Descripción general del descubrimiento del conocimiento en eSalud Reto en IberLEF 2020

Alejandro Piad-Morffisa , Yoan Gutiérrezb,c , Hian Cañizares-Diaza , Suilan Estévez-Velardea , Rafael Muñozb,c , Andrés Montoyob,c y Yudivian Almeida Cruz

a Facultad de Matemáticas e Informática, Universidad de La Habana, La Habana, 10400, Cuba

bDepartamento de Lenguaje y Sistemas Informáticos, Universidad de Alicante, Alicante, 03690, España bInstituto Universitario de Investigaciones en Computación, Universidad de Alicante, Alicante, 03690, España

Resumen

Este artículo resume los resultados de la tercera edición del desafío eHealth Knowledge Discovery (KD), organizado en el Foro Ibérico de Evaluación de Lenguas 2020. El desafío eHealth-KD propone dos tareas computacionales que involucran la identificación de entidades y relaciones semánticas en texto en lenguaje natural. , centrándose en los documentos sanitarios en lengua española. En esta edición, además del texto extraído de fuentes médicas, se introdujo en el corpus contenido de Wikipedia y se diseñó un novedoso escenario de evaluación de transferencia de aprendizaje que desafía a los participantes a crear sistemas que proporcionen una generalización entre dominios. Un total de ocho equipos participaron con una variedad de enfoques que incluyen sistemas de aprendizaje profundo de un extremo a otro, así como técnicas basadas en reglas e impulsadas por el conocimiento. Este artículo analiza los enfoques más exitosos y destaca los desafíos más interesantes para futuras investigaciones en este campo.

Palabras clave

eSalud, Descubrimiento de Conocimiento, Procesamiento del Lenguaje Natural, Aprendizaje Automático

1. Introducción

La gran cantidad de texto clínico disponible en línea ha motivado el desarrollo de sistemas automáticos de descubrimiento de conocimiento que pueden analizar estos datos y descubrir hechos relevantes. Estos descubrimientos pueden ser la base para nuevos tratamientos, la comprensión de las enfermedades y las interacciones entre medicamentos. Los sistemas computacionales diseñados para esta tarea a menudo se entrenan en corpus anotados manualmente. Para fomentar la investigación en esta área, la comunidad ha organizado desafíos competitivos para identificar, clasificar, extraer y vincular conocimientos, como en SEMEVAL[1] y campañas CLEF[2].

Actas del Foro de Evaluación de Lenguas Ibéricas (IberLEF 2020) correo electrónico: apiad@matcom.uh.cu (A. Piad-Morphis); ygutierrez@dlsi.ua.es (Y. Gutiérrez); this.canizares@matcom.uh.cu (H. Cañizares-Díaz); sestevez@matcom.uh.cu (S. Estévez-Velarde); rafael@dlsi.ua.es (R. Muñoz); montoyo@dlsi.ua.es (A. Montoyo); yudy@matcom.uh.cu (Y. Almeida-Cruz) Orcid: 0000-0001-9522-3 (A. Piad-Morphis); 0000-0002-4052-7 (Y. Gutiérrez); 0000-0002-5334-7468 (H. Cañizares-Diaz); 0000-0001-6707-1442 (S. Estévez-Velarde); 0000-0001-8127-9012 (R. Muñoz); 0000-0002-3076-0890 (A. Montoyo); 0000-0002-2345-1387 (Y. Almeida-Cruz)

ISSN 1613-0073 Actas del taller CEUR (CEUR-WS.org) 1http://alt.qcri.org/semeval2020/ 2http://www.clef-initiative.eu/

El desafío eHealth Knowledge Discovery (eHealth-KD), en su tercera edición, aprovecha un modelo semántico de lenguaje humano que codifica las expresiones más comunes de conocimiento fáctico, a través de un conjunto de cuatro tipos de entidades de propósito general y trece relaciones semánticas entre ellas. El desafío propone el diseño de sistemas que puedan anotar automáticamente entidades y relaciones en texto clínico en idioma español. En esta nueva edición, también se considera un escenario de evaluación alternativo (no relacionado con el dominio de la salud), que desafía a los participantes a diseñar sistemas que puedan transferir con éxito sus representaciones semánticas internas del dominio de la salud a un nuevo dominio arbitrario con datos de entrenamiento considerablemente reducidos. El desafío se ha celebrado en el Foro Ibérico de Evaluación de Lenguas 2020 y contó con la participación de ocho equipos de investigadores de diferentes instituciones.

Este artículo presenta el diseño del desafío así como los datos y herramientas proporcionados a los participantes, y analiza los resultados obtenidos por cada equipo. El resto del documento está organizado de la siguiente manera: la Sección 2 proporciona una descripción detallada de las tareas definidas en el desafío eHealth-KD y los datos proporcionados para la capacitación y evaluación del sistema de descubrimiento de conocimiento, así como todas las métricas de evaluación relevantes. La sección 3 describe brevemente todas las soluciones que se presentaron al desafío e introduce un conjunto de características que permiten una comparación cualitativa entre ellas. La Sección 4 presenta los principales resultados del desafío, divididos en cuatro escenarios de evaluación, y analiza los enfoques más exitosos y prometedores implementados por cada equipo. Finalmente, la Sección 5 presenta las conclusiones de la investigación y recomendaciones para futuras ediciones.

