Reducción de ruido en la supervisión distante para la extracción de relaciones

Juan Luis García Mendoza
Dr. Luis Villaseñor Pineda
Dr. Felipe Orihuela Espina

Reporte Técnico No. CCC-20-001
18 de junio de 2020

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Reducción de ruido en la supervisión distante para la extracción de relaciones

Juan Luis García Mendoza

Dr. Luis Villaseñor Pineda
Dr. Felipe Orihuela Espina

Departamento de Computación
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro # 1, Santa María Tonantzintla, Puebla, 72840, México
{juanluis,villasen,f.orihuela-espina}@inaoep.mx

Resumen

La Extracción de Información es una de las áreas del procesamiento de lenguaje natural que busca transformar la información expresada en textos (i.e. *información no estructurada*) a un formato que permita su explotación por medios computacionales (i.e. *información estructurada*). Una de las dificultades que presenta la extracción de información es la necesidad de contar con uno o varios conjuntos de datos anotados para un dominio específico. La anotación de estos datos de manera manual es muy costosa, por lo que se ha intentado realizarlo de manera automática. Una de las tareas que permite el etiquetado automático de conjuntos de datos es la *supervisión distante*. Bajo este enfoque se utiliza una base de conocimientos que define ciertas relaciones entre las entidades nombradas del dominio. El etiquetado se realiza buscando un par de entidades presentes en una sentencia en la base de conocimiento, en caso de existir, se le coloca la relación correspondiente. Este etiquetado presupone que la sentencia en cuestión expresa la relación correspondiente. Evidentemente, esto puede generar la presencia de datos ruidosos porque el par de entidades no expresa realmente la relación esperada o que la base de conocimientos no contenga información sobre éstas. Como es de esperarse, este ruido provocará un bajo rendimiento en el clasificador, de ahí la importancia de reducirlo. El objetivo principal de este trabajo es la reducción del ruido introducido por la supervisión distante en el etiquetado automático de manera que se extraigan y clasifiquen las relaciones con mayor precisión que los métodos actuales. Para ello se propone el diseño de diferentes estrategias basadas en filtros para eliminar el ruido. De igual forma, también se definirán nuevas representaciones de las instancias para identificar similitudes que permitirán la aplicación de los filtros propuestos. Los experimentos preliminares demuestran la necesidad de obtener nuevas representaciones de las instancias así como la utilidad de la aplicación de filtros generales o para cada clase.

Palabras claves: reducción de ruido, supervisión distante, extracción de información

Tabla de contenidos

Capítulo 1

Introducción

Debido a los avances tecnológicos y el abaratamiento de los dispositivos electrónicos, en los últimos años se ha incrementado considerablemente la cantidad de información digital. Esta información se presenta con frecuencia en forma de textos: noticias en línea, reportes de instituciones, opiniones de usuarios, artículos científicos, entre otras. Una mayoría de toda esta información está en forma textual, lo que dificulta que pueda analizarse por medios automáticos. De ahí la necesidad de transformar esta información no estructurada en una forma estructurada, para permitir su explotación por medios computacionales.

Una de las áreas del procesamiento de lenguaje natural que se utiliza con este fin es la Extracción de Información (EI). Esta tarea constituye el primer paso en la construcción de aplicaciones como sistemas de preguntas y respuestas, seguimiento de noticias, inteligencia de negocios, datos biomédicos, entre otras más. En Sarawagi (2007), se define la EI como la "extracción automática de información estructurada, como entidades, relaciones entre entidades y atributos que describen a las entidades desde fuentes no estructuradas". Autores como Allahyari et al. (2017) y Aggarwal (2018) plantean las tareas de reconocimiento de entidades nombradas y extracción de relaciones como parte de la EI, mientras que otros como Piskorski and Yangarber (2013) y Grishman (2015) adicionan a las anteriores, la resolución de correferencias y la extracción de eventos (ver Figura 1.1). En este trabajo solo se consideran las dos primeras tareas.

Según Sarawagi (2007), "las entidades nombradas constituyen frases nominales y comprenden uno o varios *tokens* dentro de un texto no estructurado". Son un elemento atómico o miembro de una clase semántica que puede variar con respecto al dominio de interés (Goyal et al., 2018). Es decir, las entidades nombradas en el dominio de la biomedicina pueden ser diferentes a las del dominio de los deportes. Existen tipos de entidades de dominio general como los nombres de personas, lugares, organizaciones, números, fechas, horas, entre otras (Sarawagi, 2007; Goyal et al., 2018). La tarea de reconocer (y clasificar) las entidades nombradas (NER) tiene como objetivo identificar y clasificar estas frases nominales presentes en textos no estructurados de acuerdo a los tipos de entidades predefinidos (Nadeau and Sekine, 2007; Piskorski and Yangarber, 2013). Por otra parte, las relaciones entre entidades pueden ser $n - arias$, aunque las más estudiadas en la literatura son las binarias (h, r, t), donde h y

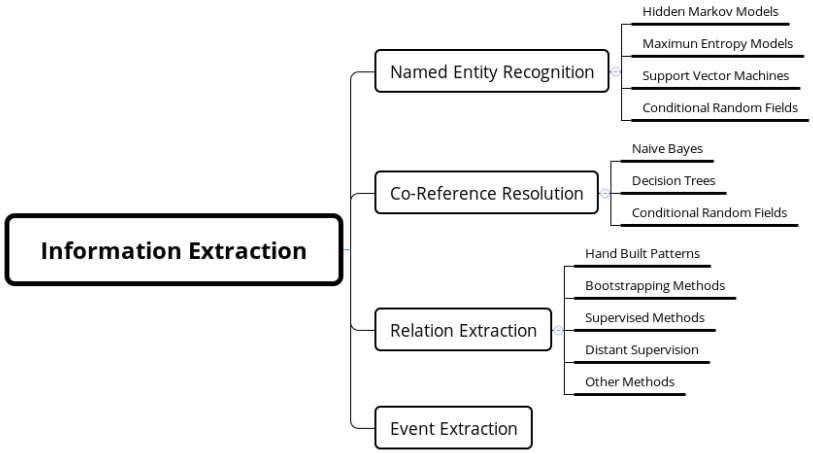


Figura 1.1: Tareas que incluye la EI comúnmente y algunos métodos utilizados en cada una de ellas (extraída de Piskorski and Yangarber (2013)).

t son dos entidades y r una relación del conjunto $\mathcal{R}$ de relaciones predefinidas o de interés (Smirnova and Cudré-Mauroux, 2018). Por último, "la extracción de relaciones es la tarea que detecta y clasifica relaciones predefinidas entre entidades identificadas en el texto" (Piskorski and Yangarber, 2013) (ver Figura 1.2).
Algunos enfoques destacados para resolver esta tarea son:

- Enfoques con datos etiquetados de manera manual tales como;

    1. Patrones construidos por expertos (Hearst, 1992): Estas técnicas requieren la construcción manual por expertos de patrones por cada relación. Debido a esto su mantenimiento es difícil y su cubrimiento es bajo porque las relaciones se basan en un conjunto de patrones finitos.
    2. Métodos semi-supervisados o basados en semillas (Brin, 1998; Agichtein and Gravano, 2000; Etzioni et al., 2005; Saha et al., 2017): Estos métodos parten de un conjunto de semillas por cada relación lo que implica que las relaciones extraídas son sensibles a este conjunto. Las semillas deben construirse de manera manual por parte de expertos en el dominio de aplicación.
    3. Métodos supervisados (Kim and Moldovan, 1993; Rillof, 1996; Soderland, 1999): Estos métodos requieren un conjunto de entrenamiento etiquetado de forma manual por parte de expertos. Es dependiente del dominio y del lenguaje.

- Enfoques con datos etiquetados de manera automática incluyendo;

4. Métodos basados en supervisión distante (Snow et al., 2005; Mintz et al., 2009): Su base inicial es que dadas dos entidades presentes en una misma sentencia se consulta una base de conocimiento para conocer si existe una relación entre ellas, de existir se etiqueta con la relación correspondiente, en caso contrario se toma como un ejemplo negativo. Esto permite la construcción de conjuntos de entrenamiento de manera automática.

- Finalmente, enfoques con datos sin etiquetar.

5. Paradigma *Open Information Extraction*[1] (Banko et al., 2007; Etzioni et al., 2011; Mausam et al., 2012; Pal and Mausam, 2016): Los métodos que se basan en este paradigma tratan de extraer todas las relaciones posibles presentes en un texto. Se han utilizado mayormente en la web.

According to Robert Callahan, president of Eastern's flight attendants union, the past practice of Eastern's parent, Houston-based Texas Air Corp., has involved ultimatums to unions to accept the carrier's terms
The unstructured source
According to ⟨Per⟩ **Robert Callahan** ⟨**/Per**⟩, president of ⟨Org⟩ **Eastern's** ⟨**/Org**⟩ igh t attendants union, the past practice of ⟨Org⟩ **Eastern's** ⟨**/Org**⟩ parent, ⟨Loc⟩ **Houston** ⟨**/Loc**⟩-based ⟨Org⟩ **Texas Air Corp.** ⟨**/Org**⟩, has involved ultimatums to unions to accept the carrier's terms
Annotated entities
Robert Callahan ⟨Employee_Of⟩ Eastern's Texas Air Corp. ⟨Located_In⟩ Houston
Relationships

Figura 1.2: Reconocimiento de entidades y extracción y clasificación de relaciones a partir de un texto plano (extraída de Sarawagi (2007)). Primero a partir de un texto no estructurado se reconocen y clasifican las entidades *Robert Callahan*, *Eastern's*, *Houston* y *Texas Air Corp.*. Luego, se extraen y clasifican las relaciones *Employee_Of* y *Located_In* entre estas entidades.

Una desventaja de los métodos que utilizan datos etiquetados de forma manual es la necesidad de anotar nuevamente los datos al cambiar de dominio. Este etiquetado se realiza por expertos en cada dominio de aplicación, lo cual es costoso. En el caso de los métodos que no requieren datos anotados no se tiene control sobre las relaciones de interés y pueden existir varias

[1] http://openie.allenai.org/ [01/08/2019]

representaciones de la misma relación semántica. Esto dificulta la población de ontologías por lo que se necesita un procesamiento posterior. El enfoque de supervisión distante se sitúa entre los dos tipos de métodos anteriores. No necesitan un conjunto de datos previamente anotado debido a que los anota de manera automática utilizando una base de conocimiento (Mintz et al., 2009; Riedel et al., 2010; Takamatsu et al., 2012; Ru et al., 2018). Este tipo de métodos puede aplicarse a cualquier dominio que cuente con una base de conocimiento. Sin embargo, provoca la aparición de falsos negativos y falsos positivos en el conjunto de entrenamiento. Esto último se debe a que un par de entidades no necesariamente expresan una relación (etiquetas ruidosas) o pueden expresar varias dependiendo del contexto Smirnova and Cudré-Mauroux (2018) (ver Figura 2.2). La existencia de estas etiquetas ruidosas puede provocar una disminución en el rendimiento de la clasificación, cambios en los requisitos de aprendizaje, aumento de la complejidad de los modelos aprendidos, dificultades para identificar características relevantes, entre otras (Frénay and Verleysen, 2014). Debido a esto, en el presente trabajo se pretende desarrollar una estrategia para reducir las etiquetas ruidosas introducidas en un conjunto de datos etiquetado de manera automática mediante la supervisión distante.

