TALP en eHealth-KD Challenge 2020: multinivel Redes neuronales recurrentes y convolucionales para Clasificación conjunta de frases clave y relaciones

Salvador Medinaa , Jordi Turmoa

aUniversidad Politécnica de Cataluña, Campus Nord, Calle de Jordi Girona, 1, 3, 08034 Barcelona, Spain

Resumen

Este artículo describe el modelo presentado por el Equipo TALP a la tarea compartida eHealth Knowledge Discovery 2020 de IberLEF[1]. El modelo repite la idea de utilizar un modelo único para identificar simultáneamente frases clave y sus relaciones. Teniendo en cuenta la nueva subtarea de transferencia y aprendizaje presentada para la edición 2020 de eHealthKD, nuestro modelo no se basa en ningún conocimiento específico de un dominio ni en características artesanales. Nuestro modelo fue competitivo en las cuatro subtareas, ubicándose en 2.ª, 3.ª, 4.ª y 1.ª posición respectivamente.

Palabras clave

NERC, extracción de relaciones, PNL de eSalud, incrustaciones contextuales

1. Introducción

Este artículo describe las opciones de diseño y la estrategia de formación detrás del modelo presentado por el equipo TALP para la tarea compartida eHealth Knowledge Discovery 2020 de IberLEF[1]. Esta tarea compartida consiste en identificar frases clave relevantes y las relaciones entre ellas en documentos de Salud Electrónica de Medline español. La edición de eHealthKD 2020 incluye dos adiciones importantes con respecto a las ediciones anteriores: un conjunto de datos conjunto creado combinando predicciones de los resultados del modelo de la edición anterior de 2019 cuando se aplican a un conjunto de datos sin etiquetar, y una nueva subtarea de aprendizaje por transferencia.

Nuestro modelo itera sobre el modelo 2019 de nuestro equipo[2] aprovechando varios modelos de representación de texto a nivel de palabra previamente entrenados, además de aprovechar el corpus etiquetado automáticamente en un paso previo al entrenamiento. En particular, utilizamos Word2Vec y FastText Medical Word Embedding para modelos en español[3] previamente entrenados del Centro de Supercomputación de Barcelona, que fueron entrenados utilizando la base de datos SciELO y un subconjunto de Wikipedia relacionado con la salud. Usamos estos dos modelos para agregar conocimiento específico del contexto a nuestro modelo, lo que creemos que fue una de las deficiencias del modelo reemplazado. Sin embargo, los resultados sugieren que el uso de estas representaciones de palabras no representa una mejora apreciable con respecto a las de propósito general para los Escenarios 1 a 3 e incluso puede ser perjudicial para el Escenario 4.

Actas del Foro Ibérico de Evaluación de Lenguas (IberLEF 2020) correo
electrónico: smedina@cs.upc.edu (S. Medina); turmo@cs.upc.edu (J. Turmo)
Orcid: 0000-0003-2473-8 (S. Medina); 0000-0002-7521-1 (J.Turmo)

ISSN 1613-0073 Actas del taller CEUR (CEUR-WS.org)

2. Descripción del sistema

Nuestro modelo espera un documento y un índice de token de origen como entrada y genera una secuencia de etiquetas para cada frase clave y clase de relación. Los documentos de entrada se analizan utilizando el analizador de dependencias de FreeLing y cada uno de sus tokens se codifica utilizando un modelo de incrustación de palabras previamente entrenado BERT, Word2Vec o FastText. Luego, el modelo aplica filtros de convolución a los tokens codificados de los documentos de entrada, combina las salidas del filtro a nivel de palabra de cada token de entrada y el token fuente especificado con incrustaciones a nivel de oración de los documentos y genera los límites de cada frase clave que contiene el token de origen, así como las probabilidades de que cualquier otro token sea el objetivo de una relación que tenga como fuente las frases clave del token de origen especificado.

Para generar todas las relaciones posibles, el modelo debe ejecutarse para cada token de entrada y combinar todas las probabilidades brutas en cada uno de ellos. Este enfoque de observar un único token de entrada a la vez está inspirado en modelos de traducción basados en la atención, como el Transformer, en el que el modelo genera el token de salida más probable, uno a la vez, condicionado a los tokens generados previamente y al documento completo sin traducir.