2. Descripción del desafío

El desafío eHealth-KD implica la identificación de entidades y relaciones semánticas en texto en lenguaje natural. Aunque en ediciones anteriores la atención se ha centrado en el ámbito de la salud, la naturaleza de las entidades y relaciones extraídas son generales y pueden aplicarse a cualquier ámbito.
La Figura 1 muestra un ejemplo de tres oraciones con las entidades y relaciones relevantes anotadas.
Se proporciona una explicación detallada del modelo de anotación en Piad-Morffis et al. [1].

La evaluación del desafío consiste en presentar un conjunto de oraciones en lenguaje natural con anotaciones producidas automáticamente por un sistema de descubrimiento de conocimiento. Los participantes reciben un conjunto de oraciones anotadas manualmente (corpus de capacitación y desarrollo) que pueden usarse para entrenamiento y/o ajuste del sistema, así como oraciones sin procesar que se usan para la evaluación (corpus de prueba). El corpus de capacitación y desarrollo se proporcionó con dos meses de anticipación, pero el corpus de prueba se publicó solo dos semanas antes de la fecha de evaluación, para desalentar cualquier ajuste en los datos de la prueba. Aunque no se requiere el código fuente real del sistema , se anima a los participantes a cargar su código en servicios abiertos para compartir código fuente como Github.

Para simplificar la evaluación y proporcionar comparaciones más detalladas, la tarea se divide en dos subtareas: una relacionada con la identificación y clasificación de entidades, y la otra relacionada con la extracción de las relaciones semánticas entre estas entidades.

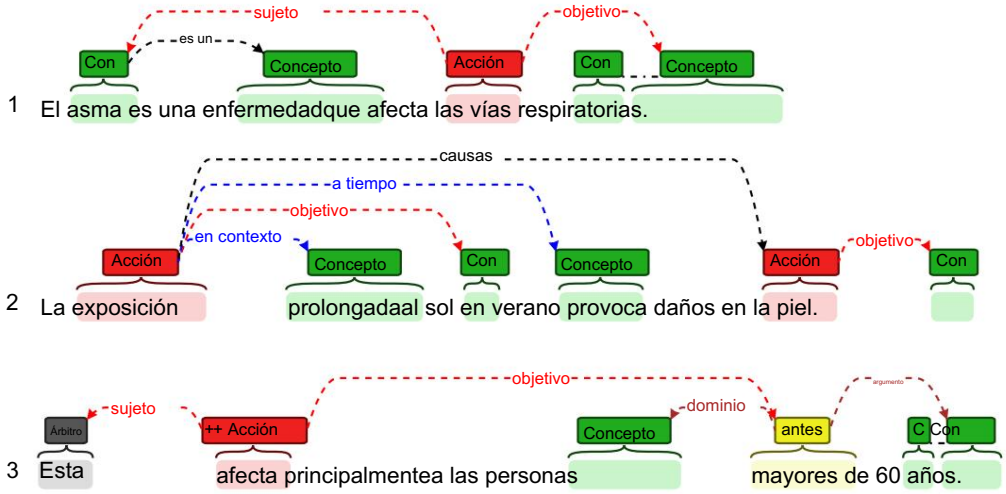

Figura 1: Anotación de ejemplo de tres oraciones del desafío eHealth-KD.

2.1. Subtarea A: Reconocimiento de entidades

Dada una lista de documentos de eSalud escritos en español, el objetivo de esta subtarea es identificar todos las entidades por documento y sus tipos. Estas entidades son todos los términos relevantes (una sola palabra o varias palabras) que representan elementos semánticamente importantes en una oración. La siguiente La figura muestra las entidades relevantes que aparecen en un conjunto de oraciones de ejemplo.

Some entities ("vías respiratorias" and "60 años") span more than one word. Entities will siempre consistirá en una o más palabras completas (es decir, no un prefijo o un sufijo de una palabra), y nunca incluya símbolos de puntuación, paréntesis, etc. circundantes. Hay cuatro tipos de entidades:

Concepto: identifica un término, concepto o idea relevante en el dominio de conocimiento de la oración.

Acción: identifica un proceso o modificación de otras entidades. Puede indicarse mediante un verbo. o construcción verbal, como "afectos", pero también por sustantivos, como "exposición". (exposición), donde denota el acto de exponerse al sol, y "daños" , donde denota el acto de dañar la piel. También se puede utilizar para indicar no verbal. relaciones funcionales, como "padre" , etc.

Predicado: identifica una función o filtro de otro conjunto de elementos, que tiene una función semántica. etiqueta en el texto, como "mayores" (mayores), y se aplica a una entidad, como "personas" (personas) con algunos argumentos adicionales como "60 años" .

Referencia: identifica un elemento textual que hace referencia a una entidad —de la misma oración o de diferente—, que puede indicarse mediante pistas textuales como "esta", "aquel", etc.

2.2. Subtarea B: Extracción de relaciones

La subtarea B continúa desde la salida de la subtarea A, vinculando las entidades detectadas y etiquetadas en el documento de entrada. El propósito de esta subtarea es reconocer todas las relaciones semánticas relevantes entre las entidades reconocidas. Ocho de las trece relaciones semánticas definidas para este desafío se pueden identificar en la Figura 1. Las relaciones semánticas se dividen en las siguientes categorías:

Relaciones generales (6): relaciones de propósito general entre dos conceptos (involucra Concepto, Acción, Predicado y Referencia) que tienen una semántica específica. Cuando se aplica cualquiera de estas relaciones, se prefiere a una relación de dominio –etiquetar una frase clave como un vínculo entre dos unidades de información–, ya que su semántica es independiente de cualquier etiqueta textual:

is-a: indica que una entidad es un subtipo, instancia o miembro de la clase identificada por el otro.

igual que: indica que dos entidades son semánticamente iguales. tiene

propiedad: indica que una entidad tiene una propiedad o característica determinada. parte

de: indica que una entidad es parte constituyente de otra. causas: indica

que una entidad provoca la existencia o aparición de otra. implica: indica que la existencia de

una entidad implica la existencia o ocurrencia de otro.

