

Aplicación de modelos extensos de lenguaje en la ambientación narrativa

Mario Gómez Alonso

Carrera de Especialización en Inteligencia Artificial

Director: Bach. Josselyn Sofía Ordóñez Olazábal

Jurados:

Jurado 1 (pertenencia) Jurado 2 (pertenencia) Jurado 3 (pertenencia)

Ciudad de Madrid, España, Junio de 2025

Resumen

En la presente memoria se describe el diseño e implementación de un prototipo que emplea modelos extensos de lenguaje para la asistencia de los usuarios en la generación de contenido narrativo. El trabajo se realizó para la aplicación de Critical Match en su desarrollo de una de sus ramas de ampliación de servicios para sus usuarios.

Para su implementación fueron imprescindibles los conocimientos adquiridos en la especialización tanto en su contenido teórico como en los procedimientos para la creación y programación de proyectos de inteligencia artificial, entre las que se destacan asignaturas como procesamiento de lenguaje natural y modelos extensos de lenguaje.

Agradecimientos

Esta sección es para agradecimientos personales y es totalmente **OPCIONAL**.

Índice general

Ke	sum	en en	I
1.	Intr	oducción general	1
	1.1.	El desafío de la creación narrativa	1
	1.2.	Motivación	2
	1.3.		3
	1.4.	Objetivos y alcance	5
2.	Intr	oducción específica	7
	2.1.	Estilo y convenciones	7
		2.1.1. Uso de mayúscula inicial para los título de secciones	7
		2.1.2. Este es el título de una subsección	7
		2.1.3. Figuras	8
		2.1.4. Tablas	9
		2.1.5. Ecuaciones	10
3.	Dise	eño e implementación	13
	3.1.	Análisis del software	13
4.	Ensa	ayos y resultados	15
	4.1.	Pruebas funcionales del hardware	15
5.		clusiones	17
	5.1.	Conclusiones generales	17
	5.2.	Próximos pasos	17
Ri	hliog	rrafía	10

Índice de figuras

1.1.	Esquema de elementos habituales en la construcción de mundos	2
2.1.	Ilustración del cuadrado azul que se eligió para el diseño del logo.	8
2.2.	Imagen tomada de la página oficial del procesador ¹	9
2.3.	¿Por qué de pronto aparece esta figura?	9
2.4.	Tres gráficos simples	ç

Índice de tablas

2.1.	caption co	orto														1	n
									•								

Dedicado a... [OPCIONAL]

Introducción general

En este capítulo se presenta el contexto general y la motivación del trabajo realizado. A su vez se realiza un análisis del estado del arte relacionado con las tecnologías que se utilizaron y se concretan los objetivos y el alcance definidos durante su realización.

1.1. El desafío de la creación narrativa

La narrativa es un género literario que relata un conjunto de sucesos protagonizados por uno o más personajes y que son presentados a través de un narrador. Se trata de una forma de expresión cultural fundamental, presente en todas las culturas y épocas, utilizado para entretener y transmitir conocimiento. Una de sus características esenciales es la presencia de elementos ficticios, ya sea de manera parcial o total, con la única excepción del subgénero de crónicas que se limita a relatar hechos reales. La narrativa sigue siendo un elemento central de la cultura y está presente en múltiples formatos como la literatura, el cine, la televisión y otros medios tanto físicos como digitales.

La inclusión de elementos ficticios en la narrativa es un proceso creativo en el que el autor recurre a su imaginación para concebir componentes que, aunque inventados, resulten verosímiles y significativos. En este ejercicio mental intervienen múltiples factores, como la coherencia interna y la relación entre los distintos elementos que conforman la historia en el espacio y el tiempo. Estas variables se vuelven especialmente complejas en narrativas con una alta carga ficticia donde se inventan, además de los personajes y sucesos, contextos completos. Aquí es donde cobra especial relevancia la construcción de mundos, un recurso narrativo fundamental que permite al escritor diseñar entornos imaginarios detallados y capaces de sostener la lógica del relato.

El proceso de creación de mundos es tan complejo como la narrativa que busca sustentar y la intención del autor de sumergir al lector dentro de la historia. En la figura 1.1 se presenta una lista amplia, aunque no completa, de componentes fundamentales en la construcción de mundos y la manera en que estos se relacionan entre sí para dar cohesión al conjunto narrativo.