2.1. Estructura interna del modelo.

En la Figura 1 se muestra una representación visual de la estructura del modelo. La red está compuesta por un conjunto de capas intermedias compartidas y dos capas de salida independientes. Las capas intermedias incluyen una capa de Unidad recurrente cerrada bidireccional seguida de un conjunto de filtros de convolución. Las salidas de convolución y de las unidades recurrentes finalmente se concatenan y se envían a una capa completamente conectada. Las capas de salida constan de una capa completamente conectada seguida de una capa de campo aleatorio condicional.

Esta estructura permite que el modelo observe los contextos local y global de cada uno de los tokens de entrada. En particular, el contexto local es capturado por la salida de las unidades recurrentes y la salida de la capa de convolución no agrupada, mientras que el contexto global es capturado por la salida de la capa de convolución máxima agrupada. Se agrega información de contexto global adicional cuando se utiliza el modelo basado en BERT concatenando la codificación del token CLS auxiliar .

La información de contexto global y la información de contexto local del token de destino se agregan a todos los pasos de tiempo antes de enviarse a la capa compartida completamente conectada. Luego, los resultados finales se generan mediante una capa de campo aleatorio condicional (CRF). Se ha demostrado que las capas CRF de salida mejoran las capacidades de las redes GRU y LSTM en tareas de etiquetado de secuencias de bajos recursos [4].

2.2. Generación y decodificación de salida.

Como se describe en la Sección 2, nuestro sistema recibe la secuencia de tokens de un documento y el índice de un token y genera los límites de la frase clave más interna a la que pertenece el token.

Estos límites se codifican y decodifican asignando una etiqueta de inicio, interior, unitaria y final a cada token incluido en esa frase clave y salida a todos los demás tokens (etiqueta BIOUE). Una limitación de este enfoque es el hecho de que solo se decodifica una frase clave para cada índice de token, pero esto no es un problema en nuestro caso, ya que las frases clave pueden subsumir pero no superponerse a otras frases clave.

Para cada token de entrada, nuestro modelo genera la lista de probabilidades de relaciones entre la entidad más interna a la que pertenece el token y cada uno de los tokens en el documento.