Relaciones contextuales (3): permiten refinar una entidad (implica Concepto, Acción, Predicado, y Referencia) adjuntando modificadores. Estos son:

en el tiempo: para indicar que algo existe, ocurre o está confinado a un período de tiempo, como en “exposición” en - tiempo “verano”. in-

place: para indicar que algo existe, ocurre o está confinado a un lugar o ubicación. in-context: para

indicar un contexto general en el que algo sucede, como un modo, manera o estado, como “exposición” en - contexto “prolongada”.

Roles de acción (2): indique qué papel juegan las entidades relacionadas con una Acción:

Asunto: indica quién realiza la acción, como en “[el] asma afecta […]”. objetivo: indica

quién recibe el efecto de la acción, como en “[…] afecta [las] vías respiratorias”. Las acciones pueden tener varios sujetos y objetivos, en cuyo caso la semántica interpretada es que la unión de los sujetos realiza la acción sobre cada uno de los objetivos.

Roles de predicado (2): indique qué rol juegan las entidades relacionadas con un Predicado:

dominio: indica la entidad principal sobre la que se aplica el predicado. arg:

indica una entidad adicional que especifica un valor para que el predicado tenga sentido. La semántica exacta de este argumento depende de la semántica de la etiqueta del predicado, como en “mayores [de] 60 años”, donde la etiqueta del predicado “mayores” indica que “60 años” es una cantidad que restringe la edad mínima para el predicado es verdadero.

2.3. Escenarios de evaluación

El Desafío eHealth-KD 2020 propone cuatro escenarios de evaluación para medir diferentes características de los sistemas participantes. Proponemos utilizar un 1 micropromediado que pondere todas las anotaciones individuales por igual, tanto entidades como relaciones. El Escenario 1 evalúa la solución de ambas tareas simultáneamente, mientras que los Escenario 2 y 3 evalúan cada tarea de forma independiente. Finalmente, el Escenario 4 desafía a los sistemas a un dominio novedoso con significativamente menos datos de entrenamiento. Esto permite una comparación más detallada entre sistemas con respecto a capacidades específicas.

2.3.1. Evaluación Principal (Escenario 1)

Este escenario evalúa ambas subtareas juntas como una canalización. La entrada consta únicamente de texto sin formato y la salida esperada es BRAT . ann archivo con todas las entidades y relaciones correspondientes encontradas.

Las medidas serán precisión, recuperación y F1 de la siguiente manera:

$$= \frac{+ + \frac{1}{2}}{+ + + + +}$$

$$= \frac{+ + \frac{1}{2}}{+ + + + +}$$

$$1 = 2 \frac{}{+}$$

La definición exacta de Correcto(C), Faltante(M), Espurio(S), Parcial(P) e Incorrecto(I) se presenta en las siguientes secciones para cada subtarea.

2.3.2. Subtarea opcional A (Escenario 2)

Este escenario solo evalúa la subtarea A. La entrada es un texto sin formato con varias oraciones y la salida es un archivo BRAT .ann con solo anotaciones de entidad (las anotaciones de relación se ignoran si están presentes).

Para calcular las puntuaciones definimos coincidencias correctas, parciales, faltantes, incorrectas y espurias.

No es necesario que los archivos de salida esperados y reales coincidan en el ID de cada entidad ni en su orden. Las coincidencias de evaluación se basan en el inicio y el final de los tramos de texto y el tipo correspondiente. A continuación se ofrece una breve descripción de las métricas:

Se informan coincidencias correctas cuando un texto en el archivo de desarrollo —DEV— coincide exactamente con un intervalo de texto correspondiente en el archivo dorado para los valores INICIO y FINAL, y también el tipo de entidad. Sólo se puede encontrar una coincidencia correcta por entrada en el archivo Gold. Por lo tanto, las entradas duplicadas contarán como espurias.

Se informan coincidencias incorrectas cuando los valores INICIO y FIN coinciden, pero no el tipo.

Se informan coincidencias parciales cuando dos intervalos [INICIO, FIN] tienen una intersección no vacía , como en el caso de “vías respiratorias” y “respiratorias” en el ejemplo anterior (y ETIQUETA coincidente). Tenga en cuenta que una frase parcial solo se comparará con una única frase correcta. Por ejemplo, “tipo de cáncer” podría ser una coincidencia parcial tanto para “tipo” como para “cáncer”, pero solo se cuenta una vez como coincidencia parcial con la palabra “tipo”. La palabra “cáncer” se cuenta entonces como Desaparecida. Esto tiene como objetivo disuadir que algunos textos grandes que cubren la mayor parte del documento obtengan una puntuación muy alta.

Las coincidencias que faltan son aquellas que aparecen en el archivo GOLD pero no en el archivo DEV.

Las coincidencias espurias son aquellas que aparecen en el archivo DEV pero no en el archivo Gold.