Capítulo 2

Trabajo relacionado

La propuesta se organiza en base a la reducción del ruido introducido por la supervisión distante en el etiquetado automático. Primero, se explican algunos modelos preentrenados que existen para representar sentencias mediante *embeddings*. Luego, se mencionan algunos métodos para la extracción y clasificación de relaciones en un escenario ideal, donde los datos no presenten ruido en las etiquetas. Por último, se presentan varios métodos y enfoques existentes en la supervisión distante.

2.1. Representaciones de sentencias mediante *embeddings* de sentencias

1. *Embeddings* propuestos por Cer et al. (2018): Se proponen dos modelos que están optimizados para textos que tengan más de una palabra como sentencias, frases o párrafos. Estos modelos devuelven, dado un texto, un vector. Un modelo se basa en el uso de transformadores (*Transformers*) (Vaswani et al., 2017) que presenta mayor exactitud a costa de mayor consumo de recursos y un modelo más complejo. El otro se basa en *deep averaging networks* (DAN) (Iyyer et al., 2015) que presenta una exactitud menor pero mayor eficiencia. Las fuentes con las que se entrenaron los modelos provienen de Wikipedia, noticias web, páginas de preguntas y respuestas y foros de discusión.

2. *Embeddings* propuestos por Peters et al. (2018): Esta representación conocida como ELMo *(Embeddings from Language Models)* captura características complejas de las palabras como la sintaxis, semántica y polisemia mediante una red profunda bidireccional. Dada una palabra retorna un vector. Estos vectores se pueden utilizar como entrada de una red profunda para obtener el *embeddings* de sentencias (Wang et al., 2018).

3. *Embeddings* propuestos por Devlin et al. (2018): BERT *(Bidirectional Encoder Representations from Transformers)* condiciona conjuntamente el contexto izquierdo y derecho en todas las capas de la red bidireccional profunda. Al igual que en (Cer et al., 2018), se utilizan transformadores (Vaswani et al., 2017). Fue preentrenado a partir del corpus *BooksCorpus* (800 millones de palabras) (Zhu et al., 2015) y Wikipedia en inglés (2500 millones de palabras).

2.2. Extracción y clasificación de relaciones

En la tarea de extracción y clasificación de relaciones se han propuesto varios modelos. En la Tabla 2.1 se muestran los resultados obtenidos por varios modelos supervisados utilizando el conjunto de datos *SemEval-2010 Task 8* (Hendrickx et al., 2010).

2.3. Supervisión distante

La supervisión distante permite el etiquetado automático de relaciones en un conjunto de datos aprovechando el conocimiento existente en varias bases de conocimiento (Mintz et al., 2009; Riedel et al., 2010; Takamatsu et al., 2012; Ru et al., 2018). Ejemplo de estas bases son *YAGO*[1] (Suchanek et al., 2008; Mahdisoltani et al., 2015) y *Freebase*[2] (Bollacker et al., 2008). En Mintz et al. (2009), se asume para la construcción del conjunto de datos con relaciones etiquetadas que "Si dos entidades pertenecen a cierta relación, toda sentencia que contenga esas dos entidades expresa esa relación". Es decir, dado un par de entidades h y t y una relación r almacenada en una base de conocimiento (BC), se etiquetan todas las sentencias del conjunto de datos que contienen ese par de entidades con r (ver Figura 2.1).

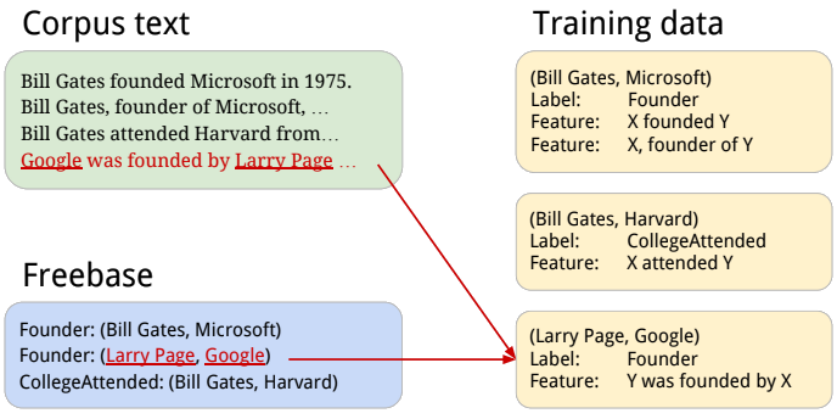


Figura 2.1: Ejemplo de construcción de un conjunto de entrenamiento (extraída del curso *CS224U: Natural Language Understanding*). A partir de las entidades *Google* y *Larry Page* se busca en la base de conocimientos la relación que existe, en este caso, *Founder* y se etiqueta la sentencia con esta.

Este planteamiento trae consigo la aparición de falsos positivos en el conjunto de entrenamiento debido a que un par de entidades presentes en una sentencia no necesariamente

[1] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/ [01/08/2019]

[2] https://developers.google.com/freebase/ [01/08/2019]

Clasificador	F1
R-BERT (*) (Wu and He, 2019)	**89.25**
Att-Pooling-CNN (Wang et al., 2016)	**88.0**
Att-Input-CNN (Wang et al., 2016)	87.5
TCA-CNN (Zhu et al., 2017)	87.3
Att-RCNN (Guo et al., 2019)	86.6
Att-ComNN (Guo et al., 2018)	86.6
BRCNN* (Cai et al., 2016)	86.3
Attention-CNN (*) (Shen and Huang, 2016)	85.9
DRNNs (Xu et al., 2016)	85.8
depLCNN+NS (Xu et al., 2015)	85.6
Entity Attention Bi-LSTM (*) (Lee et al., 2019)	85.2
CNN (Qin et al., 2016)	84.8
MixCNN + CNN (Zheng et al., 2016)	84.8
SPTree (Miwa and Bansal, 2016)	84.5
Hierarchical Attention Bi-LSTM (*) (Xiao and Liu, 2016)	84.3
Bi-LSTM* (Zhang et al., 2015)	84.3
CR-CNN (Nogueira dos Santos et al., 2015)	84.1
VOTE-BACKWATD (Nguyen and Grishman, 2015)	84.1
VOTE-BIDIRECT (Nguyen and Grishman, 2015)	84.1
Attention Bi-LSTM (*) (Zhou et al., 2016)	84.0
MixCNN+LSTM (Zheng et al., 2016)	83.8
RCNN (Zhang et al., 2018)	83.7
SDP-LSTM (Xu et al., 2015)	83.7
DepNN (Liu et al., 2015)	83.6
STACK-FROWARD (Nguyen and Grishman, 2015)	83.4
Hibrid FCM (Yu et al., 2014)	83.4
FCM (Yu et al., 2014)	83.0
CNN+softmax (Zeng et al., 2014)	82.7
MVRNN (Socher et al., 2012)	82.4
SVM (Rink and Harabagiu, 2010)	**82.2**
RNN (Socher et al., 2012)	77.6

(*) se refiere a trabajos que no están en Wang et al. (2016) y se añadieron.
F1 se refiere a la medida *F-measure* con β=1.

Tabla 2.1: Algunos resultados publicados en la literatura en la tarea de extracción y clasificación de relaciones sobre el conjunto de datos *SemEval-2010 Task 8* (Hendrickx et al., 2010) (tabla adaptada de Wang et al. (2016)). Las fuentes originales no publicaron los valores de dispersión.

expresan una relación o pueden expresar varias dependiendo del contexto (ver Figura 2.2).

debido a esto, en Riedel et al. (2010) se relaja esta suposición al plantear que: “Si dos entidades participan en una relación, al menos una sentencia que menciona estas dos entidades puede expresar esa relación”. Esta última suposición se utiliza con frecuencia en la supervisión distante. La supervisión distante, según Smirnova and Cudré-Mauroux (2018), presenta dos

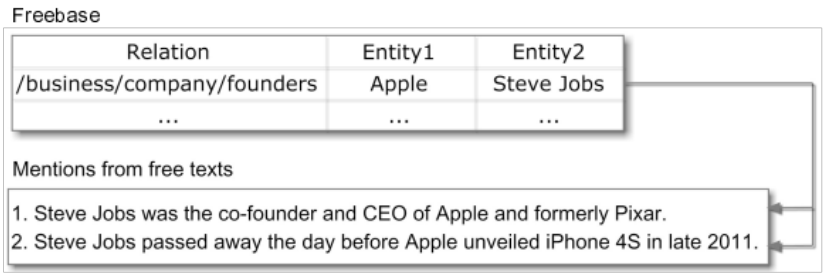


Figura 2.2: Ejemplo de un par de entidades que no expresan la misma relación. Teniendo en cuenta la relación fundador, la primera está etiquetada correctamente, mientras que la segunda no (extraída de Zeng et al. (2015)).

desafíos principales:

1. *Ruido introducido en las etiquetas por la supervisión distante:* Las etiquetas que se obtienen de manera automática con ayuda de la base de conocimiento a menudo no son correctas (etiquetas ruidosas). Esto se debe, principalmente, a que las sentencias que mencionan un par de entidades necesariamente no expresan una relación (falsos positivos) o pueden expresar varias. Este desafío es en el que se concentra este trabajo.

2. *Incompletitud de la base de conocimientos:* Si la base de conocimientos que se utiliza no contiene los pares de entidades que pudieran estar relacionados se pueden etiquetar instancias que presentan una relación como negativas y realmente no lo son (falsos negativos).

La supervisión distante se puede dividir, según Smirnova and Cudré-Mauroux (2018), en tres categorías: reducción de ruido (*Noise Reduction Approaches*) (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012), basada en *embeddings* y redes neuronales profundas (Zeng et al., 2015; Lin et al., 2016; Liu et al., 2017; Ji et al., 2017; Ru et al., 2018; Vashishth et al., 2018; Wu et al., 2018; Xu and Barbosa, 2019; Ye and Ling, 2019) y la que aprovecha información auxiliar (Wang et al., 2018) como las partes del discurso, el tipo de las entidades, entre otras. Estas categorías no son mutuamente excluyentes (Smirnova and Cudré-Mauroux, 2018), por ejemplo, se pueden utilizar *embeddings* en la reducción de ruido. Debido a esa no exclusividad, en este trabajo se decide dividir los métodos en los siguientes enfoques:

- Métodos con tolerancia a las etiquetas ruidosas (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012; Zeng et al., 2015; Lin et al., 2016; Liu et al., 2017; Ji et al., 2017; Vashishth et al., 2018; Wu et al., 2018; Xu and Barbosa, 2019; Ye and Ling, 2019): En este enfoque los clasificadores presentan mecanismos que permiten realizar la clasificación en datos con ruido en las etiquetas. Ejemplo de ello es el uso de redes neuronales profundas con mecanismos de atención sobre las entidades e instancias Lin et al. (2016); Ji et al. (2017); Vashishth et al. (2018); Jat et al. (2018); Ye and Ling (2019).

- Métodos de limpieza de etiquetas ruidosas (Wang et al., 2018; Ru et al., 2018): En este enfoque se reduce el ruido en los datos antes de clasificarlos o se definen nuevos modelos que incluyen en un primer paso la limpieza de las etiquetas.