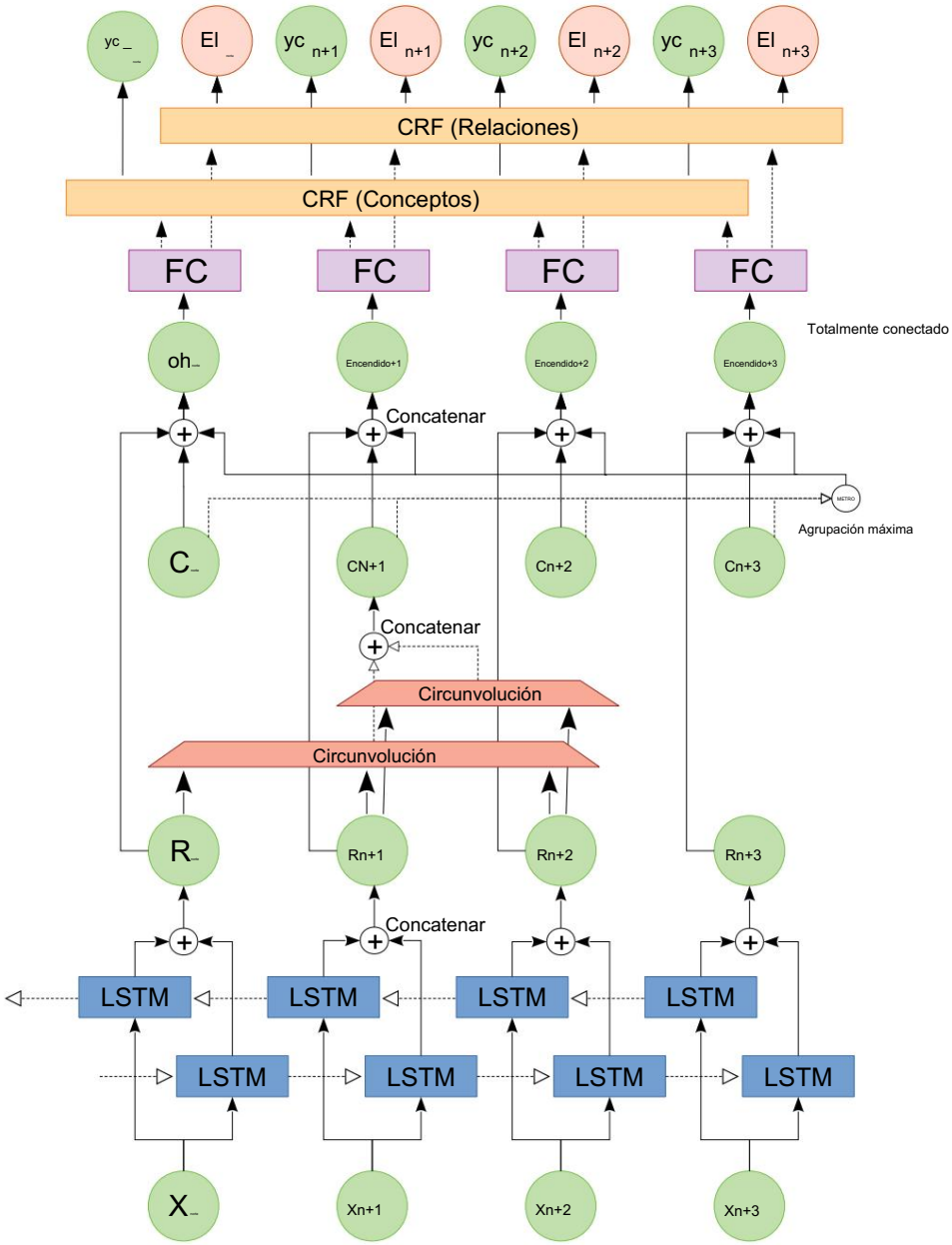

Figura 1: Arquitectura esquemática de la red neuronal artificial de identificación.

predicho. Tenga en cuenta que para el token de origen, solo consideramos la entidad más interna, mientras que para Los tokens de destino los consideramos todas las entidades principales. En consecuencia, nuestro método no permite para relaciones superpuestas del mismo token de origen. Esta restricción se impone para que el La secuencia codificada no es ambigua. Una representación visual de las predicciones de probabilidad de las relaciones. se muestra en la Figura 2. Las relaciones se predicen a partir de la frase clave objetivo si la puntuación agregada

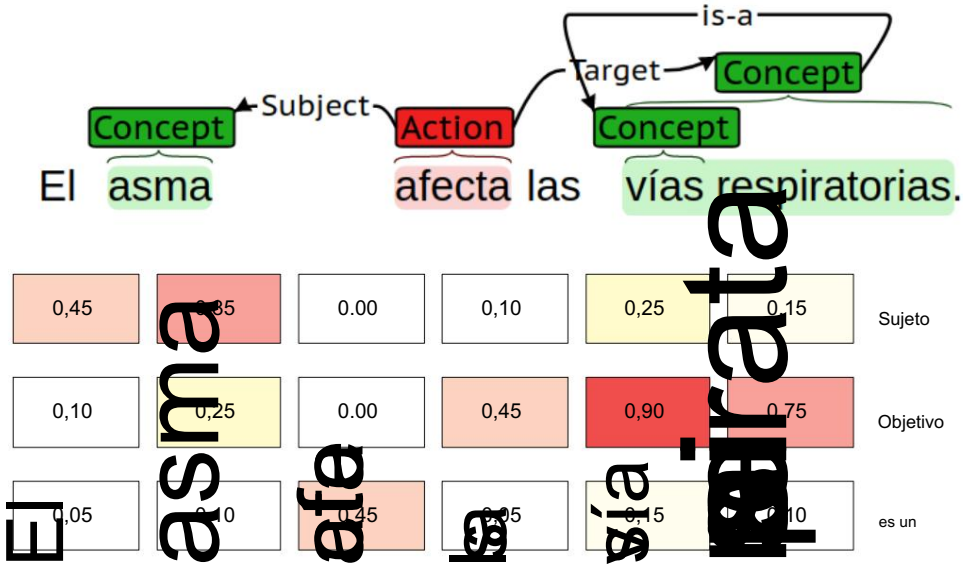

Figura 2: Representación visual de cómo la red codifica las relaciones y frases clave.

dentro de un lapso de frase clave supera un umbral. Solo se selecciona la frase clave con la puntuación más alta si se superponen varias frases clave.