A partir de estas definiciones, calculamos la precisión, la recuperación y una medida F1 estándar de la siguiente manera:

$$= \frac{+\frac{1}{2}}{+++}$$

$$= \frac{+\frac{1}{2}}{+++}$$

$$_1 = 2\ \frac{}{+}$$

2.3.3. Subtarea opcional B (Escenario 3)

Este escenario solo evalúa la subtarea B. La entrada es texto sin formato y un archivo .ann correspondiente con las entidades correctas anotadas. El resultado esperado es un archivo .ann con entidades y relaciones. Para que esto suceda, las anotaciones de entidad del archivo .ann proporcionado se pueden copiar con las anotaciones de relación adjuntas.

Para calcular las puntuaciones definimos coincidencias correctas, faltantes y espurias. No es necesario que los archivos de salida esperados y reales coincidan en el ID de cada relación (que se ignora) ni en su orden. Las coincidencias de evaluación se basan en el inicio y el final de los tramos de texto y el tipo correspondiente. A continuación se ofrece una breve descripción de las métricas:

Correcto: relaciones que coincidían exactamente con el archivo GOLD, incluido el tipo y los ID correspondientes de cada uno de los participantes.

Faltan: relaciones que están en el archivo GOLD pero no en el archivo DEV, ya sea porque el tipo es incorrecto o porque uno de los ID no coincide.

Espurias: relaciones que están en el archivo DEV pero no en el archivo gold, ya sea porque el tipo es incorrecto o porque uno de los ID no coincide.

Definimos las métricas estándar de precisión, recuperación y F1 de la siguiente manera:

$$= \frac{}{+}$$

$$= \frac{}{+}$$

$$_1 = 2 \frac{}{+}$$

2.3.4. Evaluación de dominio alternativo opcional (escenario 4)

Este escenario evalúa un conjunto de 100 oraciones de un dominio alternativo (no relacionado con la salud) para experimentar con técnicas de aprendizaje por transferencia. Cuando se publique el conjunto de prueba general, se proporcionará un pequeño conjunto de datos de desarrollo con 100 oraciones y sus anotaciones correspondientes.

Los participantes deberán entrenar sus sistemas en el corpus completo de eHealth-KD 2020 y luego aplicar algunas técnicas de ajuste en las 100 oraciones adicionales del dominio alternativo para abordar con éxito este escenario. El formato de entrada y salida y las métricas de evaluación son los mismos que para el Escenario 1.

El propósito de este escenario, que consideramos un desafío complejo, es estimular el desarrollo de sistemas que puedan generalizarse a nuevos dominios de conocimiento sin demasiados ejemplos de capacitación adicionales. Por lo tanto, animamos a los participantes a centrarse no sólo en características y técnicas específicas de esalud, sino también a considerar enfoques más generalizables.

2.4. Descripción del cuerpo

El corpus utilizado en esta edición del desafío se compone de varias fuentes reutilizadas de desafíos anteriores, así como nuevo contenido comentado. Las pautas de anotación y el procedimiento seguido fueron los descritos en Piad-Morffis et al. [2].

Se reutilizan un total de 1000 frases de formación y desarrollo de la edición anterior del desafío, que se basa en el mismo modelo de anotación y metodología. Para el corpus de prueba, se anotaron manualmente un nuevo conjunto de 300 frases de Medline. Se seleccionaron 200 oraciones adicionales de Wikinews, de las cuales 100 se proporcionaron para desarrollo y 100 para prueba en el Escenario de evaluación 4. Finalmente, basándose en las presentaciones de la edición anterior, se construyó un conjunto de 3.000 oraciones anotadas automáticamente agregando las anotaciones realizadas por los participantes anteriores. Estas oraciones no se han revisado manualmente, por lo que se proporcionan como un recurso adicional para realizar ajustes, pero deben usarse con cuidado al entrenar un nuevo sistema. Las estadísticas generales del corpus se resumen en la Tabla 1.

3. Descripción de los sistemas

Esta sección describe brevemente los ocho sistemas que se presentaron al desafío. A diferencia de ediciones anteriores, hubo un alto grado de uniformidad entre los participantes, en el sentido de que la mayoría de los enfoques implican el uso de arquitecturas de aprendizaje profundo con incorporaciones contextuales o estáticas. Sin embargo, existen diferencias interesantes entre los enfoques que resultaron significativas con respecto a los resultados obtenidos. Los equipos participantes y sus correspondientes sistemas se describen a continuación:

tabla 1

Estadísticas resumidas del eHealth-KD Corpus v2.0. Las frases clave y las etiquetas de relación se clasifican por Número de instancias en el conjunto de entrenamiento. Las colecciones de formación y desarrollo (marcadas con) tienen han sido reutilizados de ediciones anteriores.

Métrico	Entrenamiento total DEV/DEV principal/Conjunto de prueba de transferencia					
Oraciones	3400	800	200	100	300	3000
Entidades	25225	5012	1305	1242	2921	14745
- Concepto	16207	3112	797	841	1944	9513
- Acción	6431	1319	340	278	628	3866
- Predicado	1902	412	124	104	299	963
- Referencia	685	169	44	19	50	403
Relaciones	20504	4571	1204	1241	2710	10778
- objetivo	6376	1281	350	270	562	3913
- tema - en	3156	674	170	251	438	1623
contexto	2503	502	140	193	380	1288
- es un	2013	458	104	119	262	1070
- en su lugar	1250	304	77	111	237	521
- causas	890	292	71	30	92	405
- dominio	994	269	74	82	196	373
- argumento -	857	254	73	47	185	298
implica -	308	117	43	11	28	109
en el tiempo	489	126	26	81	127	129
- tiene propiedad	1088	346	134	18	18	91
- igual que			93	31	19	66
- parte de	234	67	27	9	46	85

Vicomtech [3] presentó una red neuronal profunda de extremo a extremo con modelos BERT previamente entrenados como núcleo para la representación semántica de los textos de entrada. Ellos experimentaron con dos modelos: BERT-Base Multilingual Cased y BETO, un modelo BERT previamente entrenado en Texto en español. Modelan todas las variables de salida (entidades y relaciones) al mismo tiempo, modelar todo el problema de forma conjunta. Algunas de las salidas se retroalimentan a la última capa. del modelo, conectando los resultados de las diferentes subtareas en forma de canalización.