Históricamente los autores han recurrido a diversas herramientas para llevar a cabo la construcción de mundos, tales como esquemas, mapas, cronologías, fichas de personajes y la creación de lenguas artificiales. ¹ Todos estos elementos

¹Un ejemplo popular es el de J.R.R. Tolkien, quien creó mapas, genealogías y lenguas para su obra *El Señor de los Anillos* [1].

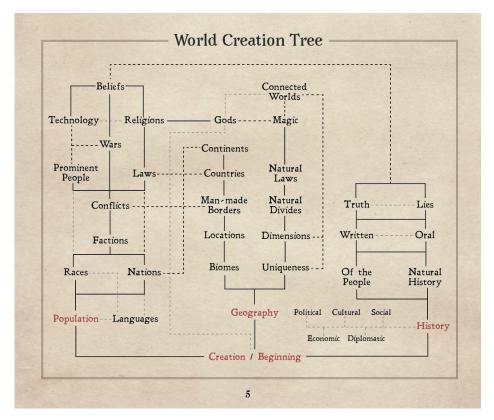


FIGURA 1.1. Esquema de elementos habituales en la construcción de mundos.

han sido producto del trabajo creativo del autor, apoyado tanto en su imaginación como en la consulta de múltiples referencias. Este proceso ha requerido una considerable inversión de tiempo y, en muchas ocasiones, afectado por bloqueos creativos a la hora de articular y cohesionar todos los componentes ficticios de la narrativa.

En los últimos años la aparición de la inteligencia artificial generativa ha abierto nuevas posibilidades dentro de los entornos creativos. Aunque originalmente no fue concebida con fines narrativos, su capacidad para producir texto coherente la convierte en una herramienta con un alto potencial para incorporarse en la construcción de mundos ficticios. Sin pretender sustituir la creatividad del autor, estas tecnologías pueden ejercer un papel de apoyo al facilitar el desarrollo de contenidos y aportar ideas que estimulen la creación literaria.

1.2. Motivación

El proyecto y consecuente trabajo se enfocan en la creación de entornos narrativos para juegos de rol y surgió de la Colaboración con la empresa Critical Match y su objetivo de ampliar las herramientas y servicios para sus usuarios de su aplicación de móvil. La *app* permite la creación y búsqueda de salas de juego que se ajusten a las necesidades y preferencias de cada usuario. En ese contexto, se busca mejorar la ambientación de las partidas proporcionando al autor de la narrativa a una serie de servicios para complementar la construcción de mundo existente.

El trabajo aborda la necesidad de los usuarios de conseguir partidas más inmersivas y reducir el trabajo de ambientación del dueño de la sala de juego. Reconociendo que no todos los jugadores tienen experiencia con la creación narrativa, esta herramienta facilita la generación de una ambientación atractiva. De este modo se optimiza el tiempo de preparación del director de juego, permitiéndole centrarse en otros aspectos de la partida. Esto también beneficia a los más experimentados al expandir sus ideas con mayor rapidez y eficacia.

Los servicios implementados en el trabajo pretenden ofrecer una experiencia claramente diferenciada de las alternativas más populares en internet ². Para ello, se proporcionan a los usuarios el acceso a consultas especializadas que solo requieren el envío de un fichero con la información narrativa, pudiendo combinarse adicionalmente con indicaciones que pueden escribirse en un cuadro de texto. Este enfoque no solo evita la incomodidad de interactuar con el modelo a través del navegador de internet del móvil, sino que ofrece respuestas precisas a sus necesidades, las cuales requerirían interacciones complejas con las alternativas existentes.

Además, este desarrollo se complementa a herramientas existentes de creación de mundos ³. Mientras que estas plataformas brindan estructura y el marco para la creación de mundos, el trabajo ofrece la capacidad de generar dinamicamente contenido narrativo específico con la información ya existente. De este modo los autores a través de ambas herramientas pueden centrarse en la visión general y cohesión del mundo mientras que la inteligencia artificial los ayuda con ideas con las que expandir la ambientación, fomentando un proceso cíclico creativo.

1.3. Estado del Arte

El avance de la inteligencia artificial generativa ha seguido un ritmo extraordinario desde la publicación del influyente artículo *Attention is All you need* [2] en 2017. La arquitectura de los transformadores revolucionó el procesamiento de lenguaje en las máquinas y permitió paralelizar los procesos a través del mecanismo de atención. Esto aumenta drásticamente la velocidad y eficacia en el proceso de aprendizaje de los modelos.