2.3.1. Métodos con tolerancia a las etiquetas ruidosas

Teniendo en cuenta lo planteado por Riedel et al. (2010), en la reducción de ruido se utiliza el enfoque de aprendizaje multi instancia (MIL) (ver Anexo A) con el objetivo de minimizar los falsos positivos. La idea es tener una bolsa con las instancias que contienen el mismo par de entidades bajo la suposición de que una de ellas es una etiqueta positiva y no es necesario que todas lo sean. Estos autores proponen un modelo gráfico no dirigido que captura y predice las relaciones entre entidades y las instancias que las expresan. Este método tiene como desventaja que no se reducen los falsos positivos directamente y falla cuando un par de entidades aparece solo una vez en el texto y expresa una relación diferente a la etiquetada (Ru et al., 2018). Además, no tiene en cuenta cuando un par de entidades participa en más de una relación (Smirnova and Cudré-Mauroux, 2018).

Con el objetivo de resolver esta última desventaja varios autores utilizan el enfoque *Multi-Instance Multi-Label* (MIML) (Hoffmann et al., 2011; Surdeanu et al., 2012). En (Hoffmann et al., 2011) (MultiR) se presenta un modelo gráfico probabilístico de aprendizaje que maneja las relaciones superpuestas. Por otra parte, en Surdeanu et al. (2012) (MIML-RE) se infiere la etiqueta para una relación en particular, pero permite más de una etiqueta entre estas entidades. Este modelo contiene dos capas, una consiste en un clasificador para determinar si el par de entidades presente en una sentencia presenta una relación y la otra son varios clasificadores binarios que determinan las relaciones a las que pertenece ese par de entidades (ver Figura 2.3).

También se han utilizado patrones negativos para remover etiquetas erróneas (Takamatsu et al., 2012). Éstos ya se conocen a priori que no expresan ninguna relación. Además, se tiene en

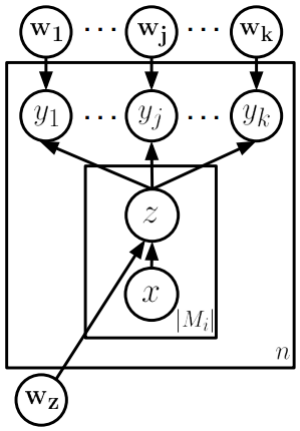


Figura 2.3: Diagrama del modelo MIML-RE (extraída de Surdeanu et al. (2012)). El cuadro externo corresponde a cada uno de los n pares de entidades en la base de conocimiento. Cada par de entidades tiene un conjunto de pares de mención M_i (cuadro interno). La variable x representa el par de mención de entrada, y representa las relaciones positivas y negativas para cada par de entidades. La variable latente z denota una predicción a nivel de mención para cada entrada. El vector de peso para el clasificador multinomial z viene dado por w_z, y hay un vector de peso w_j para cada clasificador binario y_j.

cuenta elementos como que el camino del árbol sintáctico entre las dos entidades no sea mayor a 4. El rendimiento de este método está limitado por la cantidad de instancias de entrenamiento y según Ru et al. (2018) puede ser pobre con suficientes instancias. Esto se debe al uso de patrones negativos construidos de forma manual como semillas.

Los métodos propuestos por Mintz et al. (2009), Riedel et al. (2010), Hoffmann et al. (2011) y Surdeanu et al. (2012) presentan como desventaja la utilización de herramientas existentes para extraer las características de las instancias para los modelos supervisados. De esta manera existe una acumulación del error provocado por ese uso (Zeng et al., 2015; Lin et al., 2016; Zhou et al., 2018).

Basado en el trabajo de Zeng et al. (2014), en Zeng et al. (2015) se propone la red neuronal *Piecewise Convolutional Neural Networks* (PCNN) utilizando el enfoque MIL para el manejo del ruido (PCNN+MIL). La entrada de esta red está conformada por *embeddings* de palabras y de posiciones de cada una de las palabras presentes en las sentencias. La principal modificación con respecto a lo planteado en Zeng et al. (2014), es que se realiza el *max pooling* en tres segmentos determinados por la posición en que se encuentra el par de entidades (ver Figura 2.4). Esta modificación se debe a que *single max pooling* reduce el tamaño de las capas ocultas muy rápido y no captura la información estructural entre las entidades (Zeng

et al., 2015). Este método asume que una bolsa es etiquetada correctamente si y solo si a

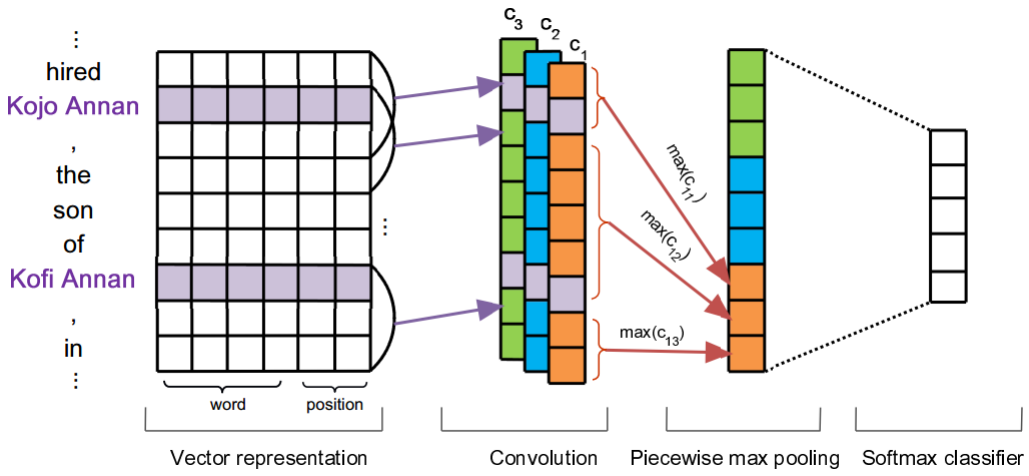


Figura 2.4: Arquitectura de la red PCNN que ilustra el procedimiento para una sentencia de una bolsa que representa al par de entidades *Kojo Annan* y *Kofi Annan* (extraída de Zeng et al. (2015)).

una de las sentencias que contiene se le asigna una etiqueta positiva. Falla cuando a una de las sentencias de la bolsa se le asigna una etiqueta positiva y todas las de la bolsa son falsos positivos (Qin et al., 2018a). También falla si existe más de una sentencia en la bolsa que exprese una relación.

En Lin et al. (2016) se propone atención a nivel de sentencias en múltiples instancias con el objetivo de utilizar la información de todas las sentencias presentes en la bolsa. Este modelo calcula la probabilidad de cada relación r a partir de un conjunto de sentencias $\{x_1, x_2, \ldots, x_n\}$ y el par de entidades correspondiente. Para obtener la representación de cada sentencia x_i se utiliza una red neuronal convolucional (CNN). Luego, cuando se obtienen las representaciones de cada sentencia se utiliza la atención para seleccionar aquellas que expresan la relación correspondiente. Otro método que utiliza atención a nivel de sentencias es el propuesto por Ji et al. (2017), el cual nombran APCNN. Estos autores se basan en la red PCNN (Zeng et al., 2015) (ver Figura 2.4) y además de incorporar el módulo de atención incluyen información sobre las entidades para proporcionar conocimiento que ya se tiene. Cuando se utiliza CNN para codificar la información adicional a partir de las entidades lo nombran APCNN+D. Los métodos APCNN y APCNN+D permiten reconocer varias sentencias en una bolsa como válidas. En Jat et al. (2018) se utiliza atención a nivel de palabras y de entidades.

Los modelos propuestos por Zeng et al. (2015), Lin et al. (2016) y Ji et al. (2017) presentan como desventaja que sus bolsas pueden contener cientos de sentencias (Zhou et al., 2018)

mientras que docenas de sentencias relevantes son suficientes para determinar la relación de un par de entidades. Además, todas las palabras de las sentencias tiene la misma importancia e ignoran que existen palabras que tiene mayor valor para determinadas relaciones. En Zhou et al. (2018) (HSAN) se corrigen estas deficiencias al seleccionar de la bolsa varias instancias relacionadas con la etiqueta para predecir la relación y emplear un mecanismo de atención a nivel de palabras para resaltar dinámicamente partes importantes de la oración. En el caso de los métodos propuestos por Lin et al. (2016) y Ji et al. (2017), se añaden como deficiencias que solo una relación se utiliza para calcular el peso de atención de cada bolsa para obtener su representación en el entrenamiento e ignora cuando todas las sentencias de la bolsa son falsos positivos. Estas dos últimas desventajas también están presentes en Jat et al. (2018) y se intentan corregir en Ye and Ling (2019) utilizando atención a nivel de bolsa y de sentencia de manera conjunta.

En Vashishth et al. (2018) se propone una red neuronal que utiliza información presente en la base de conocimientos con el objetivo de mejorar el etiquetado. Principalmente, se utiliza el tipo de la entidad y alias de las relaciones. Además, se utilizan redes convolucionales de grafos (GCN) (Defferrard et al., 2016) para la modelación de la información sintáctica. Este método, nombrado por los autores como RESIDE, consta de tres componentes. La primera, codificación sintáctica de sentencias (*Syntactic Sentence Encoding*) codifica cada sentencia en la bolsa concatenando las representaciones vectoriales de las palabras y posiciones obtenidos de Bi-GRU y GCN para cada token y le unen la atención sobre ésta, adquisición de información auxiliar (*Side Information Acquisition*) y agregación del conjunto de instancias (*Instance Set Aggregation*).

2.3.2. Métodos de limpieza de etiquetas ruidosas

En Wang et al. (2018) se propone un método libre de etiquetas (*LFDS*) para el etiquetado de los pares de entidades. Para ello utilizan el conocimiento que se tiene a priori de la base de conocimientos y codifican (calculan el *embeddings*) las tripletas (h, r, t) en un espacio continuo de baja dimensión utilizando el método TransE propuesto en Bordes et al. (2013). Las predicciones se realizan calculando el *embeddings* de la sentencia y se compara con el de las relaciones. Se escoge como resultado la relación más cercana. Para construir los *embeddings* de las sentencias se utiliza el modelo *Piecewise Convolutional Neural Networks* (PCNN) (ver Figura 2.4) propuesto en Zeng et al. (2015). La ventaja de este método, además de mejorar los resultados, es que no incluye un modelo extra para manejar el problema del etiquetado. Su principal desventaja es que no detecta correctamente las relaciones que presentan solapamiento

como *place_of_birth* y *place_lived*.

En Ru et al. (2018) se utiliza un algoritmo que calcula la similitud semántica entre fragmentos de texto para reducir las etiquetas erróneas. La idea es calcular la similitud semántica que existe entre la tripleta almacenada en la base de conocimiento que representa la relación (representada mediante su frase de dependencia principal) y la frase de dependencia entre las dos entidades. Ambas frases se representan mediante vectores. Además, se sustituye la entrada de la red convolucional propuesta en Zeng et al. (2014) por la representación de la frase de dependencia principal. En la Figura 2.5 se aprecia el diagrama de lo planteado en Ru et al. (2018).