TALP-UPC [4] presentó una red neuronal profunda de extremo a extremo, para identificar simultáneamente frases clave y sus relaciones, que no se basan en ningún conocimiento específico del dominio ni rasgos artesanales. Los documentos de entrada se analizan usando FreeLing y se codifican usando ya sea un modelo de incrustación de palabras previamente entrenado BERT, Word2Vec o FastText. Con el fin de generar todas las relaciones posibles, el modelo debe ejecutarse para cada token de entrada y tener todas las probabilidades brutas combinadas en cada uno de ellos.

UH-MAJA-KD [5] presentó un modelo híbrido para la subtarea A que utiliza Stacked Bidireccional Capas LSTM como codificadores contextuales y campos aleatorios condicionales de cadena lineal como decodificadores de etiquetas. El sistema aborda la subtarea B mediante una consulta por pares, codificando información sobre la oración y el par de entidades dado usando estructuras sintácticas derivado del árbol de análisis de dependencia, por medio de Recurrent Neural basado en LSTM

Redes.

IXA-NER-RE [6] presentó un modelo de dos pasos para las subtareas NER y RE, cada uno de ellos desarrollado independientemente del otro. La tarea de reconocimiento de entidades de nombres se ha concebido como un sistema seq2seq básico que aplica un modelo de lenguaje de propósito general e incrustaciones estáticas. En la subtarea de RE, se exploraron dos enfoques: transferir métodos de aprendizaje y Matching the Blank para abordar el problema del tamaño reducido del corpus de entrenamiento al producir representaciones de relaciones directamente a partir de texto sin etiquetar.

UH-MatCom [7] presentó varios modelos de aprendizaje profundo entrenados y ensamblados para extraer automáticamente las entidades y relaciones. Sus modelos utilizan una combinación de técnicas de última generación como BERT, Bi-LSTM y CRF. También exploran el uso de fuentes de conocimiento externas como ConceptNet.

SINAI [8] presentó una red neuronal BiLSTM+CRF donde se combinan diferentes incrustaciones de palabras como entrada a la arquitectura: incrustaciones médicas generadas a medida, incrustaciones no médicas contextualizadas e incrustaciones no médicas previamente entrenadas basadas en transformadores.

HAPLAP [9] presentó una red neuronal AB-LSTM conjunta que combina un Bi-LSTM con agrupación máxima y un Bi-LSTM atento para la tarea de extracción de relaciones. El Joint AB-LSTM se alimenta con las oraciones preprocesadas, sus entidades y relaciones entre ellas, e incrustaciones a distancia.

ExSim [10] presentó un enfoque de recuperación de información en el que las entidades y relaciones en el conjunto de entrenamiento se comparan mediante similitud de incrustación de palabras para determinar la etiqueta más probable.

Baseline es una implementación básica que almacena todos los pares de entidades y etiquetas, y todos los tripletes de dos entidades y etiquetas de relación que se encuentran en el conjunto de entrenamiento, y simplemente genera para el conjunto de prueba una etiqueta si encuentra una coincidencia exacta. El propósito de la línea de base es brindar a los participantes un punto de partida que ya se encargue de cargar los datos, analizar el formato de anotación y producir el resultado correcto.

Con diferencia, el tipo de enfoque más común corresponde a arquitecturas recurrentes de aprendizaje profundo (por ejemplo, capas LSTM) con incrustaciones contextuales (por ejemplo, BERT). Esta combinación es la base de siete de los ocho sistemas participantes. Esto no es sorprendente dado el éxito reciente de estos enfoques en varias tareas de PNL y, de hecho, así se sugirió en la descripción general de ediciones anteriores del eHealth-KD Challenge [11] [12]. Las variaciones dentro de esta tendencia incluyen el uso de incorporaciones personalizadas en lugar de incorporaciones previamente entrenadas y la introducción de funciones basadas en el conocimiento.
Sin embargo, la diferencia de enfoque más significativa corresponde a sistemas que realizan una estrategia de extremo a extremo versus sistemas que resuelven cada subtarea por separado. En las dos ediciones anteriores del desafío, el sistema con mejor rendimiento ha utilizado una estrategia de extremo a extremo. En esta edición, dos equipos (Vicomtech y TALP) implementan diferentes estrategias de extremo a extremo.

3.1. Características de los sistemas

Para describir cada sistema definimos un conjunto de características que agrupan los diferentes enfoques utilizados por los participantes. Estas características abarcan desde conceptos abstractos como el uso de conocimiento externo hasta detalles de implementación como el uso de transformadores u otras incorporaciones contextuales. El propósito de estas características es analizar qué hay en común entre los sistemas que mejor funcionan en cada escenario y posiblemente identificar técnicas interesantes o inexploradas. Las características se describen a continuación.

