Machine Translated by Google

HapLap en eHealth-KD Challenge 2020

Sergio Santanaa , Alicia Péreza and Arantza Casillasa

Centro aHiTZ - Ixa, Universidad del País Vasco UPV/EHU, Manuel Lardizabal 1, 20080 Donostia, España

Resumen

Presentamos el trabajo realizado por el grupo HapLap en la subtarea B del concurso eHealth-KD 2020. La extracción de relaciones se abordó con un sistema de canalización que utiliza una red neuronal Joint AB-LSTM junto con una fase de preproceso y postproceso. Obtuvimos un resultado de 0,316 en el Escenario 3.

Palabras clave

Reconocimiento de entidades, Extracción de relaciones, Red neuronal conjunta AB-LSTM.

1. Introducción

Presentamos el trabajo realizado por el grupo HapLap en la tarea eHealth-KD 2020 [1]. En esta tercera edición el objetivo de la tarea es extraer automáticamente conocimiento, representado mediante trece relaciones semánticas, de documentos electrónicos sanitarios españoles. Hemos participado en la subtarea opcional B: la entrada es un texto plano con anotaciones de entidades en un archivo BRAT y la salida es el archivo BRAT anterior con las entidades y las relaciones. Para abordar esto, hemos implementado un sistema de canalización que utiliza una red neuronal Joint AB-LSTM junto con una fase de preproceso y postproceso.

2. Trabajo relacionado

En los últimos años han ido surgiendo diversos concursos relacionados con la extracción de relaciones como: Semeval 2018 tarea 7 [2] para extraer relaciones de textos científicos; eHealthKD 2018 [3], eHealthKD 2019 [4] o BioNLP [5] para extraer y clasificar relaciones clínicas a partir de textos clínicos. Así, el problema de la extracción de relaciones está despertando interés en diferentes áreas y también en el área de documentación clínica. Desde el resurgimiento de las redes neuronales, se han implementado diferentes enfoques para extraer relaciones clínicas. El sistema DET-BLSTM [6] utiliza una red Bi-LSTM. En [7], los autores presentaron una combinación de dos redes diferentes, unidad recurrente cerrada (GRU) y red neuronal convolucional (CNN) para detectar relaciones clínicas. En [8] también se utiliza una red neuronal convolucional para clasificar relaciones. En [9] se utiliza una red neuronal conjunta AB-LSTM para extraer relaciones de reacciones adversas a medicamentos. En este artículo presentamos una articulación neuronal AB-LSTM, un

Actas del Foro de Evaluación de Lenguas Ibéricas (IberLEF 2020) correo
electrónico: ssantana005@ikasle.ehu.eus (S. Santana); alicia.perez@ehu.eus (A. Pérez); arantza.casillas@ehu.eus (A. Casillas)
orcid: 0000-0003-2638-9598 (A. Pérez); 0000-0003-4248-8182 (A. Casillas)
© 2020 Copyright de este artículo por parte de sus autores. Uso permitido bajo la Licencia Creative Commons Attribution 4.0 International (CC BY 4.0).
ISSN 1613-0073 Actas del taller CEUR (CEUR-WS.org)

CEUR-WS.org/Vol-2664/eHealth-KD_paper6.pdf

Machine Translated by Google

modificación del trabajo presentado en la red [10] para la extracción de relaciones clínicas en el contexto del concurso eHealthKD 2020.

3. Materiales y métodos

Para este trabajo hemos dividido el sistema en tres fases: Primero el preproceso, donde adaptamos el formato de datos para usar con el Joint AB-LSTM. Después de eso tenemos la fase de entrenamiento, donde entrenamos y evaluamos la red neuronal y obtenemos la predicción. Y después de obtener las predicciones, tenemos el posproceso, donde convertimos esas predicciones al formato de datos que se utiliza en la competencia.

3.1. Preproceso

En el preproceso realizamos las siguientes operaciones:

- Convierta la entrada del formato de enfrentamiento Brat al formato utilizado en eHeathKD 2019 desafío.
- Convertir los datos en el formato eHealthKD 2019 al formato utilizado por Joint AB-LSTM.
- Crear las relaciones NO_RELATION.

En la primera parte del sistema hemos preprocesado las relaciones de entrada. Hemos convertido el formato de relación de entrada de Brat Standoff (también conocido como ann) al formato utilizado en la competencia anterior eHealthKD 2019 mediante los scripts ann2txt (https://github.com/ conocimiento-aprendizaje/ehealthkd-2019/blob/ master/scripts/ann2txt.py) proporcionado allí. A continuación, necesitábamos adaptarlo a lo que requiere el Joint AB-LSTM. Se han implementado tres programas para el preprocesamiento y su código se ha publicado en GitHub (https://github.com/Porobu/ HAPLAP-MAL). Estos tres programas cargan las instancias que están en el formato de datos eHealthKD 2019 y las unen en un solo archivo.