En el transcurso del año siguiente surgieron los primeros modelos extensos de lenguaje que adoptaron la arquitectura de los transformadores. Con el lanzamiento de GPT por OpenAI y de BERT por Google, ambos modelos se convirtieron en los pioneros y pilares fundamentales de los modelos extensos de lenguaje. Mientras BERT se especializó en la comprensión del contexto del texto de manera bidireccional, GPT fue desarrollado para la generación de texto.

Desde entonces la evolución de los modelos no se ha detenido. El número de parámetros ha crecido exponencialmente, lo que les ha otorgado una capacidad sin precedentes para comprender y producir texto coherente y fluido similar al humano. Además, este desarrollo ha sido transversal y está dotando a los modelos más recientes la habilidad para generar no solo texto, sino también imágenes, audio e incluso video.

²Algunos ejemplos son ChatGPT de OpenAI, Gemini de Google o MetaAI.

³Por ejemplo: WorldAnvil, Obsidian Portal o Legend Keeper.

Estos avances están actualmente al alcance del publico general a través de internet, con ejemplos destacados como ChatGPT, Gemini o Meta AI. ChatGPT fue el pionero y el que abarcó mayor popularidad, con una interfaz gráfica similar a la de un chat con la que el usuario podía interactuar de forma intuitiva y sencilla con el modelo. La utilización de estos modelos se masificó rápidamente y se implementaron variantes en multitud de sectores, principalmente en forma de asistentes virtuales.

En el ámbito creativo, el potencial de la inteligencia artificial generativa es tan amplio como discutido. Dependiendo de su aplicación, los modelos pueden ser catalizadores del proceso creativo al ayudar a superar bloqueos y explorar nuevas ideas con rapidez o bien reemplazar por completo la labor del artista. Este dilema plantea importantes debates sobre la autoría, los derechos de propiedad intelectual y el futuro de diversas profesiones creativas. Actualmente, la relación entre la inteligencia artificial y los artistas se encuentra en un proceso de redefinición de su paradigma.

A pesar de sus impresionantes capacidades, los modelos generativos se enfrentan a retos muy significativos que se enumeran a continuación:

- Sensibilidad a los prompts: Los modelos generativos son muy dependientes de los datos de entrada. Pequeñas variaciones en la información inicial pueden llevar a cambios drásticos en el resultado generado, lo que afecta la precisión de las respuestas y hace que sean menos fiables para usuarios sin experiencia previa.
- Alucinaciones e inconsistencias a largo plazo: Persisten las dificultades para mantener la coherencia y consistencia en textos extensos debido a la capacidad finita de su memoria. (ventana de contexto) También existe la posibilidad de que la salida se vuelva incongruente o generar información ficticia debido a factores internos o a la complejidad de la solicitud.
- Sesgos en los datos de entrenamiento: Los modelos pueden perpetuar o amplificar prejuicios sociales presentes en los datos con los que fueron entrenados.

Para mitigar estos problemas se están utilizando una serie de tecnicas:

- Generación aumentada por recuperación: Esta técnica permite a los modelos consultar fuentes de datos externas y ampliar la información de la entrada con elementos que están fuera de su entrenamiento inicial o de su ventana de contexto inmediata. Esto reduce las alucinaciones y mejora la verosimilitud de la respuesta.
- Filtrado de datos de entrenamiento: Se implementan procesos más rigurosos para limpiar y despolarizar los conjuntos de datos de entrenamiento, para minimizar la presencia de sesgos.
- *Fine-tuning*: Consiste en la adaptación de los modelos a tareas específicas a través de un proceso de reentrenamiento parcial.
- Prompt engineering: Se basa en la formulación precisa de instrucciones en la entrada para guiar la generación de la respuesta y obtener mejores resultados.

Este trabajo se fundamenta en este marco de conocimiento y se emplearon varias técnicas expuestas para mitigar el impacto de las limitaciones actuales de los modelos generativos.

1.4. Objetivos y alcance

El propósito del trabajo es el de desplegar un componente que aloje los modelos extensos de lenguaje con los que generar contenido narrativo a partir de una entrada. Este servidor también es capaz de administrar los modelos, añadir nuevos y elegir cuáles están activos.