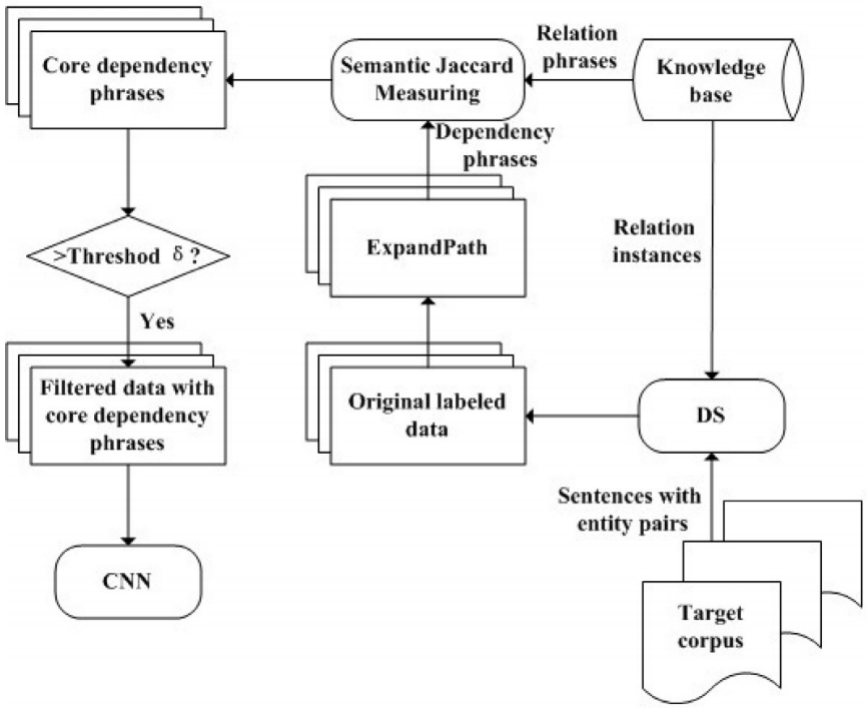


Figura 2.5: Diagrama del modelo propuesto por Ru et al. (2018) (extraída de (Ru et al., 2018)).

En la Tabla 2.2 se muestran los resultados de varios trabajos relacionados con la supervisión distante. La evaluación en la supervisión distante se realiza mediante las curvas de precisión y recuerdo y la precisión en K elementos (Precisión@K) (ver Anexos C.2 y C.3). Las curvas de precisión y recuerdo en la supervisión distante, según (Mintz et al., 2009), son un intento de medir el comportamiento de cada uno de los métodos. Sin embargo, no son concluyentes en cuanto a cuál método es mejor debido a que se basan en las etiquetas originales, las cuáles pueden presentar ruido. Esta presencia de etiquetas ruidosas en el conjunto de evaluación no

permite utilizar las medidas de precisión y recuerdo. Debido a esto se utiliza la precisión en K elementos, donde se mide de manera manual cuántos de los K elementos presentan etiquetas correctas. Lo anterior constituye una de las principales dificultades del área al no ser definitivos los resultados de las curvas de precisión y recuerdo y el costo de la revisión manual. Esta es la razón principal por la que se necesita un conjunto de datos para la supervisión distante en el cuál se controle el ruido, es decir, se conozcan cuáles son las instancias con etiquetas ruidosas.

Clasificador	P@100	P@200	P@300	P@500	Promedio
Evaluación manual					
LFDS (Wang et al., 2018)	0.90	0.88	-	0.83	0.869
SEE (He et al., 2018)	0.91	0.87	-	0.77	0.850
APCNN+D (Ji et al., 2017)	0.87	0.83	-	0.74	0.813
PCNN+ATT (Lin et al., 2016) (+)	0.86	0.83	-	0.73	0.807
APCNN (Ji et al., 2017)	0.87	0.82	-	0.72	0.803
PCNN+MIL+D (Ji et al., 2017)	0.86	0.82	-	0.71	0.797
PCNN+MIL (Zeng et al., 2015)	0.86	0.80	-	0.69	0.783
MIML (Surdeanu et al., 2012) (*)	0.85	0.75	-	0.61	0.737
MultiR (Hoffmann et al., 2011) (*)	0.83	0.74	-	0.59	0.720
Mintz (Mintz et al., 2009) (*)	0.77	0.71	-	0.55	0.676
Evaluación automática					
Intra- and Inter-Bag (Ye and Ling, 2019) (-)	0.918	0.84	0.787	-	0.848
PCNN+ATT+sof-label (Liu et al., 2017) (-)	0.87	0.845	0.77	-	0.828
PCNN+noise_convert+cond_opt (Wu et al., 2018) (-)	0.85	0.82	0.77	-	0.813
RESIDE (Vashishth et al., 2018) (-)	0.84	0.785	0.756	-	0.794
PCNN+ATT (Lin et al., 2016) (-)	0.762	0.731	0.674	-	0.722

(*) Tomado de (Zeng et al., 2015; Wang et al., 2018; He et al., 2018). **(-)** No se especifica por los autores si la evaluación fue manual como en los otros casos, por tanto se toma como automática. **(+)** Tomado de (He et al., 2018) porque se especifica que fue una evaluación manual.

Tabla 2.2: Listado de algunos resultados publicados en la literatura sobre el conjunto de datos New York Times 2010 (Riedel et al., 2010) relacionado con la supervisión distante. Es necesario señalar que no se publican los valores de dispersión.

2.3.3. Discusión

Según Frénay and Verleysen (2014), comúnmente se emplean tres aproximaciones para lidiar con etiquetas ruidosas en los problemas de clasificación: métodos robustos a las etiquetas ruidosas, métodos tolerantes a las etiquetas ruidosas y métodos de limpieza de etiquetas ruidosas. En el caso de la supervisión distante se utilizan los últimos dos enfoques. Este

trabajo se enfoca en los métodos de limpieza de etiquetas ruidosas debido a que los métodos tolerantes al ruido, a pesar de que permiten usar y aprovechar el conocimiento previo (Frénay and Verleysen, 2014), aumentan la complejidad de los algoritmos. Por el contrario, según (Frénay and Verleysen, 2014), los métodos que limpian las etiquetas ruidosas (ver Figura 2.6) son fáciles de implementar pero pueden eliminar demasiadas instancias lo que pudiera reducir el rendimiento de los clasificadores (Matic et al., 1992). No obstante, en Brodley and Friedl (1999) se sugiere que es preferible eliminar varias instancias etiquetadas de manera correcta que mantener instancias con etiquetas ruidosas. Otro elemento de los métodos que limpian las etiquetas ruidosas es que permiten aplicar un método con tolerancia al ruido una vez concluida la fase de limpieza. Debido a que el paso de limpieza o filtrado no garantiza la eliminación total del ruido, una propuesta de solución puede ser utilizar métodos de filtrado y luego aplicar tolerantes al ruido, es decir, vincular los dos enfoques. En los métodos de limpieza, una vez detectada una instancia con etiqueta ruidosa se puede eliminar o reetiquetar (Brodley and Friedl, 1996). En experimentos realizados por Miranda et al. (2009) remover las etiquetas ruidosas resultó más efectivo que reetiquetarlas o utilizar un enfoque híbrido.

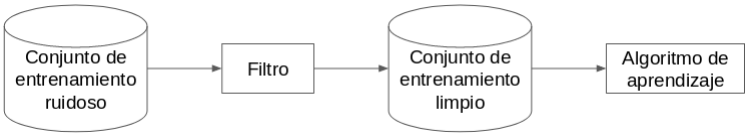


Figura 2.6: Diagrama para la limpieza de conjuntos de datos con etiquetas ruidosas utilizando filtros (Brodley and Friedl, 1999) (extraído de (Frénay and Verleysen, 2014)).

Es necesario señalar que en Frénay and Verleysen (2014) se refieren a la clasificación automática en presencia de ruido, lo cual presenta diferencias con la supervisión distante. En la supervisión distante no se debe confiar en ninguna de las etiquetas debido al etiquetado automático, lo que impide el uso de los algoritmos de detección de valores atípicos tradicionales. Estos algoritmos son utilizados, comúnmente, en datos etiquetados por expertos de manera manual para detectar errores cometidos por estos. Esto es posible porque existen instancias correctamente etiquetadas y la presencia de ruido debe ser menor.

Capítulo 3

Propuesta de Investigación

3.1. Problema de Investigación

Dado un conjunto de sentencias $\mathcal{S}$ y una base de conocimientos Ψ con:

1. Un conjunto de entidades $\mathcal{E}_\Psi$ presentes en Ψ.
2. Un conjunto de entidades de interés $\mathcal{E}$, $\mathcal{E} \subseteq \mathcal{E}_\Psi$.
3. Un conjunto de relaciones $\mathcal{R}_\Psi$ presentes en Ψ.
4. Un conjunto de relaciones de interés $\mathcal{R}$, $\mathcal{R} \subseteq \mathcal{R}_\Psi$.
5. Un conjunto de tripletas $\mathit{\Gamma}_\Psi = \{(h, r, t)|h, t \in \mathcal{E}_\Psi, r \in \mathcal{R}_\Psi\}$ presentes en Ψ.
6. Un conjunto de tripletas de interés $\mathit{\Gamma} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$, $\mathit{\Gamma} \subseteq \mathit{\Gamma}_\Psi$.

La supervisión distante utiliza la información contenida en Ψ para construir un conjunto de entrenamiento etiquetado $\mathcal{D}$, donde a cada sentencia $\mathcal{S}_i$ que contenga una tripleta $(h, r, t) \in \mathit{\Gamma}$ se le asigna la etiqueta r o $\mathcal{NA}$ en caso de no existir la tripleta. Una de las desventajas de esto es la introducción de etiquetas ruidosas $r_{ruidosas}$ debido a que $\mathcal{S}_i$ no exprese la relación r.
Planteamiento del problema: Dado un conjunto de entrenamiento $\mathcal{D}$, etiquetado mediante supervisión distante, se necesita reducir la cantidad de etiquetas ruidosas $r_{ruidosas}$ presentes en $\mathcal{D}$ tal que se incremente la precisión de subsecuentes clasificadores en la extracción y clasificación de relaciones.

3.2. Preguntas, hipótesis, objetivos y contribuciones

3.2.1. Preguntas de Investigación

1. ¿De qué manera se pueden obtener nuevas representaciones de las sentencias utilizando información auxiliar para obtener una mayor similitud entre sentencias que representen la misma relación?
2. A partir de la representación obtenida, ¿se podrían reducir las etiquetas ruidosas utilizando métodos de limpieza?

3. ¿La utilización de métodos de limpieza para la reducción del ruido y métodos de extracción y clasificación tolerantes al ruido pudieran incrementan los valores de precisión con respecto al uso de este último solamente?

4. ¿Cómo se puede utilizar la salida de un clasificador como retroalimentación del método para la reducción de etiquetas ruidosas de manera que se incremente la precisión en la extracción y clasificación de relaciones?

3.2.2. Hipótesis

Utilizando métodos de limpieza de conjunto con métodos tolerantes al ruido se puede reducir el ruido introducido por la supervisión distante durante el etiquetado automático de conjuntos de datos, lo que permite incrementar la precisión en la extracción y clasificación de relaciones.

3.2.3. Objetivos

Objetivo General

Reducir el ruido introducido por la supervisión distante durante el etiquetado automático de conjuntos de datos mediante la combinación de métodos de limpieza y tolerantes al ruido para extraer y clasificar las relaciones con mayor precisión que los métodos actuales.

Objetivos específicos

1. Proponer una nueva forma de obtener representaciones de sentencias utilizando la semántica y sintaxis de las palabras además de información auxiliar sobre las entidades.

2. Definir un método para reducir las etiquetas ruidosas, específicamente los falsos positivos, generados por el enfoque de supervisión distante que permita incrementar la precisión en la extracción y clasificación de relaciones.

3. Incorporar un mecanismo de retroalimentación entre el método de extracción y clasificación de relaciones y el utilizado para reducir el ruido en las etiquetas.