2.3. Funciones de entrada

Como se mencionó anteriormente, nuestro modelo procesa los documentos a nivel de token. Representamos cada token mediante un vector, que resulta de la concatenación de las características que se enumeran a continuación:

- Codificación One-Hot de los campos de categoría y tipo de la etiqueta de parte del discurso del token de Conjunto de etiquetas de FreeLing.
- Vector normalizado que codifica las dependencias encontradas en la ruta entre el token y el token de destino (el que se está decodificando). Se calcula sumando la representación de codificación one-hot de la clase de dependencia para cada salto en la ruta de dependencia y normalizando el vector resultante, sin considerar su dirección. Por ejemplo, la representación de la ficha "I" en "Yo como pescado" cuando la ficha objetivo es "pez" sería un vector con $\sqrt{2}$ en las posiciones correspondientes a "subj" (sujeto) y "cd" (complemento directo); mientras que para "comer" sería un vector con un solo 1 en la posición "cd".
- Codificación One-Hot de la distancia entre el token y el token objetivo.

- Incrustación de palabras del token. Consideramos 4 incrustaciones alternativas de palabras previamente entrenadas modelos:
  - Concatenación de las últimas capas de salida de un BERT multilingüe de uso general[5] model1 sin ajuste fino.
  - Word2Vec y FastText Medical Word Embedding para modelos españoles de Barcelona Centro de supercomputación2 [3].
  - Corporas sin anotaciones en español FastText de SUC3 [6]

2.4. Preentrenamiento con el corpus del conjunto.

Debido a la cantidad comparativamente grande de parámetros en nuestro modelo con respecto al tamaño del conjunto de datos de entrenamiento, el sobreajuste puede ser un problema. Esto lo evitamos utilizando un conjunto relativamente más grande pero inexacto en una fase previa al entrenamiento. Para no permitir que las variables de nuestro modelo caigan en mínimos locales que harían que nuestro modelo imite modelos de años anteriores, agregamos aleatoriamente documentos del corpus de entrenamiento de IberLEF 2020. Además, aumentamos la deserción y disminuimos gradualmente la tasa de aprendizaje para la capacitación y ajustamos los pasos de la capacitación.

2.5. Entrenamiento y ajuste de un solo escenario

En el escenario de evaluación general, la función de pérdida tiene que equilibrar la precisión tanto para el reconocimiento de frases clave como para las tareas de extracción de relaciones. Esto puede resultar problemático, ya que las actualizaciones de parámetros realizadas por el optimizador para mejorar una tarea pueden ser perjudiciales para la otra. Sin embargo, en los escenarios de evaluación 2 y 3, es decir, tareas independientes de reconocimiento de frases clave y extracción de relaciones, el modelo no tiene que generar ambos resultados. En consecuencia, por un lado, podemos utilizar una función de pérdidas sin concesiones. Por otro lado, esto significa no poder explotar la correlación entre tareas, por lo que también podría conducir a un peor rendimiento.

Para estudiar este efecto, sugerimos diferentes estrategias de entrenamiento en un solo escenario: usar el modelo general sin alteración alguna, ajustar las salidas del modelo general con una función de pérdida independiente durante algunas épocas o entrenar el modelo específico desde cero. Tenga en cuenta que en el caso del escenario 3, decodificamos las frases clave utilizando la verdad dorada en lugar del resultado del modelo para las tres estrategias; y concatenar una codificación one-hot de las etiquetas de frases clave a la entrada para la estrategia desde cero . La Tabla 2 muestra los resultados de las tres estrategias de capacitación de un solo escenario.

