con trabajo de (Hipólito, 2018) explorando tema de Inteligencia Artificial.

N	Término	Definición
1	#IA#	la IA es una rama de las ciencias computacionales encargada de estudiar modelos de cómputo capaces de realizar actividades propias de los seres humanos en base a dos de sus características primordiales : el razonamiento y la conducta. [5] <np>
2	#representación#	La representación es una cuestión clave a la hora de encontrar soluciones adecuadas a los problemas planteados.
3	#IA#	La IA es una de las disciplinas más nuevas junto_con la genética moderna.
4	#inteligencia#	la inteligencia es un programa capaz de ser ejecutado independientemente de la máquina que lo ejecute, computador o cerebro ».
5	#navegación#	de caminos, la navegación es un subcampo de la IA de el juego que se centra en dar a los PNJ la capacidad de navegar en su entorno, la búsqueda de un camino hacia un objetivo, evitando colisiones con otras entidades o colaborar con ellos.
6	#tablero#	el tablero es una estructura de datos de tipo matriz que contiene unas casillas las cuales están ocupadas por un jugador o vacías.
7	#partida#	Una partida es una secuencia de estados por los que pasa un tablero.

Como se mencionó anteriormente, hay algunos términos que no se consideran útiles, como los expuestos en 2, 4, 5, 6 y 7 que se observan en la tabla 18, esto se debe a que cuentan con la estructura de un Contexto definitiorio, no son viables para la generación de preguntas debido a la discordancia con el tema. Por todas estas razones, este método fue descartado.

Watson NLU

La segunda herramienta con la que se experimentó fue Watson NLU ("Watson Natural Language Understanding", 2019), que es una colección de API que ofrece análisis de texto a través del Procesamiento de lenguaje natural. Este conjunto de API puede analizar texto ayudando a extraer sentimientos, emociones, palabras clave, entidades, categorías, conceptos, sintaxis y roles semánticos. Además, puede crear un modelo personalizado para algunas API para obtener resultados específicos que se adaptan al tema deseado, estos son sus modelos de paga, los cuales deben ser entrenados con textos de los temas deseados con la finalidad de tener mayor eficiencia en la identificación de conceptos, palabras clave entre otros.

Para poder utilizar esta herramienta, mediante la librería de Watson NLU se le envía la URL de Wikipedia sobre el tema que se realizarán las preguntas, y por medio de la opción de conceptos, palabras clave y entidades se podrán realizar dichas preguntas, como nota, el número mayor de palabras clave que puede extraer NLU de la URL del tema son 50, estas vienen acompañadas de un valor entre 0.0 y 1.0 que indica la relevancia sobre el tema, y de la cual, no se menciona cómo se genera. El número mayor de conceptos que se pueden extraer son 8, calificados con el mismo rango de relevancia. En la extracción de entidades el rango mayor es de 50 y calificados de la misma forma, la única diferencia que incluye es la entidad a la que pertenece, por ejemplo: “IBM - .95 – Company”, donde la palabra es *IBM* con un valor .95 de relevancia y la entidad a la que pertenece es *company*.

Tomando en cuenta los 10 temas de los que se hicieron pruebas como con el método anterior, Watson NLU arrojó un total de 489 datos, donde 409 fueron palabras clave y 80 conceptos. Analizando a detalle y tratando de generar preguntas conceptuales, se notó que la mayoría de las palabras no eran funcionales un ejemplo de esto, se observa en la tabla 19

Tabla 19. Resultados de consulta a Watson NLU con temas de Ciencias de la Computación.