En un intento de permitir que la red neuronal aprenda a discriminar entre relaciones positivas y negativas (ausencia de relación), se deben proporcionar ambos tipos de instancias en la etapa de inferencia. Para ello, en el preprocesamiento también se creó una clase de relación auxiliar, NO_RELATION. Por lo tanto, un punto crítico es cómo elegir instancias que contengan pares de entidades que podrían estar relacionadas y, por lo tanto, sean relaciones candidatas y etiquetarlas como instancias negativas. Tanto la selección como las proporciones pueden ser cruciales. Hemos utilizado una forma sencilla de elegirlos, que sólo crea relaciones negativas (NO_RELATION) entre pares de entidades que tienen al menos una instancia de relación positiva en el conjunto de datos. Para reducir aún más las relaciones negativas, solo las hemos creado entre pares de entidades en la misma oración.

En esta etapa tenemos un conjunto de datos con los candidatos marcados como relacionados o no relacionados. En este punto, un enfoque de clases múltiples nos permite predecir si un par candidato está relacionado con algunas de las clases de relación disponibles (incluida NO_RELATION). Este fue, de hecho, nuestro enfoque-1: un par de entidades que podrían estar relacionadas (son una relación-candidata) se clasifican directamente mediante el Joint AB-LSTM.

No hace falta decir que en la muestra antes mencionada los casos negativos superan sustancialmente a los positivos, lo que conduce a una distribución de clases sesgada. En la tabla 1 podemos ver el número de positivos.

Machine Translated by Google

tabla 1

Número de relaciones positivas, negativas y totales en los conjuntos de datos de capacitación y desarrollo.

conjunto de datos	Relaciones Positivas	Relaciones Negativas	Total
Capacitación	8597	50812	59409
Desarrollo	1204	7144	8348

y relaciones negativas en nuestros conjuntos de datos de capacitación y desarrollo. Debemos recordar que en nuestro enfoque de clasificación de clases múltiples (método 1), el número de relación positiva contiene las trece clases, lo que distorsiona aún más los datos. La inferencia tiende a estar sesgada hacia la clase mayoritaria. Para hacer frente a esto, propusimos abordar la clasificación en dos etapas (nuestro enfoque-2):

- En la primera fase hemos creado el conjunto de datos binarios, y todas las relaciones positivas (objetivo, causas...) se han agrupado en la clase RELACIÓN . En esta fase filtramos todas las relaciones negativas, para reducir el desequilibrio.

- En la segunda fase ahora tenemos sólo el conjunto de datos con las relaciones positivas (arg, objetivo, sujeto...), y entrenamos el sistema para predecir la relación.

Ambos enfoques (y ambas fases del segundo enfoque) se implementaron mediante el enfoque conjunto AB-LSTM. Se dan más detalles en la siguiente sección.

3.2. Red conjunta AB-LSTM

Después de preprocesar las instancias, las cargamos en la red neuronal Joint AB-LSTM. La red neuronal Joint AB-LSTM se implementó mediante Tensorflow. La red también realiza su propio preprocesamiento. Primero, todos los tokens están en minúsculas.

La red empleó incrustaciones de palabras como característica principal. Para este trabajo hemos utilizado incrustaciones previamente entrenadas del dominio clínico. Las incorporaciones han sido capacitadas en corpus que consisten en EHR (registros médicos electrónicos) que no están disponibles públicamente debido a problemas de confidencialidad. Otras opciones podrían haber resultado más apropiadas que la nuestra, ya que la cantidad y el tipo de datos empleados tienen un gran impacto en las incrustaciones resultantes. Además de las incrustaciones de palabras, la red emplea otra característica poderosa: las incrustaciones a distancia. La distancia se calcula simplemente como el número de tokens entre cada palabra anotada en la oración y la entidad de la palabra objetivo.

Teniendo las relaciones completamente preprocesadas, se entrena la red neuronal. Esta red combina dos redes neuronales ampliamente utilizadas en PNL: una Bi-LSTM con agrupación máxima y una Bi-LSTM atenta. El Joint AB-LSTM se alimenta con las oraciones preprocesadas, sus entidades y relaciones entre ellas, y las incrustaciones a distancia creadas previamente.

Hemos optimizado dos hiperparámetros de la red neuronal, la tasa de abandono y de aprendizaje para obtener el modelo final. Hemos entrenado el modelo con una combinación de los conjuntos de datos train+dev de eHealthKD 2019 y eHealthKD 2020, y hemos utilizado el conjunto de datos de desarrollo de eHealthKD 2020 como validación. Tenga en cuenta que esta optimización se ha realizado sobre el llamado conjunto de datos multiclase (método 1), no sobre el conjunto de datos binarios (método 2). Después de realizar la optimización, establecimos 0,001 como tasa de aprendizaje y no utilizamos abandonos.

Machine Translated by Google

Tabla 2

Resultados del conjunto de datos de desarrollo de eHealthKD 2020 obtenidos con el Enfoque 1 (multiclase) y el Enfoque 2 (trabajando en dos fases para filtrar relaciones binarias).