3.2.4. Principales contribuciones

La principal contribución de esta investigación doctoral es una nueva solución al problema de la reducción de etiquetas ruidosas en conjuntos de datos de entrenamiento etiquetados mediante supervisión distante.

Otras contribuciones que se esperan obtener son las siguientes:

1. Un conjunto de datos para la evaluación de supervisión distante debido a la no existencia de uno en el cual el ruido esté controlado.

2. Una nueva forma de obtener representaciones de sentencias que permitan determinar la similitud entre sentencias que expresen la misma relación.

3. Un método que permita, de manera automática, el filtrado de los falsos positivos introducidos por la supervisión distante.

4. Un mecanismo que permita la retroalimentación de los métodos de limpieza con los resultados del clasificador.

3.3. Metodología

En esta sección se detalla la metodología propuesta para alcanzar los objetivos planteados. La metodología planteada consta de los siguientes pasos:

1. **Estudio del estado del arte.**

   a) Identificar y obtener conjuntos de datos etiquetados relacionados con la supervisión distante y la extracción y clasificación de relaciones.

   b) Analizar las características de los conjuntos de datos obtenidos.

   c) Puesta a punto de algoritmos del estado del arte relacionados con la supervisión distante y la extracción de relaciones sobre estos conjuntos de datos.

2. **Proponer una nueva forma de obtener representaciones de sentencias.**
   El objetivo de la nueva representación es que al realizar el agrupamiento por relación se minimice la distancia intra-cluster (sentencias con la misma relación, más cercanas) y maximizar la distancia inter-cluster (sentencias con diferentes relaciones, más lejanas).

   a) Evaluar diferentes representaciones para las sentencias.
      Se evaluarán varias representaciones de sentencias, entre las cuáles se encuentran DAN y TRANSF (Cer et al., 2018), ELMo (Peters et al., 2018) y BERT (Devlin et al., 2018). La evaluación se realizará en los conjuntos de datos descritos en la Sección 4.1 con varias distancias, como la coseno, utilizada en (Wang et al., 2016; Cer et al., 2018).

   b) Definir que información de las sentencias y relaciones se puede utilizar para obtener nuevas representaciones.
      Existe información de sentencias y relaciones que se ha utilizado de manera independiente en diferentes tareas de procesamiento de lenguaje natural. Esta se puede combinar y ser usada como entrada de redes para obtener nuevas representaciones. Estos elementos son:

i) Descripción de las entidades (Ji et al., 2017). La descripción de las entidades puede aportar información de importancia en la extracción y clasificación de relaciones.

ii) El tipo de las entidades (Liu et al., 2014). El tipo de las entidades pueden ser indicadores importantes para decidir las relaciones entre ellas.

iii) La fortaleza de la correlación contextual y las conexiones entre cada palabra con respecto a cada entidad (Wang et al., 2016). Se captura la relevancia de las palabras con respecto a las entidades objetivo.

iv) Frases de dependencia entre las entidades (Ru et al., 2018). En muchos casos, las relaciones entre dos entidades se describen mediante estas frases (Vo and Bagheri, 2018).

v) *POS embeddings* (Rotsztejn et al., 2018). Las etiquetas POS expresan como se utilizan las palabras en la sentencia, lo que puede ayudar a identificar patrones en las relaciones.

Esta información se añadirá como entrada a redes neuronales con el objetivo de observar la influencia que presentan sobre la extracción y clasificación de relaciones.

c) Definir la arquitectura de la red neuronal para obtener nuevas representaciones.
Se decide utilizar redes neuronales no supervisadas debido al ruido que presentan las instancias en sus etiquetas, lo que no las hace confiables. Ejemplo de estas redes son los *autoencoders*, el cual es uno de los métodos más utilizados (Min et al., 2018). De igual manera, se puede utilizar agrupamiento no supervisado (*deepclustering*) (Aljalbout et al., 2018).

d) Validar las representaciones obtenidas.
La validación de las nuevas representaciones se realizará en conjuntos de datos que se han utilizado para la extracción de relaciones y que se encuentran anotados de manera manual. Esta validación se realizará mediante:

i) El coeficiente silhouette (ver Anexo B) para medir la calidad del agrupamiento. Este se realiza de manera manual por las etiquetas de las instancias.

ii) La cantidad de etiquetas que se logran etiquetar correctamente utilizando una k-vecindad.

3. Proponer un método para reducir las etiquetas ruidosas.
El objetivo de este método es la reducción de etiquetas ruidosas, lo que permite que se mejore la extracción y clasificación de relaciones en cuanto a precisión. Una vez detectadas

las etiquetas ruidosas se pueden eliminar del conjunto, cambiarlas a la etiqueta negativa o corregirlas.

a) Diseñar estrategias para eliminar etiquetas ruidosas basadas en la cercanías de las instancias.
   Se diseñarán diferentes estrategias para eliminar etiquetas ruidosas basadas en el cálculo de los K vecinos más cercanos. Además, se utilizarán grupos formados por instancias con el mismo par de entidades y/o con la misma relación. Algunas de estas estrategias son:

   i. Agrupando por relación y/o por pares de entidades de manera manual.
      A. Si la etiqueta mayoritaria de los k vecinos más cercanos no coincide con la del elemento, se considera ruidosa.
      B. Si la distancia al elemento más cercano supera la media del grupo se considera ruidosa.
      C. Si la distancia a todos los elementos del grupo supera la media del mismo $\pm$ 3 desviaciones estándar se considera ruidosa.
      D. Si la distancia al centro del grupo supera la media del mismo se considera ruidosa.
      E. Formar grupos por el par de entidades y aplicar algunas de las técnicas anteriores. Además, se agrupan las instancias por la relación que expresan. En caso de que exista solo una instancia con el par de entidades se verifica si la distancia de esta al grupo que representa la relación que expresa no supera la media.

   ii. Utilizando diferentes algoritmos de agrupamiento para la asignación de las etiquetas.
      A. Se partirá de un conjunto de datos preetiquetados mediante una base de conocimientos y se realizará un agrupamiento no supervisado.
      B. Se partirá de un conjunto de datos con un subconjunto de relaciones bien etiquetadas entre pares de entidades (semillas). Se toman estas semillas como centros de los grupos y se intentan agrupar el resto de posibles relaciones en cada uno de ellos. En caso de no superar un umbral de similitud determinado se crea un nuevo grupo con esa relación.

b) Establecer una función que permita el cálculo de un índice de confianza sobre cada instancia etiquetada de manera automática.

   Con el índice de confianza sobre cada instancia etiquetada de manera automática se puede determinar cuan correcta es. A partir de este valor se puede realizar un corte por un umbral para determinar cuáles instancias contienen etiquetas ruidosas. Este índice puede ser obtenido mediante redes neuronales profundas (Ji et al., 2017) o la probabilidad que brindan los clasificadores de que la instancia pertenezca a la clase.

c) Diseñar estrategias para eliminar etiquetas ruidosas basadas en redes neuronales profundas.

   En la actualidad varios trabajos relacionados con la supervisión distante proponen redes neuronales que incluyen el manejo del ruido. Estas redes utilizan el planteamiento de Riedel et al. (2010) y varios niveles de atención (Lin et al., 2016; Ji et al., 2017; Jat et al., 2018; Ye and Ling, 2019).

d) Diseñar ensambles de estrategias para la reducción de etiquetas ruidosas. Debido a la naturaleza del lenguaje y las características de cada relación pueden ser necesarias más de una estrategia de reducción de etiquetas ruidosas, es decir, ensambles de estrategias. Se utilizarán las variantes *bagging*, voto pesado o un nuevo clasificador para combinar esos resultados.

e) Evaluar las estrategias para la reducción de las etiquetas ruidosas en un conjunto de datos controlado. Las estrategias anteriores se evaluarán en conjuntos de datos donde se conozcan las etiquetas ruidosas. De esta manera, también se podrán utilizar las medidas de precisión, recuerdo y F1 para realizar la comparación de estrategias así como la tasa de reducción de ruido.

4. **Incorporar un mecanismo de retroalimentación.**

a) Definir mediante que función el clasificador evaluará el rendimiento de las estrategias de reducción de etiquetas ruidosas.

   Se definirá la función que le permita al clasificador determinar cuan efectivas fueron las estrategias de reducción. Debido a que las etiquetas pueden ser ruidosas no se pueden utilizar las medidas de precisión y recuerdo. Una solución a esto es utilizar las probabilidades de pertenencia a cada clase.

b) Incorporar el clasificador al conjunto de ensambles de estrategias de reducción de ruido. El clasificador puede formar parte de las estrategias de ensamble, tomando como

instancias con etiquetas ruidosas aquellas en las que menos confía. La desventaja de esto es que las etiquetas pueden contener ruido.

c) Diseñar un método iterativo que involucre las estrategias de reducción de etiquetas ruidosas con los algoritmos de extracción y clasificación de relaciones.
A partir de la salida del clasificador se puede determinar cuan efectivas fueron las estrategias de reducción y, en caso de ser necesario, corregir los parámetros de éstas. El criterio de parada puede ser cuando en la próxima iteración no se mejore la anterior.

3.4. Plan de Trabajo

	Tareas	2018	2019				2020				2021				2022		
		4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3
1	Revisión del estado del arte																
2	Propuesta de doctorado																
3	Defensa de la propuesta doctorado																
4	Evaluaciones																
5	Identificación, localización y adquisición de las bases de de datos más significativas.																
6	Obtención de una nueva representación de sentencias.																
7	Obtención de un nuevo método para la reducción de ruido en la supervisión distante																
8	Construcción de un conjunto de datos																
9	Agregarle al método propuesto la salida del clasificador																
10	Publicaciones																
11	Elaboración de la tesis doctoral																
12	Defensa del doctorado																

Figura 3.1: Plan de Trabajo durante el doctorado.

3.5. Plan de Publicaciones

1. **Conference** Distant supervision for relation extraction using k-nearest neighbor and CNNs. *The 28 th International Conference on Computational Linguistics (COLING)*, **deadline:** April 2020.

2. **Journal** Sentences Representation for the relation extraction with deep learning. *Information Systems*, *Factor de impacto:* 2.066.

3. **Journal** Distance supervision for relation extraction combining cleaning methods with deep learning. *Information Processing Management*, *Factor de impacto:* 3.892.

Capítulo 4

Resultados preliminares

El trabajo realizado hasta hoy consiste en lo siguiente:

1. Identificar y obtener los conjuntos de datos.
2. Evaluar la influencia de diferentes representaciones de los datos de entrada sobre la extracción y clasificación de relaciones utilizando redes convolucionales.
3. Evaluación de diferentes representaciones de sentencias.
4. Evaluar estrategias de los k vecinos más cercanos para eliminar las etiquetas ruidosas.
5. Trabajo en la construcción de un conjunto de datos en un dominio de ejemplo para la extracción y clasificación de relaciones utilizando la supervisión distante.