A su vez, se dispone de un simple servidor web que actúa como la interfaz gráfica entre el usuario y la inteligencia artificial. Su función es recibir y mostrar la información al usuario, procesar sus instrucciones y el archivo adjunto y transformarlo en una instrucción personalizada según el servicio de generación narrativa solicitado. Esta instrucción será enviada al modelo de inteligencia artificial generativa, y su respuesta se reenviará al usuario.

Ambos servidores conforman un prototipo de pruebas que se entregará al cliente, junto con este documento, con el propósito de servir como estudio inicial y como base para futuros desarrollos dentro de su aplicación.

El alcance del trabajo incluye los siguientes elementos:

- Desarrollo de un servidor web.
- Descripción de los modelos extensos de lenguaje utilizados.
- Definición, implementación y descripción del Prompt engineering para cada servicio.
- La integración de un conjunto esencial de servicios REST.

El alcance del trabajo no incluye:

- Diseño de la página que cumpla con los estándares más habituales.
- Securización del servidor web ni de ninguno de sus servicios REST.
- Rendimiento de los servicios que garanticen tiempos de respuesta cercanos a tiempo real.

Introducción específica

Todos los capítulos deben comenzar con un breve párrafo introductorio que indique cuál es el contenido que se encontrará al leerlo. La redacción sobre el contenido de la memoria debe hacerse en presente y todo lo referido al proyecto en pasado, siempre de modo impersonal.

2.1. Estilo y convenciones

2.1.1. Uso de mayúscula inicial para los título de secciones

Si en el texto se hace alusión a diferentes partes del trabajo referirse a ellas como capítulo, sección o subsección según corresponda. Por ejemplo: "En el capítulo 1 se explica tal cosa", o "En la sección 2.1 se presenta lo que sea", o "En la subsección 2.1.2 se discute otra cosa".

Cuando se quiere poner una lista tabulada, se hace así:

- Este es el primer elemento de la lista.
- Este es el segundo elemento de la lista.

Notar el uso de las mayúsculas y el punto al final de cada elemento.

Si se desea poner una lista numerada el formato es este:

- 1. Este es el primer elemento de la lista.
- 2. Este es el segundo elemento de la lista.

Notar el uso de las mayúsculas y el punto al final de cada elemento.

2.1.2. Este es el título de una subsección

Se recomienda no utilizar **texto en negritas** en ningún párrafo, ni tampoco texto <u>subrayado</u>. En cambio sí se debe utilizar *texto en itálicas* para palabras en un idioma extranjero, al menos la primera vez que aparecen en el texto. En el caso de palabras que estamos inventando se deben utilizar "comillas", así como también para citas textuales. Por ejemplo, un *digital filter* es una especie de "selector" que permite separar ciertos componentes armónicos en particular.

La escritura debe ser impersonal. Por ejemplo, no utilizar "el diseño del firmware lo hice de acuerdo con tal principio", sino "el firmware fue diseñado utilizando tal principio".

El trabajo es algo que al momento de escribir la memoria se supone que ya está concluido, entonces todo lo que se refiera a hacer el trabajo se narra en tiempo pasado, porque es algo que ya ocurrió. Por ejemplo, "se diseñó el firmware empleando la técnica de test driven development".

En cambio, la memoria es algo que está vivo cada vez que el lector la lee. Por eso transcurre siempre en tiempo presente, como por ejemplo:

"En el presente capítulo se da una visión global sobre las distintas pruebas realizadas y los resultados obtenidos. Se explica el modo en que fueron llevados a cabo los test unitarios y las pruebas del sistema".

Se recomienda no utilizar una sección de glosario sino colocar la descripción de las abreviaturas como parte del mismo cuerpo del texto. Por ejemplo, RTOS (*Real Time Operating System*, Sistema Operativo de Tiempo Real) o en caso de considerarlo apropiado mediante notas a pie de página.

Si se desea indicar alguna página web utilizar el siguiente formato de referencias bibliográficas, dónde las referencias se detallan en la sección de bibliografía de la memoria, utilizado el formato establecido por IEEE en [3]. Por ejemplo, "el presente trabajo se basa en la plataforma EDU-CIAA-NXP [4], la cual...".

2.1.3. Figuras

Al insertar figuras en la memoria se deben considerar determinadas pautas. Para empezar, usar siempre tipografía claramente legible. Luego, tener claro que **es incorrecto** escribir por ejemplo esto: "El diseño elegido es un cuadrado, como se ve en la siguiente figura:"



La forma correcta de utilizar una figura es con referencias cruzadas, por ejemplo: "Se eligió utilizar un cuadrado azul para el logo, como puede observarse en la figura 2.1".