PNL: uso de funciones y estrategias clásicas de procesamiento del lenguaje natural, como codificación TF-IDF , derivación, lematización, análisis de dependencias, etc.

Incrustaciones estáticas: uso de incrustaciones de palabras previamente entrenadas, como Word2Vec o Glove, entrenadas en corpus estándar.

Incrustaciones contextuales: uso de incrustaciones contextuales como BERT o GPT, entrenadas en corpus estándar.

Incrustaciones personalizadas: usar cualquier tipo de incrustación con un conjunto de datos personalizado seleccionado para esto tarea o un proceso de ajuste fino.

Red recurrente: utilizando cualquier variante de redes neuronales recurrentes, como GRU o LSTM, posiblemente combinado con otras arquitecturas de aprendizaje profundo.

Bases de conocimiento: utilizar cualquier fuente de conocimiento semántico externo para definir características o para enriquecer el conjunto de entrenamiento.

Extremo a extremo: Diseñar un sistema único que se entrene simultáneamente en ambas subtareas y comparta al menos una parte de las características, representación o parámetros de aprendizaje tanto para entidades como para relaciones.

4. Resultados

La Tabla 2 resume los resultados obtenidos por cada participante en cada escenario de evaluación. Los resultados se ordenan por 1 en el Escenario 1, que se considera la evaluación principal. Los tres mejores resultados en cada escenario están resaltados en negrita.

En general, el sistema con mejor rendimiento lo presentó Vicomtech [3] , que no sólo obtiene el mejor resultado en el Escenario 1 (por un margen significativo), sino que también se sitúa entre los tres primeros en todos los escenarios. Asimismo, el sistema propuesto por Talp-UPC [4] obtiene el mejor resultado en el Escenario 4, considerado el más difícil dado el escaso número de ejemplos de formación. También vale la pena mencionar los resultados obtenidos por UH-MAJA-KD, que también se ubica entre los mejores resultados en todos los escenarios, y la diferencia con el mejor resultado anterior es inferior a 0,001 en dos escenarios, lo que puede considerarse estadísticamente insignificante.

Finalmente, es interesante notar que los sistemas que obtuvieron los mejores resultados para cada tarea individual (es decir, SINAI en el Escenario 2 e IXA-NER-RE en el Escenario 3) no se ubican entre los tres primeros en los escenarios generales. Esto sugiere una interesante compensación entre centrarse en resolver una tarea específica o diseñar un sistema que funcione bien en general.

Tabla 2

Resultados (métrica) en cada escenario, ordenados por Escenario 1 (columna Puntuación). Los mejores resultados por escenario están resaltados en negrita.

Equipo	Puntuación (1) Scn 1	Scn 2	Scn 3	Scn 4	Características
Vicomtech	0,665	0,820	0,583	0,563	Red recurrente, integración contextual, de extremo a extremo
Talp-UPC	0,626	0,815	0,574	0,583	Red recurrente, incrustación contextual, incrustación estática, PNL, de extremo a extremo
UH-MAJA-KD	0,625	0,814	0,598	0,547	Red recurrente, incrustación contextual, PNL
IXA-NER-RE	0,557	0,691	0,633	0,478	Red recurrente, incrustación contextual, incrustación personalizada
UH-MatCom	0,556	0,794	0,545	0,420	0.373 Red recurrente, incrustación contextual, PNL, bases de conocimiento
SINAÍ		0,825	0,461	0,395	0.281 Red recurrente, incrustación contextual, incrustación personalizada, bases de conocimientos
FLIPBOARD		0,541	0,316	0,137	Red recurrente, incrustación contextual
exsim		0,245	0,314	0,131	0.122 PNL, incrustación estática

4.1. Análisis del desempeño de los sistemas.

De acuerdo con las características definidas en la Sección 3.1, realizamos un análisis cualitativo de las estrategias más exitosas en cada escenario. La figura 2 muestra un diagrama de caja del ranking obtenidos por sistemas con cada una de las características antes definidas, según escenario de evaluación. El diagrama de caja muestra la media, los rangos intercuartiles y la puntuación mínima y máxima. entre todos los sistemas con una determinada característica.

Como se observó, la estrategia común de utilizar incrustaciones contextuales y redes recurrentes es capaz de producir resultados en toda la gama de rankings. Sin embargo, varios sistemas tienen Implementamos y adaptamos esta estrategia, produciendo resultados con una variedad de variaciones. De ahí que el uso La combinación de capas BERT o LSTM por sí sola no garantiza una estrategia exitosa. Asimismo, como se observa en ediciones anteriores, el uso de incrustaciones personalizadas parece incurrir en una desventaja marginal, quizás dado que es difícil entrenar incorporaciones de alta calidad en corpus de dominios específicos. Por otro lado, el uso de bases de conocimiento internas para enriquecer las representaciones semánticas. parece ser útil en la subtarea de reconocimiento de entidades, como lo ejemplifica el resultado obtenido por el SINÁI [8]. El enfoque más exitoso parece ser el diseño de extremo a extremo. arquitecturas en lugar de resolver ambas subtareas por separado. Esta ha sido una tendencia en todos los ediciones del desafío eHealth-KD y es uno de los conocimientos más importantes. El hecho que los sistemas de extremo a extremo superen consistentemente a otros enfoques indica que existe una interesante interacción entre la representación semántica de entidades y relaciones. Ambos Los enfoques integrales presentados proporcionan una ventaja importante en términos de retroalimentación interna. intercambio al resolver la Subtarea A y la Subtarea B, mejorando el descubrimiento de entidades y relaciones. Este enfoque apoya la idea de que ambas subtareas no son completamente independientes de entre sí. Sin embargo, como se explica en la Sección 4, si bien los sistemas de extremo a extremo superan a todos los demás enfoques en los Escenario 1 y 4, donde se realizan ambas subtareas, hay subtareas específicas enfoques que funcionan mejor cuando sólo se evalúa una de las tareas.