Lenguaje de programación	Matemáticas discretas	Teoría de la computación	Inteligencia artificial	Ingeniería de software	Ciencias de la computación	Sistemas distribuidos	Base de datos	Red neuronal artificial	Diagramas de flujo
Paradigma de programación	matemáticas discretas	Nachum Dershowitz	Inteligencia	proceso unificado	computer science	Grid Services Architecture	datos informaciones	largo plazo	von neumann
Compilador	geometría discreta	Yuri Gurevich	inteligencia artificial	Software Craftmanship	calculadora mecánica	«sistemas distribuidos»	modelo jerárquico	redes neuronales cmac	único punto
Lenguaje de programación	ciencia computacional teórica	cálculo lambda	sistemas expertos	World Wide Web	Computación IEEE	World Community Grid	OpenOffice.org Base	redes neuronales llamadas	Triángulo boca
Programación imperativa	tópicos continuos	cómputo estudia abstracciones	«la inteligencia	única solución	International Business Machines	Grid Engine	múltiples sitios	neuronales profundas feedforward	Microsoft Visio
Programación funcional	distribuciones discretas	«A natural axiomatization	John McCarthy	enterprise architects	Computer History Museum	múltiples computadoras	Database Management System	neuronales biológicas modelos	«crea diagramas
Programador	probabilidad discreta	clásicos modelos	término inteligencia	Herramienta CASE	Segunda Guerra Mundial	Single System Image	axiomas base	corto plazo	Formcraft Standard Register
Lenguaje formal	colecciones discretas	Alan Turing	«inteligencia artificial»	mejores resultados	término ciencias	computación zombi	Structured Query Language	módulos neuronales abstractos	john von neumann
alto nivel	Eric W	Principales subramas	Marvin Minsky	Computer Aided	frase «Las ciencias	New Computing Infrastructure	Información Pública Gubernamental	neuronales controladores	actividades uml
lenguaje ensamblador	prolífico leonhard euler	estado abstracto	Alan Turing	modelo cocomo	Mayores logros	Data Distribution Service	Rafael Llanos Ferraris	redes funcionales multicapas	Gómez Cejas
semántica estática	fórmulas combinatorias explícitas	inclusive lenguajes	«existirá inteligencia	application developmen t	Howard Aiken	Fail Over	amplio rango	múltiples capas	Gómez Rondón
tipos explícitos	topología combinatoria	mayores logros	«what is artificial	Business Process Model	Computer Programs	entorno multiusuario	nodo padre	Redes discretas	lógica requerida
lenguajes multiparadigma	Clay Mathematics Institute	código secreto	Semantic Information	cabo numerosas tareas	Kurt Gödel	Globus Toolkit	SGBD participantes	,el aprendizaje	actividad uml

En dicha tabla se puede observar en la columna del tema “ciencias de la computación” el concepto extraído “segunda guerra mundial” y con esto se deduce que, al momento de hacer la pregunta, la respuesta podría desviarse del tema, dicho de otro modo, se tiene desambiguación y discordancia con la información del tema que se envió. También se observó que clasificar las palabras por relevancia no era útil, ya que se encontraban palabras con relevancia mayores a .90 y no guardaban relación con el tema, aplica tanto para entidades, conceptos y palabras clave. Esto se ve reflejado en la tabla 20 donde el tema “Lenguaje de programación” obtuvo 50 palabras clave, pero el rango de relevancia entre ellas es de .52 y .92, dejando palabras clave importantes como “Lenguaje Fortran” con una relevancia de .63 y “tipo implícitas” con .89.

En la Tabla 20 se muestra el resumen de los resultados de las pruebas, a pesar de tener un apartado donde se consideran las palabras importantes, estas son difíciles de extraer, debido que al momento de obtenerse estas vienen de forma aleatoria y deben ser seleccionadas por un usuario. Es por esto que se descartó utilizar esta herramienta.

Tabla 20. Resultado de la búsqueda en temas de Ciencias de la Computación con Watson NLU.

Tema	Total				Considerados útiles			
	keywords	relevancia	conceptos	relevancia	keywords	relevancia	conceptos	relevancia
Lenguaje de programación	50	.52-.92	8	.99-1	10	.61-.92	0	.99-1
Matemáticas discretas	38	.51-.99	8	.97-.99	4	.61-.99	0	.97-.99
Teoría de la computación	17	.64-.93	8	.98-1	4	.88-.93	0	.98-1
Inteligencia artificial	50	.30-.99	8	.99-1	6	.31-.99	0	.99-1
Ingeniería de software	50	.72-.96	8	.95-.99	3	.83-.96	0	.95-.99
Ciencias de la computación	50	.46-.97	8	.97-.99	6	.63-.97	0	.97-.99
Sistemas distribuidos	26	.55-.95	8	.94-1	13	.59-.95	0	.94-1
Base de datos	36	.59-.97	8	.92-1	8	.69-.97	0	.92-1
Red neuronal artificial	50	.53-.91	8	.97-.99	6	.62-.92	0	.97-.99
Diagramas de flujo	42	.51-.98	8	.94-1	4	.73-.98	0	.94-1

Bibliografía

Amidi, S. (Otoño de 2018). *Stanford University*. Obtenido de https://stanford.edu/~shervine/l/es/teaching/cs-229/hoja-referencia-aprendizaje-automatico-consejos-trucos

Arrufat, M. J., Sánchez, V. G., & Santiuste, E. G. (2015). Tendencias en la evaluación del aprendizaje en cursos en línea a masivos y abiertos. *Educación XXI*, 77-96.