2.6. Parámetros entrenables y recursos computacionales.

Todos los modelos se entrenaron utilizando el marco TensorFlow® 1.15 para Python® 3.6 en una CPU Intel® Xeon® E5-2620 v4 de 8 núcleos a 2,10 GHz, 16 GB de RAM DDR4 y una GPU GeForce® GTX 1070.

1Utilizamos el modelo BERT-Base, Multilingual Cased (104 idiomas, 12 capas, 768 ocultos, 12 cabezas, 110 millones de parámetros) del repositorio de los autores (https://github.com/google-research/bert)

2Utilizamos SciELO + Wikipedia del 15 de abril de 2020, versión de 300 dimensiones de Medical Word Embedding para español, que se puede descargar desde https://zenodo.org/record/ 3744326 3Utilizamos el modelo binario de subpalabras de 300 dimensiones de https://github.com/dccuchile/spanish-word-embeddings/blob/master/emb-from-suc.md

Modelo	escenario 1			Escenario 4		
	PAG	R	1	PAG	R	1
Vicomtech	0,679364	0,652315	0,665564	0,594009	0,535521	0,563251
UH-MAJA-KD	0,634542	0,615741	0,625	0,608321	0,49813	0,547739
Talp-UPC (presentación)	0,626969	0,626389	0,626679	0,604724	0,563772	0,58353
UPC base (BERT)	0,629630	0,627306	0,628466	0,464271	0,555970	0,506000
Talp-UPC (BERT FT)	0,602778	0,600000	0,601386	0,551309	0,618937	0,583169
Talp-UPC (Salud W2V)	0,573148	0,606268	0,589243	0,382219	0,485488	0,427708
Tapl-UPC (Salud de FastText)	0,569444	0,598832	0,583768	0,400291	0,496906	0,443396
Tapl-UPC (texto rápido general)	0,574537	0,592646	0,583451	0,418363	0,496059	0,453910

tabla 1

Resultados finales de la evaluación de los escenarios 1 y 4 de eHealth-KD de IberLEF2020. También incluimos la evaluación para los corpus específicos del contexto. BERT FT se perfecciona utilizando el desarrollo de aprendizaje por transferencia cuerpo.

y un disco duro Seagate® de 1 TB a 7200 rpm.

Se entrenaron modelos basados en BERT y Word2Vec/FastText para un total de 128 y 96 épocas respectivamente, divididas entre los pasos de preentrenamiento, entrenamiento y ajuste. Capacitación las épocas se distribuyeron uniformemente entre los pasos de entrenamiento previo y de entrenamiento para los modelos sin ajuste fino. Cuando se aplicó el ajuste fino (aprendizaje por transferencia o escenarios de tarea única), la capacitación previa se acortó en 16 épocas.

Para cada modelo de representación de palabras, se entrenaron modelos independientes con 8, 32 y 64 filtros de convolución de tamaños 3 y 5; y 8, 32 y 64 unidades recurrentes monocapa.

3. Resultados

Las tablas 1 y 2 muestran los resultados para los cuatro escenarios del Conocimiento eSalud 2020 de IberLEF Tarea compartida de Discovery para las presentaciones con mejor puntuación. Nuestro modelo es competitivo en todos los escenarios, aterrizando en segunda, tercera, cuarta y primera posición respectivamente.

Las puntuaciones obtenidas al utilizar Word2Vec (W2V Health), FastText específico del contexto (FastText Health) y los modelos FastText de propósito general (FastText General) se muestran en la mitad inferior de la Tabla 1. Con estos modelos, se observa una caída de 0,04 en 1 puntuación en el escenario 1 con respecto a el modelo basado en BERT y un 0,13 en el escenario 4. La diferencia en la puntuación para el escenario 1 entre los modelos de contexto específico y de propósito general es insignificante, mientras que vemos una caída de 0,01 en 1 el escenario 4.

La Tabla 1 también muestra los resultados independientes de nuestro modelo basado en BERT cuando se ajusta con el corpus de desarrollo de transferencia-aprendizaje. El modelo ajustado ve un aumento de 0,08 en la puntuación para 1 el Escenario 4, mientras que solo ve una caída de 0,03 en el Escenario 1.