5. Conclusiones y trabajo futuro

El eHealth-KD 2020 propuso –al igual que las ediciones anteriores eHealth-KD 2019[11] y eHealth-KD 2018[12]– la modelización del lenguaje humano en un escenario en el que la salud electrónica española Los documentos podrían ser legibles por máquina desde un punto de vista semántico. Con esta tarea, nosotros

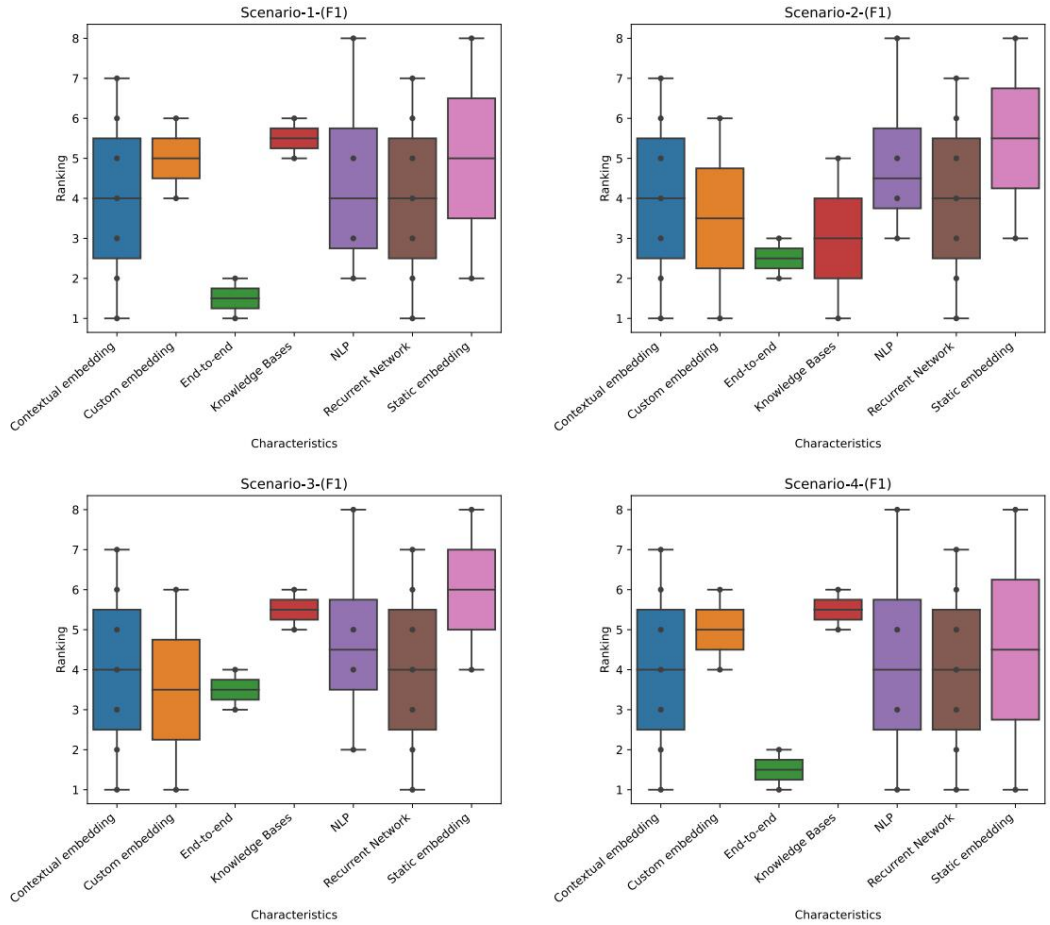

Figura 2: Diagrama de caja de la distribución de la clasificación para los sistemas que aplicaron cada uno de los enfoques definidos en la Sección 3.1

.

Se espera que fomente el desarrollo de tecnologías de software para extraer automáticamente una gran variedad de conocimientos de documentos de eSalud escritos en español. Para ello , se anotó manualmente un nuevo corpus en lengua española. Asimismo, proporcionamos herramientas para simplificar la construcción de sistemas de descubrimiento de conocimiento basados en este corpus.

En el desafío se presentaron ocho sistemas logrando una puntuación máxima F1 de 0,665. Todos los participantes presentaron algoritmos en todos los escenarios, obteniendo los mejores resultados los sistemas de extremo a extremo. El cambio significativo más utilizado en la edición de 2020 con respecto a las anteriores es el uso de incrustaciones contextuales (es decir, arquitecturas transformadoras y específicamente BERT) como reemplazo de incrustaciones de palabras estáticas. Los resultados indican que, aunque en el desafío se presentaron enfoques prometedores , la extracción de relaciones semánticas de propósito general a partir de textos en lenguaje natural sigue siendo un área de investigación abierta. Además, aunque los enfoques modernos de aprendizaje profundo son los más exitosos, creemos que todavía hay margen de mejo

incorporando componentes basados en el conocimiento que puedan explotar la estructura del modelo de anotación.