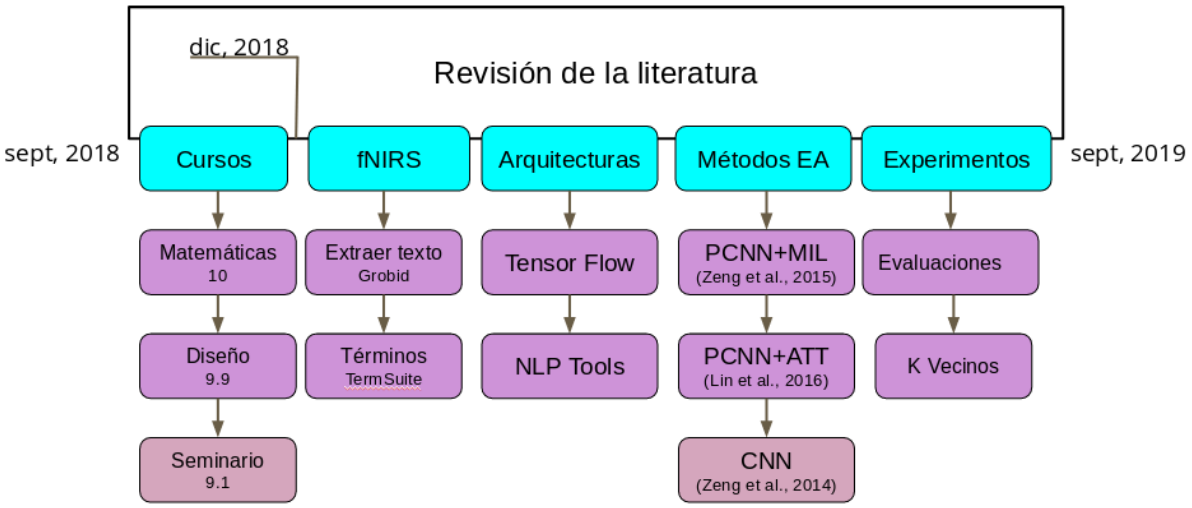


Figura 4.1: Trabajo realizado durante el primer año del doctorado.

4.1. Conjuntos de datos

4.1.1. Conjuntos de datos para la supervisión distante

En el caso de la supervisión distante, el conjunto de datos que se utiliza en varios trabajos es el propuesto por Riedel et al. (2010) basado en noticias del periódico *The New York Times*

(NYT2010) (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012; Takamatsu et al., 2012; Zeng et al., 2015; Lin et al., 2016; Liu et al., 2017; Ji et al., 2017; Ru et al., 2018; Vashishth et al., 2018; Wang et al., 2018; Jat et al., 2018; Ye and Ling, 2019) que se encuentra alineado con la base de conocimientos *Freebase*[1] (Bollacker et al., 2008). Recientemente, Jat et al. (2018) crearon un conjunto de datos, nombrado *Google Distant Supervision (GDS)*, para la supervisión distante basado en el conjunto para la extracción de relaciones *Google Relation Extraction*. Los autores garantizan que dentro del conjunto de instancias que comparten el mismo par de entidades existe al menos una que realmente exprese la relación. Esto no lo garantiza, hasta el momento, ningún conjunto de datos dedicado a esta tarea.

New York Times (NYT2010)

Este conjunto de datos[2] fue creado por (Riedel et al., 2010) para la tarea de supervisión distante. Cuenta con 53 tipos de relaciones, incluyendo $\mathcal{NA}$, que indica la no existencia de una relación. La Tabla 4.1 muestra la cantidad de instancias, pares de entidades y relaciones diferentes de $\mathcal{NA}$ que contienen los conjuntos de entrenamiento y de evaluación. Existen sentencias en este conjunto que no contiene el par de entidades definido por lo cuál se tomaron como incorrectas y no se incluyeron.

	Conjunto de entrenamiento		**Conjunto de evaluación**	
	Total	**Correctas**	**Total**	**Correctas**
Cantidad de instancias	522611	521793	172448	172415
Pares de entidades	279226	279079	96678	96655
Relaciones	136379	135811	6444	6441

Tabla 4.1: Cantidad de instancias del conjunto de datos NYT2010.

Google Distant Supervision (GDS)

Este conjunto de datos[3], creado por Jat et al. (2018), cuenta con cinco tipos de relaciones, incluyendo $\mathcal{NA}$, que indica la no existencia de una relación (ver Tabla 4.2). Los autores de este conjunto lo dividieron en tres particiones, entrenamiento (11297 instancias y 6498 pares de entidades), prueba (1864 y 1082) y evaluación (5663 y 3247).

[1] https://developers.google.com/freebase/ [01/08/2019]
[2] http://iesl.cs.umass.edu/riedel/ecml/ [01/08/2019]
[3] https://github.com/SharmisthaJat/RE-DS-Word-Attention-Models [01/082019]

	Entrenamiento		Prueba		Evaluación	
Relación	**I**	**PE**	**I**	**PE**	**S**	**PE**
perGraduatedInstitution	2652	1575	439	262	1365	787
perHasDegree	1785	860	290	143	894	429
perPlaceOfBirth	2001	1296	323	216	1032	647
perPlaceOfDeath	2088	1169	365	195	1016	584
$\mathcal{NA}$	2771	1601	447	266	1356	800
Total	11297	6501	1864	1082	5663	3247

I: instancias. **PE:** pares de entidades.

Tabla 4.2: Cantidad de sentencias y pares de entidades por relación del conjunto GDS.

4.1.2. Conjuntos de datos para la extracción y clasificación de relaciones de manera supervisada

Para la extracción y clasificación de relaciones se escoge *SemEval-2010 Task 8* propuesto por Hendrickx et al. (2010) por ser de los más utilizados para esta tarea en la literatura (ver Tabla 2.1). Recientemente, se publicó el conjunto de datos *SemEval-2018 Task 7* propuesto por Gábor et al. (2018) para la extracción de relaciones en resúmenes de artículos científicos.

SemEval-2010 Task 8 (SemEval2010)

Este conjunto de datos[4] fue liberado como parte de la tarea ocho, nombrada *Multi-Way Classification of Semantic Relations Between Pairs of Nominals*, del evento SemEval-2010 (Hendrickx et al., 2010). Cuenta con 10 relaciones, incluyendo "Other" que representa $\mathcal{NA}$. De estos tipos de relaciones, nueve se representan de manera bidireccional convirtiéndose en 18 tipos más "Other". Las Tablas 4.3 y 4.4 muestran la cantidad de instancias por cada relación en las particiones de entrenamiento (8000) y evaluación (2717) del SemEval2010 respectivamente. Este conjunto al no estar orientado a la supervisión distante la cantidad de pares de entidades están cercanas a la cantidad de instancias.

4.2. Experimentos

4.2.1. Evaluación de la influencia de representaciones más ricas como datos de entrada

Objetivo: Evaluar la influencia de diferentes representaciones de los datos de entrada sobre la extracción y clasificación de relaciones utilizando redes convolucionales.
Conjunto de datos: SemEval2010.

[4]https://drive.google.com/file/d/0B_jQiLugGTAkMDQ5ZjZiMTUtMzQ1Yy00YWNmLWJlZDYtOWY1ZDMwY2U4YjFk /view?sort=name&layout=list&num=50 [01/08/2019]

Relación	Instancias	Relación	Instancias	Total
Cause-Effect(e1,e2)	344	*Cause-Effect(e2,e1)*	659	1003
Component-Whole(e1,e2)	470	*Component-Whole(e2,e1)*	471	941
Content-Container(e1,e2)	374	*Content-Container(e2,e1)*	166	540
Instrument-Agency(e1,e2)	97	*Instrument-Agency(e2,e1)*	407	504
Entity-Destination(e1,e2)	844	*Entity-Destination(e2,e1)*	1	845
Entity-Origin(e1,e2)	568	*Entity-Origin(e2,e1)*	148	716
Member-Collection(e1,e2)	78	*Member-Collection(e2,e1)*	612	690
Message-Topic(e1,e2)	490	*Message-Topic(e2,e1)*	144	634
Product-Producer(e1,e2)	323	*Product-Producer(e2,e1)*	394	717
Other ($\mathcal{NA}$)	1410			1410

e1: entidad 1 (h). **e2:** entidad 2 (t).

Tabla 4.3: Cantidad de instancias por relación en el conjunto de entrenamiento de SemEval2010.

Relación	Instancias	Relación	Instancias	Total
Cause-Effect(e1,e2)	134	*Cause-Effect(e2,e1)*	194	328
Component-Whole(e1,e2)	162	*Component-Whole(e2,e1)*	150	312
Content-Container(e1,e2)	153	*Content-Container(e2,e1)*	39	192
Instrument-Agency(e1,e2)	22	*Instrument-Agency(e2,e1)*	134	156
Entity-Destination(e1,e2)	291	*Entity-Destination(e2,e1)*	1	292
Entity-Origin(e1,e2)	211	*Entity-Origin(e2,e1)*	47	258
Member-Collection(e1,e2)	32	*Member-Collection(e2,e1)*	201	233
Message-Topic(e1,e2)	51	*Message-Topic(e2,e1)*	210	261
Product-Producer(e1,e2)	108	*Product-Producer(e2,e1)*	123	231
Other ($\mathcal{NA}$)	454			454

e1: entidad 1 (h). **e2:** entidad 2 (t).

Tabla 4.4: Cantidad de instancias por relación en el conjunto de evaluación de SemEval2010.

Red utilizada: Red convolucional propuesta por Zeng et al. (2014) y disponible en *Github*[5]
Resultados esperados: Que se incremente el rendimiento de la red al incorporar representaciones más ricas de los datos como entrada.

La elección de la arquitectura propuesta por Zeng et al. (2014) se debe a que no tiene componentes adicionales como niveles de atención. De esta manera, se puede observar solo la influencia de los datos de entrada sobre la extracción y clasificación de relaciones. La configuración de esta red se mantiene igual a la implementación de *Github* (ver Tabla 4.5). El

[5]https://github.com/roomylee/cnn-relation-extraction [01/08/2019]

embedding de palabras que se utilizó es el publicado por Google[6], entrenado sobre un conjunto de noticias de Google que contiene cerca de 100 mil millones de palabras. Este modelo contiene vectores de 300 dimensiones para 3 millones de palabras y frases. El *embeddings* de partes de la sentencia que se utilizó fue generado a partir de 800 000 entradas de Wikipedia sustituyendo cada palabra por su etiqueta POS y presenta 50 dimensiones.

Parámetro	**Valor**	**Parámetro**	**Valor**
Tamaño del WE	300	Número de épocas	100
Tamaño del PF	50	Tamaño de los filtros	2,3,4,5
Tamaño del POSE	50	Cantidad de filtros por tamaño	128
Optimizador	*AdadeltaOptimizer*	Tamaño de los lotes	20
Índice de aprendizaje	1.0		

WE: *embeddings* de palabras. **PF:** *embeddings* de posiciones. **POSE:** *embeddings* de partes de la sentencia.

Tabla 4.5: Valores iniciales de los parámetros de la red convolucional.

En la Tabla 4.6 se puede observar que la configuración con mejor comportamiento, en cuanto a macro-F1, está compuesta por WE, PF y POSE. Los valores obtenidos cuando se utilizan WE solamente, WE y PF (entrada original de la red) o WE y POSE son similares. A partir de la mejor combinación (CNN + WE + PF + POSE), se evaluó el impacto de los tamaños 20, 50, 100 y 200 del PF en los resultados. En la Tabla 4.7 se muestra que presentan un comportamiento similar. Esto indica que, a pesar de que PF ayuda en la extracción, utilizar un tamaño mayor del *embeddings* no garantiza mejores resultados.