Como se puede observar en la Tabla 2, nuestro modelo es menos competitivo en el escenario 3, en el que es superado por el modelo de IXA-NER-RE por un margen de casi 0,06 en 1 puntuación. Sin embargo, esto La puntuación coincide con el modelo desde cero no presentado descrito en la Sección 2.5. De manera similar, el La estrategia del modelo general es comparable al modelo del SINAI en el escenario 2.

Modelo	Escenario 2			Escenario 3		
	PAG	R	1	PAG	R	1
SINAI	0,844633	0,806655	0,825207	0,627063	0,365385	0,461725
Vicomtech	0,821622	0,820144	0,820882	0,671679	0,515385	0,583243
IXA-NER-RE	0,726733	0,660072	0.6918	0,647887	0,619231	0,633235
UH-MAJA-KD	0,820255	0,808453	0,814312	0,629237	0,571154	0,59879
Talp-UPC (afinado)	0,807218	0,82464	0,815836	0,646635	0,517308	0,574786
Talp-UPC (general)	0,841727	0,808290	0,824670	0,501923	0,617021	0,553552
Talp-UPC (desde cero)	0,821942	0,810284	0,816071	0,592308	0,678414	0,632444

Tabla 2

Resultados finales de la evaluación de los Escenarios 2 y 3 de eSalud-KD de IberLEF2020. También incluimos la evaluación para las dos estrategias de capacitación adicionales de un solo escenario, que no fueron presentadas.

4. Discusión

El modelo conjunto de clasificación de frases clave y extracción de relaciones presentado por nuestro equipo para La anterior edición de la tarea compartida eHealth Knowledge Discovery de IberLEF superó en rendimiento todos los demás modelos participantes por un amplio margen. Esto confirmó nuestra creencia de que un modelo conjunto tiene el potencial de explotar la información mutua entre las dos tareas y proporcionar mejores resultados de evaluación que la arquitectura tradicional paso a paso. La mejora fue, sin embargo, menos apreciable para la tarea de clasificación de frases clave.

Después de comparar nuestro modelo con el resto de las presentaciones de los participantes, planteamos la hipótesis de que Una de nuestras principales deficiencias fue la absoluta falta de conocimiento específico del contexto. Para En la edición de este año decidimos explorar diferentes alternativas para abordar esto. Pero desde una nueva Se agregó un escenario de transferencia de aprendizaje, cuyo puntaje de evaluación probablemente se vería comprometido. Si el modelo fuente dependía demasiado de características específicas del contexto, optamos por agregar esto. información específica del contexto de una manera que no altere significativamente la estructura del modelo ni hacerlo menos general con reglas hechas a mano. En particular, optamos por cambiar el modelo de representación de palabras de propósito general por uno específico de salud.

Desafortunadamente, los resultados muestran que el uso de incrustaciones de palabras específicas del contexto no mejorar sustancialmente las incrustaciones de uso general e incluso conduce a peores resultados en el escenario de transferencia-aprendizaje. No sólo eso, sino que también hemos mostrado que la palabra contextual Las incrustaciones como BERT y XLNet superan significativamente la incrustación predictiva de palabras. modelos como Word2Vec y FastText. Además, la concatenación de esta segunda palabra La representación no parece aportar ninguna información adicional a la original, si bien hace que el modelo sea más complejo en términos de la cantidad de parámetros entrenables.

Varias hipótesis pueden explicar estos resultados insatisfactorios. En primer lugar, sostenemos que aunque El registro lingüístico de los documentos es formal, el uso de términos técnicos es limitado. Similarmente, Las clases de relaciones y especialmente las categorías de frases clave son posiblemente generales, como lo señala el resultados obtenidos en el Escenario 4. En segundo lugar, es posible que los modelos predictivos de incrustación de palabras no puedan capturar la información semántica de los términos médicos en un grado que pueda ser utilizado por nuestro modelo, pero pueden ser preferibles características bastante más explícitas.