FIGURA 2.1. Ilustración del cuadrado azul que se eligió para el diseño del logo.

El texto de las figuras debe estar siempre en español, excepto que se decida reproducir una figura original tomada de alguna referencia. En ese caso la referencia de la cual se tomó la figura debe ser indicada en el epígrafe de la figura e incluida como una nota al pie, como se ilustra en la figura 2.2.



FIGURA 2.2. Imagen tomada de la página oficial del procesador¹.

La figura y el epígrafe deben conformar una unidad cuyo significado principal pueda ser comprendido por el lector sin necesidad de leer el cuerpo central de la memoria. Para eso es necesario que el epígrafe sea todo lo detallado que corresponda y si en la figura se utilizan abreviaturas entonces aclarar su significado en el epígrafe o en la misma figura.



FIGURA 2.3. ¿Por qué de pronto aparece esta figura?

Nunca colocar una figura en el documento antes de hacer la primera referencia a ella, como se ilustra con la figura 2.3, porque sino el lector no comprenderá por qué de pronto aparece la figura en el documento, lo que distraerá su atención.

Otra posibilidad es utilizar el entorno *subfigure* para incluir más de una figura, como se puede ver en la figura 2.4. Notar que se pueden referenciar también las figuras internas individualmente de esta manera: 2.4a, 2.4b y 2.4c.



FIGURA 2.4. Tres gráficos simples.

El código para generar las imágenes se encuentra disponible para su reutilización en el archivo **Chapter2.tex**.

2.1.4. Tablas

Para las tablas utilizar el mismo formato que para las figuras, sólo que el epígrafe se debe colocar arriba de la tabla, como se ilustra en la tabla 2.1. Observar que sólo algunas filas van con líneas visibles y notar el uso de las negritas para los encabezados. La referencia se logra utilizando el comando \ref{<label>} donde label debe estar definida dentro del entorno de la tabla.

¹Imagen tomada de https://goo.gl/images/i7C70w

```
\begin{table}[h]
\centering
\caption[caption corto]{caption largo más descriptivo}
 \begin{tabular}{l c c}
  \toprule
  \textbf{Especie} & \textbf{Tamaño} & \textbf{Valor}\\
  \midrule
 Amphiprion Ocellaris & 10 cm
Hepatus Blue Tang & 15 cm
Zebrasoma Xanthurus & 12 cm
                                             & \$ 6.000 \\
                                             & \$ 7.000 \\
                                             & \$ 6.800 \\
  \bottomrule
  \hline
 \end{tabular}
 \label{tab:peces}
\end{table}
```

TABLA 2.1. caption largo más descriptivo.

Especie	Tamaño	Valor
Amphiprion Ocellaris	10 cm	\$ 6.000
Hepatus Blue Tang	15 cm	\$ 7.000
Zebrasoma Xanthurus	12 cm	\$ 6.800

En cada capítulo se debe reiniciar el número de conteo de las figuras y las tablas, por ejemplo, figura 2.1 o tabla 2.1, pero no se debe reiniciar el conteo en cada sección. Por suerte la plantilla se encarga de esto por nosotros.

2.1.5. Ecuaciones

Al insertar ecuaciones en la memoria dentro de un entorno *equation*, éstas se numeran en forma automática y se pueden referir al igual que como se hace con las figuras y tablas, por ejemplo ver la ecuación 2.1.

$$ds^{2} = c^{2}dt^{2} \left(\frac{d\sigma^{2}}{1 - k\sigma^{2}} + \sigma^{2} \left[d\theta^{2} + \sin^{2}\theta d\phi^{2} \right] \right)$$
 (2.1)

Es importante tener presente que si bien las ecuaciones pueden ser referidas por su número, también es correcto utilizar los dos puntos, como por ejemplo "la expresión matemática que describe este comportamiento es la siguiente:"

$$\frac{\hbar^2}{2m}\nabla^2\Psi + V(\mathbf{r})\Psi = -i\hbar\frac{\partial\Psi}{\partial t}$$
(2.2)