Modelo	**Macro-Precisión**	**Macro-Recuerdo**	**Macro-F1**
CNN + WE	80.38 ± 0.80	82.08 ± 0.86	81.12 ± 0.36
CNN + WE + PF	79.73 ± 0.72	82.73 ± 0.93	81.12 ± 0.37
CNN + WE + POSE	79.71 ± 0.67	82.93 ± 0.75	81.19 ± 0.34
CNN + WE + PF + POSE	$\mathbf{80.83 \pm 0.55}$	$\mathbf{83.49 \pm 0.95}$	$\mathbf{82.04 \pm 0.35}$

WE: *embeddings* de palabras. **PF:** *embeddings* de posiciones. **POSE:** *embeddings* de partes de la sentencia.

Tabla 4.6: Resultados de 40 ejecuciones de CNN y varios *embeddings* como entrada de la red.

Conclusiones del experimento 1

1. La incorporación de una representación más rica como entrada, en este caso, el POSE incrementó el rendimiento del clasificador.

[6]https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing [01/08/2019]

Modelo	Precisión	Recuerdo	Macro-F1
CNN + WE + PF(20) + POSE	80.61 ± 0.53	83.59 ± 0.93	82.00 ± 0.43
CNN + WE + PF(50) + POSE	80.83 ± 0.55	83.49 ± 0.95	82.04 ± 0.35
CNN + WE + PF(100) + POSE	80.70 ± 0.53	83.64 ± 1.00	82.06 ± 0.44
CNN + WE + PF(200) + POSE	80.80 ± 0.55	83.55 ± 1.00	82.06 ± 0.48
CNN + WE + PF(300) + POSE	80.79 ± 0.47	83.63 ± 0.74	82.10 ± 0.37

WE: *embeddings* de palabras. **PF:** *embeddings* de posiciones. **POSE:** *embeddings* de partes de la sentencia.

Tabla 4.7: Resultados de 40 ejecuciones de la red CNN tomando como entrada, además de WE y POSE, las dimensiones 20, 50, 100, 200 y 300 del PF.

2. El tamaño del PF no influyó en el rendimiento del clasificador.

4.2.2. Evaluación de diferentes representaciones de sentencias

Objetivo: Evaluar diferentes representaciones de sentencias mediante el coeficiente silhouette (ver Anexo B) al agrupar por relación y diferentes distancias y funciones de similitud.
Conjuntos de datos: GDS y SemEval2010.
Representaciones utilizadas: DAN y TRANSF (Cer et al., 2018).
Resultados esperados: Que las sentencias con igual relación queden cercanas entre ellas y alejadas de las demás. Que al incluir las entidades se mejore el coeficiente de silhouette para el conjunto GDS.

El agrupamiento se realiza de manera manual teniendo en cuenta la relación que expresa cada instancia. Es decir, las instancias con igual relación pertenecen al mismo grupo. La relación $\mathcal{NA}$ (ninguna relación) no se tuvo en cuenta en el agrupamiento. Las representaciones que se utilizan para las instancias son DAN y TRANSF (Cer et al., 2018) tomando todo el texto (Largo), el texto entre las dos entidades incluyéndolas (Corto_E) y sin incluirlas (Corto).
En las Tablas 4.8 y 4.9 se puede observar que el mejor comportamiento se obtiene con la distancia coseno y la representación con TRANSF en los conjuntos de datos GDS y SemEval2010. La representación TRANSF funcionó mejor que DAN para los tres textos, obteniéndose el valor más alto en el texto entre las dos entidades incluyendo a éstas. Esto pudiera deberse a que si se toma todo el texto se tienen en cuenta elementos que no son de importancia para representar la relación. De igual manera, si se toma el texto entre las dos entidades sin incluirlas se pierde información sobre éstas. Además, puede ocurrir que varias relaciones se representen de igual manera y la diferencia esté marcada por las entidades. Por

otra parte, con ambas representaciones se obtiene un coeficiente de silhouette cercano a 0, lo que indica que las distancias entre los elementos de un mismo grupo y de estos hacia los otros son cercanas. Por tanto, no se logra separar correctamente las instancias que presentan una misma relación de las demás.

	GDS					
	Corto		Corto_E		Largo	
Distancia	**DAN**	**TRANSF**	**DAN**	**TRANSF**	**DAN**	**TRANSF**
Coseno	0.026	**0.031**	0.023	**0.032**	0.021	**0.026**
Euclidiana	0.016	0.020	0.014	0.018	0.012	0.015
Manhattan	0.016	0.020	0.014	0.019	0.013	0.016
Chebyshev	0.008	0.012	0.004	0.012	0.001	0.008

Corto: se toma el texto que existe entre las dos entidades sin incluirlas. **Corto_E:** se toma el texto que existe entre las dos entidades incluyéndolas. **Largo:** se toma todo el texto.

Tabla 4.8: Coeficiente silhouette para las funciones de distancia coseno, euclidiana, manhattan y chebyshev utilizando los modelos DAN y TRANSF para la representación y agrupando por relación.

	Semeval2010					
	Corto		Corto_E		Largo	
Distancia	**DAN**	**TRANSF**	**DAN**	**TRANSF**	**DAN**	**TRANSF**
Coseno	-0.039	**-0.030**	0.010	**0.016**	-0.003	**0.000**
Euclidiana	-0.009	-0.003	0.006	0.010	-0.001	0.000
Manhattan	-0.010	-0.004	0.007	0.009	-0.001	0.000
Chebyshev	-0.009	-0.009	0.000	0.006	-0.002	-0.001

Corto: se toma el texto que existe entre las dos entidades sin incluirlas.
Corto_E: se toma el texto que existe entre las dos entidades incluyéndolas.
Largo: se toma todo el texto.

Tabla 4.9: Coeficiente silhouette para las funciones de distancia coseno, euclidiana, manhattan y chebyshev utilizando los modelos DAN y TRANSF para la representación y agrupando por relación.

Conclusiones del experimento 2

1. No se logra separar correctamente las instancias que presentan una misma relación de las demás con estas dos representaciones.

2. Se obtienen los mayores coeficientes de silhouette con la distancia coseno y tomando el texto entre las dos entidades incluyendo a éstas.

3. Valorar utilizar otra medida para la evaluación, por ejemplo, ver cuántas instancias logro etiquetar correctamente en una K-vecindad determinada.

4.2.3. Estrategia de los k vecinos más cercanos para eliminar las etiquetas ruidosas.

Objetivo: Evaluar la estrategia de k vecinos más cercanos para determinar las etiquetas ruidosas.
Conjunto de datos: GDS.
Representaciones utilizadas: DAN y TRANSF (Cer et al. (2018)) tomando el texto que se encuentra entre las dos entidades sin incluirlas (Corto) e incluyéndolas (Corto_E)
Vecinos utilizados: 1, 2, 3, 5, 7 y 9.
Estrategias para el manejo de instancias con etiquetas ruidosas: Cambiar a NA o eliminar instancias.
Clasificadores: Redes convolucionales CNN y PCNN+ATT.
Resultados esperados: Que se incremente la precisión del clasificador sobre los conjuntos resultantes con respecto al conjunto original.
Una de las estrategias que se evaluará para determinar las etiquetas ruidosas es un filtro que utilice a los k vecinos más cercanos. Esta estrategia se utiliza, a pesar de lo resultados del Experimento 4.2.2, porque solo se utilizará una k vecindad determinada y no todo el grupo. Si la etiqueta que predomina en los k vecinos más cercanos no coincide con la del elemento, se debe utilizar una de las estrategias para manejar las etiquetas ruidosas. Si existe empate, se promedia la distancia al elemento de los vecinos con la misma etiqueta y se toma la etiqueta de los vecinos con el menor promedio. Esta estrategia se muestra en el Algoritmo 1, el cual retorna el conjunto de instancias final luego de aplicar una de las estrategias a las etiquetas detectadas como ruidosas.
Para obtener los nuevos conjuntos de datos con menos ruido se aplicó el Algoritmo 1 solamente a la partición de entrenamiento de GDS para los valores de k, *1, 2, 3, 5, 7 y 9* (ver Figura 4.2). Los parámetros de la red CNN (Zeng et al., 2014) que se utilizaron son los mismos que se presentan en la Tabla 4.5 con excepción del índice de aprendizaje que se cambió a 0.001. Por último, se realiza la evaluación mediante curvas de precisión y recuerdo y la precisión@k y se comparan con los obtenidos por el *baseline* (ver Figura 4.3) que consiste en aplicar la red sobre el conjunto de datos original.

Algorithm 1 Eliminar etiquetas ruidosas utilizando k vecinos más cercanos.

Require: $instancias$: instancias con etiquetas diferentes a $\mathcal{NA}$.
Require: K: cantidad de vecinos a tener en cuenta.
Require: n_iter: cantidad de iteraciones.

$i \leftarrow 0$
repeat
  $instancias\_ruidosas \leftarrow []$
  for all $instancia$ **in** $instancias$ **do**
    $k\_vecinos \leftarrow K\ vecinos\ más\ cercanos\ de\ instancia$
    $etiqueta\_k\_vecinos \leftarrow etiqueta\ mayoritaria\ de\ k\_vecinos$
    if $instancia[etiqueta] \neq etiqueta\_k\_vecinos$ **then**
      $instancias\_ruidosas \leftarrow instancia$
    end if
  end for
  if $instancias\_ruidosas \cap instancias = \emptyset$ **then**
    break
  end if
  for all $instancia$ **in** $instancias\_ruidosas$ **do**
    $instancia[etiqueta] = \mathcal{NA}\ \ o\ \ eliminar\ instancia$
  end for
  $i \leftarrow i + 1$
until $i = n\_iter$
return $instancias$

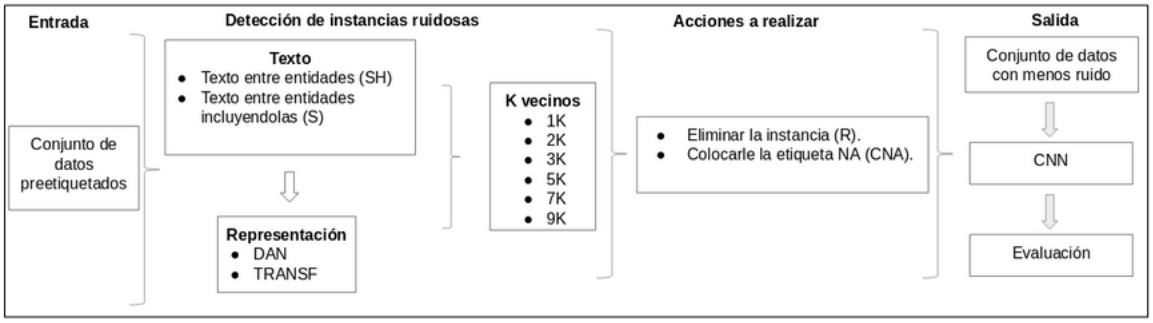


Figura 4.2: Esquema para obtener conjuntos de datos con menos etiquetas ruidosas.

Esta red se entrena con los nuevos conjuntos de datos obtenidos con los diferentes valores de k y diversas configuraciones. En cada época se evalúa con la partición de prueba de GDS para obtener el mejor modelo. Por último, se realiza la evaluación del mejor modelo con la partición de evaluación. En las Figuras 4.4 y 4.5 se muestran las curvas de precisión y recuerdo para los modelos DAN y TRANSF. Esta forma de evaluación asume que las etiquetas del conjunto

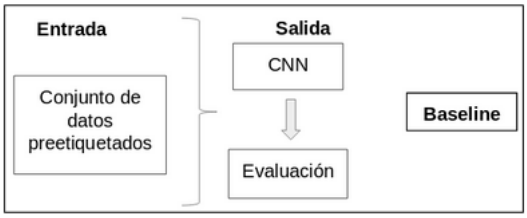


Figura 4.3: *Baseline* para realizar la comparación.

de evaluación son correctas y no presentan etiquetas ruidosas, lo cual no se garantiza.