	Recuperación de precisión	F1
Enfoque 1	0.336	0,298 0,316
Enfoque 2	0.328	0,306 0,316

3.3. Proceso después de

Después de obtener las predicciones de la red neuronal, las postprocesamos para obtener las relaciones de salida en el formato Brat Standoff, respetando los ID de las entidades doradas.

4. Resultados

Como se describe en la sección 3.1, proporcionamos dos enfoques diferentes. Los resultados obtenidos con cada uno de ellos se dan en la tabla 2.

El método 1 supera al método 2 en términos de precisión, pero cuando se produce el retiro, opuesto. Sin embargo, para ambos enfoques la medida F1 tiene el mismo valor.

5. Conclusiones

La extracción de relaciones se abordó con un enfoque neuronal, la red Joint AB-LSTM. Aplicamos dos enfoques de preprocesamiento simples para obtener instancias tanto positivas como negativas. Esta etapa podría resultar ingenua por la forma en que se realizó el muestreo y las proporciones seleccionadas. Exploramos dos enfoques de preprocesamiento: uno directo, el enfoque 1, que simplemente aborda problemas de clases múltiples; uno filtrado (enfoque 2) que intentaba deshacerse de los candidatos negativos antes de la etapa multiclase. Ninguno de ellos superó significativamente al otro. Para trabajos futuros, deberíamos explorar las incorporaciones proporcionadas a la red. Las incrustaciones son la principal fuente de conocimiento en esta etapa con conjuntos de capacitación limitados y demostraron ser significativamente influyentes en trabajos relacionados.

Expresiones de gratitud

Este trabajo fue parcialmente apoyado por el Ministerio de Ciencia y Tecnología de España PAD- MED (PID2019-106942RB-C31) y por el Gobierno Vasco (IXA IT-1343-19 y una Beca para el estudiante Sergio Santana publicada el 03/12/ 2020 BOPV).

Referencias

[1] Piad-Morffis A, Gutiérrez Y, Cañizares-Diaz H, Estevez-Velarde S, Almeida-Cruz Y, Munoz R, Montoyo A, Descripción general del eHealth Knowledge Discovery Challenge en IberLEF 2020, en: Proceedings of the Iberian Languages Foro de Evaluación ubicado conjuntamente con

Machine Translated by Google

36º Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, IberLEF@SEPLN 2020, España, septiembre de 2020., 2020.

[2] K. Gábor, D. Buscaldi, A.-K. Schumann, B. QasemiZadeh, H. Zargayouna, T. Charnois, Semeval-2018 tarea 7: Extracción y clasificación de relaciones semánticas en artículos científicos, en: Actas del 12º Taller Internacional sobre Evaluación Semántica, 2018, págs. 679–688.

[3] E. Cámara Martínez, Y. Almeida Cruz, MC Díaz Galliano, S. Estevez-Velarde, M.A. García Cumbreras, Mª García Vega, Y. Gutiérrez, A. Montejo Ráez, A. Montoyo, R. Muñoz, et al.

[4] 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 (IberLEF 2019). Actas del taller CEUR , CEUR-WS. org, 2019.

[5] D. Demner-Fushman, KB Cohen, S. Ananiadou, J. Tsujii (Eds.), Actas del 18.° taller de BioNLP y tarea compartida, Asociación de Lingüística Computacional, Florencia, Italia, 2019. URL: https:// / www.aclweb.org/anthology/W19-5000.

[6] L. Li, J. Zheng, J. Wan, D. Huang, X. Lin, Extracción de eventos biomédicos a través de redes de memoria a largo plazo a lo largo de un árbol extendido dinámico, en: Bioinformática y Biomedicina (BIBM), Conferencia Internacional IEEE 2016 en adelante, IEEE, 2016, págs.

[7] B. He, Y. Guan, R. Dai, Unidades recurrentes cerradas convolucionales para la clasificación de relaciones médicas , en: Conferencia Internacional IEEE sobre Bioinformática y Biomedicina (BIBM) de 2018, IEEE, 2018, págs.

[8] S. Medina Herrera, J. Turmo Borras, Clasificación conjunta de frases clave y relaciones en documentos electrónicos de salud, en: Actas de TASS 2018: Taller sobre análisis semántico en SEPLN (TASS 2018) ubicado conjuntamente con la 34a Conferencia SEPLN (SEPLN 2018): Sevilla, España, 18 de septiembre de 2018, CEUR-WS. org, 2018, págs. 83–88.

[9] S. Santiso, A. Pérez, A. Casillas, Explorando el ab-lstm conjunto con lemas integrados para el descubrimiento de reacciones adversas a medicamentos, revista IEEE de informática biomédica y de salud (2018).

[10] S. Santiso González, Extracción de reacciones adversas a medicamentos en historias clínicas electrónicas escritas en español (2019).