Barberá, E. (2016). Aportación de la tecnología a la e-Evaluación. *Revista de educación a distancia*, 50-60.

Benotti, L., Martinez, M. C., & Schapachnik, F. (2017). A tool for introducing computer science with automatic formative. *IEEE Transactions on Learning technologies*, 179-192.

Binda, M. d. (2006). Consideraciones sobre el Examen de Preguntas de Opción Múltiples(Multiple choice). *Revista Argentina de Radiología*, 337-339.

Blanco, S., Vélez, O. F., Marcela, C., Tobón, Z., & Jairo, J. (2015). Evaluación de conocimineto con exámenes de selección múltiple: ¿tres o cuatro opciones de respuesta? Experiencia con el examen de admisión a posgrados médico-quirúrgicos en la Universidad de Antioquia. *latreia (Medellin)*, 300-311.

Centeno Brambila, D. A., & Lira Obando, A. (2015). Sistema de evaluaciones en línea como herramienta para los niveles de educación media superior. *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo, vol. 6, núm. 11*.

Crisp, G., Guárdia, L., & Hillier, M. (2016). Using e-Assessment to enhance student learning and evidence learning outcomes. *Int J Educ Technol High Educ*, 13-18.

Erostarbe, I. I., & Albonigamayor, J. J. (2007). Auto-evaluación a través de Internet: variables metacognitivas y rendimiento académico. *Revista Latinoamericana de Tecnología Educativa*, 59-76.

Fracchia, C., & Roger, S. (2003). El lenguaje natural en plataformas de educación a distancia. *Red UNCI*, 90-97.

Galli, A., Roiter, H., De Mollein, D., Swieszkowski, S., Atamañuk, N., Guerrero, A. A., . . . Barrero, C. (2011). Evaluación de la calidad de las preguntas de selección múltiple utilizadas en los exámenes de Certificación y Recertificación en Cardiología en el año 2009. *Revista Argentina de Cardiología*, 419-422.

Grondona, N., Mazza, M., & Dorfman, P. (2012). Asistentes virtuales de clase en la Educación Universitaria. *Una visió crític: III Congrés Europeu de Tecnológies de la Informació en l'Educació i en la Societat*, 294-296.

Haladyna, T. M., Haladyna, R., & Soto, C. M. (2002). Preparación de preguntas de opcion múltiples para medir el aprendizaje de los estudiantes. *OEI-Revista Iberoamericana de Educación*, 1-17.

Hernández, J. M. (2016). Análisis automático de textos en español utilizando NLTK.

Martínez, G. S., Martínez, R. A., & Aguilar, C. A. (2006). Extracción automática de contextos definitorios en textos especializados. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 351-352.

Menvielle, M. A., Groppo, M. A., Marciszack, M. M., & Analía Guzmán, K. L. (2016). Detección de conceptos y relaciones para evaluación de respuestas. *Facultad Regional Córdoba - Universidad Tecnológica Nacional*, 1-8.

Menvielle, M. A., Groppo, M. A., Marciszack, M. M., Guzmán, A., Ligorria, K., & Cassatti, M. (2018). Arquitectura y Operatoria. *Arquitectura y operatoria de un sistema de corrección de exámenes automatizado, utilizando grafos dirigidos*, 29-44.

Moreno, A. (17 de octubre de 2017). *Instituto de Ingeniería del Conocimiento*. Obtenido de https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/

Rivera Jiménez, J., Flores Hernández, F., Alpuche Hernández, A., & Martínez González, A. (2017). Evaluación de reactivos de opción múltiple en medicina. Evidencia de validez de un instrumento. *Investigación en Educación Médica, Vol. 6, núm. 21*, 8-15.

Salazar Blanco, O. F. (2015). Evaluación conocimientos con exámenes de selección múltiple: ¿tres o cuato opcioooooo de respuesta? Exxperiencia con el amen de admisión a posgrados médico-uirúrgicos en la Universidad de Antioquia. *Iatreia*.

Samuel Rocha, N. C.-S. (2016). Identificacion de las relaciones de hiperonimia en contextos definitorios por medio de un autómata determinista. *Pontificia Universidad Católica*.

Sierra, G. (2009). Extracción de contextos definitorios en textos de especialidad a partir del conocimiento de patrones lingüísticos. *LinguaMÁTICA*, 13-38.

Zue, V., & Glass, J. (2000). Conversational interfaces: Advance and challenges. *Proceedings of the IEEE 88(8)*, 1166-1180.