5. Conclusiones

En este artículo hemos descrito las principales características del modelo que hemos desarrollado para la presentación del equipo TALP a la tarea compartida 2020 eHealth Knowledge Discovery de IberLEF. Nuestro modelo sigue la tendencia iniciada por el modelo de 2018 de nuestro equipo, que consiste en utilizar una única red con pesos compartidos que realiza conjuntamente las tareas de reconocimiento de frases clave y extracción de relaciones para aprovechar la información mutua entre los dos. Ha demostrado ser competitivo frente al modelo de otros participantes, especialmente en los escenarios general y de transferencia de aprendizaje, ubicándose en segunda y primera posición respectivamente. El escenario de transferencia de aprendizaje destaca particularmente la adaptabilidad y la independencia del contexto de nuestro modelo.

Se realizaron tres mejoras principales con respecto al modelo del año anterior: tasa de aprendizaje adaptativo para el entrenamiento previo, ajuste de escenario único y representaciones de vectores de palabras específicas del contexto. Sin embargo, lo último ha sido bastante decepcionante y concluimos que agregar información específica del contexto a nuestro modelo sigue siendo un problema sin resolver.

Además de la limitación antes mencionada, vemos otras deficiencias en nuestro modelo que aún deben abordarse para capturar con mayor precisión la información mutua entre las dos tareas de descubrimiento de conocimiento. Entre estas mejoras, nos gustaría señalar dos que creemos que son más prometedoras:

- Utilice una función de combinación entrenable para los resultados generados por el modelo para diferentes tokens de origen en un documento. Nuestro modelo actual, por otro lado, utiliza una operación de unión simple para unir las predicciones de los diferentes tokens de una sola frase clave.
- Uso de un modelo de incrustación de palabras contextuales específico para cada contexto. El uso de incrustaciones de palabras predictivas específicas del contexto no ha demostrado ser exitoso para nuestro modelo, pero las incrustaciones de palabras contextuales de propósito general se pueden ajustar con corpus sin etiquetar específicos del contexto.

Expresiones de gratitud

Esta contribución ha sido financiada parcialmente por el Ministerio de Economía de España (MINECO) y la Unión Europea (TIN2016-77820-C3-3-R y AEI/FEDER,UE).

Referencias

[1] Piad-Morffis, Y. Gutierrez, H. Cañizares-Diaz, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz y A. Montoyo, Overview of the eHealth Knowledge Discovery Challenge en IberLEF 2020, en : Actas del Foro Ibérico de Evaluación de Lenguas compartido con la 36ª Conferencia de la Sociedad Española para el Procesamiento del Lenguaje Natural, IberLEF@SEPLN 2020, España, septiembre de 2020.

[2] S. Medina Herrera, J. Turmo Borras, Talp-upc en ehealth-kd Challenge 2019: Un modelo conjunto con incrustaciones contextuales para la extracción de información clínica, en: Actas del Foro Ibérico de Evaluación de Lenguas (IberLEF 2019): co- ubicado junto a la 35ª Conferencia de la

Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN 2019): Bilbao, España, 24 de septiembre de 2019, CEUR-WS. org, 2019, págs. 78–84.

[3] F. Soares, M. Villegas, A. González-Agirre, M. Krallinger, J. Armengol-Estapé, Incorporaciones de palabras médicas para español: desarrollo y evaluación, en: Actas del 2º Taller clínico de procesamiento del lenguaje natural, Asociación para Lingüística Computacional, Minneapolis, Minnesota, EE. UU., 2019, págs. 124-133. URL: https://www.aclweb.org/anthology/ W19-1916. doi:10.18653/v1/W19-1916.

[4] Z. Huang, W. Xu, K. Yu, Modelos lstm-crf bidireccionales para etiquetado de secuencias, preimpresión de arXiv arXiv :1508.01991 (2015).

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Entrenamiento previo de transformadores bidireccionales profundos para la comprensión del lenguaje, preimpresión de arXiv arXiv:1810.04805 (2018).

[6] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriquecimiento de vectores de palabras con información de subpalabras, Transacciones de la Asociación de Lingüística Computacional 5 (2017) 135-146 .