Para generar la ecuación 2.1 se utilizó el siguiente código:

```
\begin{equation}
\label{eq:metric}
ds^2 = c^2 dt^2 \left( \frac{d\sigma^2}{1-k\sigma^2} + \sigma^2\left[ d\theta^2 + \sin^2\theta d\phi^2 \right] \right)
\end{equation}
```

Y para la ecuación 2.2:

```
\begin{equation}
  \label{eq:schrodinger}
  \frac{\hbar^2}{2m}\nabla^2\Psi + V(\mathbf{r})\Psi =
  -i\hbar \frac{\partial\Psi}{\partial t}
  \end{equation}
```

Diseño e implementación

Todos los capítulos deben comenzar con un breve párrafo introductorio que indique cuál es el contenido que se encontrará al leerlo. La redacción sobre el contenido de la memoria debe hacerse en presente y todo lo referido al proyecto en pasado, siempre de modo impersonal.

3.1. Análisis del software

19

20 21

22 23 24

26

updateSensors();

updateAlarms();

controlActuators();

vTaskDelayUntil(&ticks, period);

La idea de esta sección es resaltar los problemas encontrados, los criterios utilizados y la justificación de las decisiones que se hayan tomado.

Se puede agregar código o pseudocódigo dentro de un entorno lstlisting con el siguiente código:

```
\begin{lstlisting}[caption= "un epígrafe descriptivo"]
  las líneas de código irían aquí...
 \end{lstlisting}
 A modo de ejemplo:
1 #define MAX_SENSOR_NUMBER 3
2 #define MAX_ALARM_NUMBER 6
3 #define MAX_ACTUATOR_NUMBER 6
5 uint32_t sensorValue[MAX_SENSOR_NUMBER];
6 FunctionalState alarmControl[MAX_ALARM_NUMBER]; //ENABLE or DISABLE
void vControl() {
   initGlobalVariables();
12
13
   period = 500 ms;
14
15
   while (1) {
16
17
     ticks = xTaskGetTickCount();
```

28 }

CÓDIGO 3.1. Pseudocódigo del lazo principal de control.

Ensayos y resultados

Todos los capítulos deben comenzar con un breve párrafo introductorio que indique cuál es el contenido que se encontrará al leerlo. La redacción sobre el contenido de la memoria debe hacerse en presente y todo lo referido al proyecto en pasado, siempre de modo impersonal.

4.1. Pruebas funcionales del hardware

La idea de esta sección es explicar cómo se hicieron los ensayos, qué resultados se obtuvieron y analizarlos.

Conclusiones

Todos los capítulos deben comenzar con un breve párrafo introductorio que indique cuál es el contenido que se encontrará al leerlo. La redacción sobre el contenido de la memoria debe hacerse en presente y todo lo referido al proyecto en pasado, siempre de modo impersonal.

5.1. Conclusiones generales

La idea de esta sección es resaltar cuáles son los principales aportes del trabajo realizado y cómo se podría continuar. Debe ser especialmente breve y concisa. Es buena idea usar un listado para enumerar los logros obtenidos.

En esta sección no se deben incluir ni tablas ni gráficos.

Algunas preguntas que pueden servir para completar este capítulo:

- ¿Cuál es el grado de cumplimiento de los requerimientos?
- ¿Cuán fielmente se puedo seguir la planificación original (cronograma incluido)?
- ¿Se manifestó algunos de los riesgos identificados en la planificación? ¿Fue efectivo el plan de mitigación? ¿Se debió aplicar alguna otra acción no contemplada previamente?
- Si se debieron hacer modificaciones a lo planificado ¿Cuáles fueron las causas y los efectos?
- ¿Qué técnicas resultaron útiles para el desarrollo del proyecto y cuáles no tanto?

5.2. Próximos pasos

Acá se indica cómo se podría continuar el trabajo más adelante.

Bibliografía

- [1] J.R.R. Tolkien. *The Letters of J.R.R. Tolkien*. Houghton Mifflin, 1981.
- [2] Ashish Vaswani et al. «Attention is all you need». En: (). arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.
- [3] IEEE. *IEEE Citation Reference*. 1.ª ed. IEEE Publications, 2016. URL: http://www.ieee.org/documents/ieeecitationref.pdf (visitado 26-09-2016).
- [4] Proyecto CIAA. *Computadora Industrial Abierta Argentina*. Visitado el 2016-06-25. 2014. URL: http://proyecto-ciaa.com.ar/devwiki/doku.php?id=start.