Expresiones de gratitud

This research has been partially supported by the University of Alicante and University of Havana, the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) and the Spanish Government through the projects SIIA (P R O M E T E O / 2 0 1 8 / 0 8 9, P R O M E T E U / 2 0 1 8 / 0 8 9) and LIVING-LANG (R T I 2 0 1 8 - 0 9 4 6 5 3 - B - C 2 2).

Referencias

[1] A. Piad-Morffis, Y. Guitérrez, S. Estevez-Velarde, R. Muñoz, Un modelo de anotación de propósito general para el descubrimiento de conocimiento: Estudio de caso en texto clínico español, en: Actas del 2º Procesamiento Clínico del Lenguaje Natural. Taller, 2019, págs. 79–88.

[2] A. Piad-Morffis, Y. Gutiérrez, Y. Almeida-Cruz, R. Muñoz, Un ecosistema computacional para apoyar tecnologías de descubrimiento de conocimiento en esalud en español, Journal of Biomedical Informatics (2020) 103517.

[3] A. García-Pablos, N. Perez, M. Cuadros, E. Zotova, Vicomtech en eHealth-KD Challenge 2020: Modelo profundo de extremo a extremo para la extracción de entidades y relaciones en textos médicos, en: Actas de la Foro Ibérico de Evaluación de Lenguas compartido con el 36º Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, IberLEF@SEPLN 2020, 2020.

[4] S. Medina, J. Turmo, TALP en eHealth-KD Challenge 2020: Redes neuronales convolucionales y recurrentes de múltiples niveles para la clasificación conjunta de frases clave y relaciones en: Actas del Foro de Evaluación de Lenguas Ibéricas ubicado conjuntamente con el 36º Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, IberLEF@SEPLN 2020, 2020.

[5] A. Rodríguez Pérez, E. Quevedo Knight, J. Mederos Alvarado, R. Cruz-Linares, JP Consuegra-Ayala, UH-MAJA-KD en eHealth-KD Challenge 2020: Modelos de aprendizaje profundo para el descubrimiento de conocimientos en documentos de eSalud en español, en: Actas del Foro de Evaluación de Lenguas Ibéricas ubicado junto con la 36ª Conferencia de la Sociedad Española de Procesamiento del Lenguaje Natural , IberLEF@SEPLN 2020, 2020.

[6] E. Andrés, O. Sainz, A. Atutxa, O. López de Lacalle, IXA-NER-RE en eHealth-KD Challenge 2020: Aprendizaje por transferencia translingüe para la extracción de relaciones médicas, en: Actas de las lenguas ibéricas Foro de Evaluación compartido con el 36º Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, IberLEF@SEPLN 2020, 2020.

[7] JP Consuegra-Ayala, M. Palomar, UH-MatCom en eHealth-KD Challenge 2020: Modelos de aprendizaje profundo y conjunto para el descubrimiento de conocimientos en documentos en español, en: Actas del Foro de Evaluación de Lenguas Ibéricas ubicado conjuntamente con la 36ª Conferencia de la Sociedad Española de Procesamiento del Lenguaje Natural, IberLEF@SEPLN 2020, 2020.

[8] López-Úbeda P, Perea-Ortega JM, DG. Manuel C., MT Martin-Valdivia, LA Ureña- López, SINAI en eHealth-KD Challenge 2020: Combinación de incrustaciones de palabras para el reconocimiento de entidades nombradas en registros médicos españoles, en: Proceedings of the Iberian Languages

Foro de Evaluación compartido con el 36º Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, IberLEF@SEPLN 2020, 2020.

[9] S. Santana, A. Pérez, A. Casillas, HapLap en eHealth-KD Challenge 2020, en: Actas del Foro de Evaluación de Lenguas Ibéricas co-ubicado con la 36ª Conferencia de la Sociedad Española de Procesamiento del Lenguaje Natural, IberLEF@SEPLN 2020, 2020.

[10] Z. Hamzah Almugbel, ExSim en eHealth-KD Challenge 2020, en: Actas del Foro de Evaluación de Lenguas Ibéricas ubicado junto con la 36ª Conferencia de la Sociedad Española para el Procesamiento del Lenguaje Natural, IberLEF@SEPLN 2020, 2020.

[11] Piad-Morffis A, Gutiérrez Y, Consuegra-Ayala JP, Estevez-Velarde S, Almeida-Cruz Y, Muñoz R, Montoyo A , en: Actas del Foro de Evaluación de Lenguas Ibéricas ubicado conjuntamente con la 35ª Conferencia de la Sociedad Española de Procesamiento del Lenguaje Natural, IberLEF@SEPLN 1–16. URL: http://ceur-ws.org/Vol-2421/ eHealth-KD_overview.pdf.

[12] Cámara EM, Almeida-Cruz Y, Díaz-Galiano MC, Estevez-Velarde S, Á. Ver artículo PubMed Google Scholar Cumbreras, MG Vega, Y, Gutiérrez, A, Montejo-Raez, A, Montoyo, R, Muñoz, A, Piad-Morffis, J, Villena-Román, Actas de TASS 2018: Taller sobre análisis semántico en SEPLN, TASS@SEPLN 2018, ubicado conjuntamente con la 34ª Conferencia SEPLN (SEPLN 2018), Sevilla, España, 18 de septiembre de 2018, 2018, págs. 10-11.
13–27. URL: http://ceur-ws.org/Vol-2172/p0_overview_tass2018.pdf.