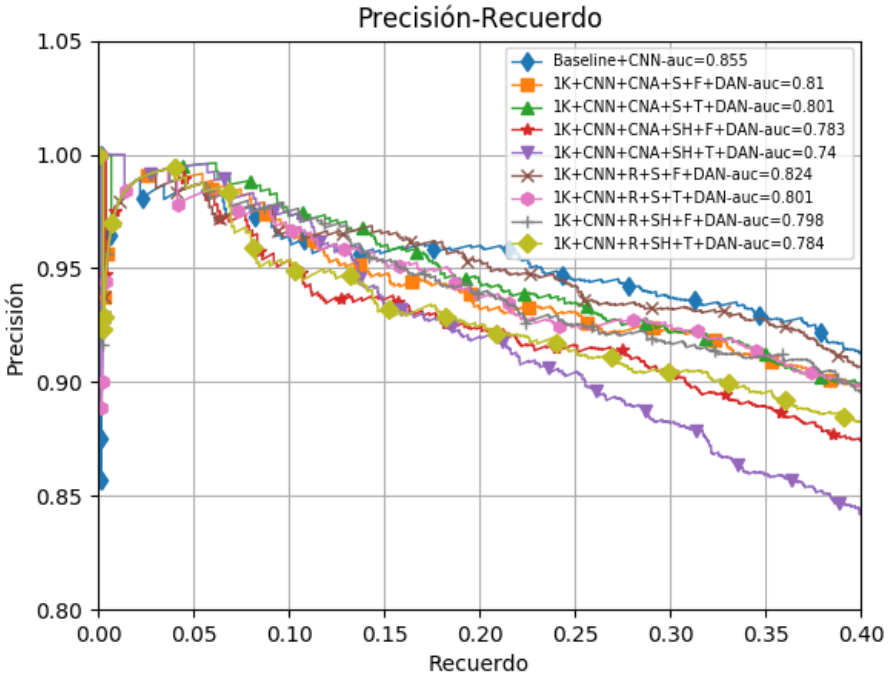


Figura 4.4: Curvas de precisión y recuerdo para diferentes configuraciones y valores de k de la estrategia k vecinos más cercanos utilizando el modelo DAN.

Baseline: se entrenó la red CNN con los datos originales. **CNN:** red utilizada. **#K:** número de vecinos utilizados para reducir el ruido. **CNA:** acción de cambiar a $\mathcal{NA}$ las etiquetas consideradas ruidosas. **R:** acción de remover las instancias consideradas ruidosas. **S:** texto entre entidades incluyéndolas. **SH:** texto entre entidades sin incluirlas. **F:** no incluye las instancias con la relación $\mathcal{NA}$ en el cálculo de los k vecinos. **T:** incluye las instancias con la relación $\mathcal{NA}$ en el cálculo de los k vecinos. **auc:** área bajo la curva.

Ninguna de las configuraciones utilizadas se encuentran por encima del *baseline* tomando como recuerdo 1.0. Sin embargo, si se realiza un corte de este valor sobre 0.4, existen intervalos donde la precisión de estas configuraciones es mayor. Tomando las curvas hasta un

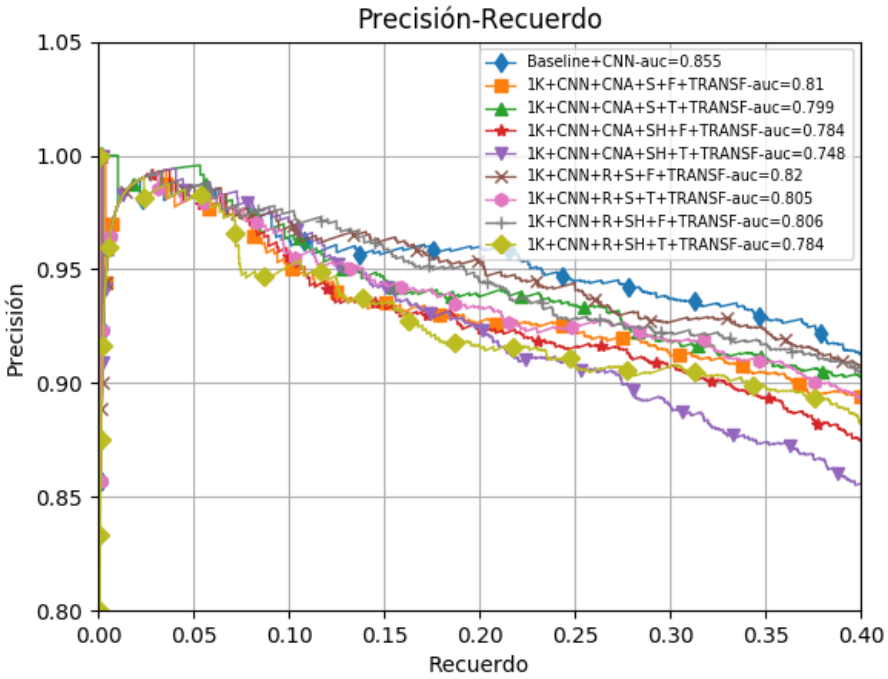


Figura 4.5: Curvas de precisión y recuerdo para diferentes configuraciones y valores de k de la estrategia k vecinos más cercanos utilizando el modelo TRANSF.

Baseline: se entrenó la red CNN con los datos originales. **CNN:** red utilizada. **#K:** número de vecinos utilizados para reducir el ruido. **CNA:** acción de cambiar a $\mathcal{NA}$ las etiquetas consideradas ruidosas. **R:** acción de remover las instancias consideradas ruidosas. **S:** texto entre entidades incluyéndolas. **SH:** texto entre entidades sin incluirlas. **F:** no incluye las instancias con la relación $\mathcal{NA}$ en el cálculo de los k vecinos. **T:** incluye las instancias con la relación $\mathcal{NA}$ en el cálculo de los k vecinos. **auc:** área bajo la curva.

valor de recuerdo de 1.0, los mejores resultados, en cuanto al área bajo la curva, se obtienen por las configuraciones 1K+CNN+R+S+F+DAN y 1K+CNN+R+S+F+TRANSF para los modelos DAN y TRANSF respectivamente. Esto puede indicar que luego de aplicar el filtro de los k vecinos se mejora la precisión tomando valores de recuerdo pequeños. Además, que eliminar las instancias ruidosas, no incluir las instancias con la etiqueta $\mathcal{NA}$ en el cálculo de los k vecinos y el texto entre las dos entidades incluyéndolas tuvieron un área bajo la curva superior a cambiarle la etiqueta a $\mathcal{NA}$, incluir las instancias con etiqueta $\mathcal{NA}$ y tomar el texto entre entidades sin incluirlas respectivamente. Debido a que las etiquetas del conjunto de evaluación pueden contener ruido, las curvas de precisión y recuerdo se utilizan principalmente para obtener una idea del funcionamiento del método y para el ajuste de parámetros (Mintz et al., 2009). Por lo anterior, se realiza una evaluación manual (Mintz et al., 2009) de los

	Baseline		1K+CNN+R+S+F+DAN		1K+CNN+R+S+F+TRANSF	
Relación	Precisión@30	Precisión@50	Precisión@30	Precisión@50	Precisión@30	Precisión@50
perGraduatedInstitution	1.00	1.00	1.00	1.00	1.00	1.00
perHasDegree	**0.23**	**0.22**	**0.17**	**0.16**	**0.23**	**0.18**
perPlaceOfBirth	1.00	0.98	1.00	1.00	1.00	1.00
perPlaceOfDeath	0.97	0.90	0.93	0.96	0.93	0.94
Todas	**0.23**	**0.22**	**0.17**	**0.16**	**0.23**	**0.18**

Tabla 4.10: Evaluación manual por clases del conjunto de evaluación GDS utilizando la medida Precision@K para los valores 30 y 50.

Configuración	**Precisión@50**	**Precisión@100**	**Precisión@200**	**Precisión@300**
Baseline	**0.22**	**0.15**	0.14	0.14
1K+CNN+R+S+F+DAN	0.16	0.13	0.17	0.18
1K+CNN+R+S+F+TRANSF	0.18	0.14	**0.18**	**0.20**

Tabla 4.11: Evaluación manual del conjunto de evaluación GDS utilizando la medida Precision@K para los valores 50, 100, 200 y 300.

resultados utilizando la medida Precision@K (ver Tablas 4.10 y 4.11). Esta evaluación se realiza ordenando las instancias por el valor de confianza que otorga la red neuronal a su predicción.

La Tabla 4.10 muestra que la clase que presenta más ruido en las etiquetas es *perHasDegree* al presentar la red muchas predicciones correctas con respecto a la etiqueta original y sin embargo, luego de una revisión manual, no expresar realmente dicha relación. La precisión, en esta clase, no alcanza el valor de 0.25 en ninguna de las configuraciones en contraste con las otras tres clases que si alcanzan valores cercanos a uno. Dentro de las 50 instancias en las cuáles la red neuronal presenta mayor confianza solo existen instancias de la clase *perHasDegree*. Esto pudiera deberse a que es la clase que menor proporción de pares de entidades presenta con respecto al total de instancias de la misma (ver Tabla 4.2). Esto significa que parte con el menor por ciento de instancias correctas entre las clases, al garantizar los autores que por cada par de entidades se garantiza que al menos una instancia presenta la relación. Por tanto, se puede concluir que la aplicación de filtros específicos para cada clase puede ser de utilidad.

Por otra parte, en la Tabla 4.11 se muestra que, tomando las 100 instancias con mayor confianza por parte de la red, se obtienen valores de precisión con los datos originales superiores a los obtenidos con los datos filtrados. Sin embargo, su precisión en K disminuye para las 200 y 300 primeras instancias obteniéndose con la configuración 1K+CNN+R+S+F+TRANSF el mejor valor. Esto se debe a que la predicción en las 300 primeras instancias del conjunto de datos

original pertenecen a la clase *perHasDegree* que es ruidosa. Esto no ocurre luego de aplicarle filtros a estos datos. Esto justifica la utilidad de la aplicación de filtros en la reducción de ruido. También, se utilizó la red neuronal propuesta por Lin et al. (2016) que se basa en la red PCNN (Zeng et al., 2015) (ver Figura 2.4) e incorpora un mecanismo de atención sobre las instancias (PCNN+ATT). La implementación utilizada fue tomada de *github*[7]. Los parámetros de la red se mantienen como aparecen en el proyecto original. Ejemplo de algunos son el uso del gradiente descendiente, 0.5 como índice de aprendizaje, *softmax* como función de activación y entropía cruzada como función de pérdida. Esta red se entrena de manera similar a lo realizado con la red CNN y se vuelve a asumir que las etiquetas del conjunto de entrenamiento son correctas. En las Figuras 4.6 y 4.7 se muestran las curvas de precisión y recuerdo para los modelos DAN y TRANSF teniendo en cuenta un solo vecino. El comportamiento de esta red es similar al de CNN en cuanto a las mejores configuraciones (ver Anexos E y F).

Por último, en Vashishth et al. (2018) se presentan resultados de los métodos propuestos por Zeng et al. (2015), Lin et al. (2016), Jat et al. (2018) y Vashishth et al. (2018) sobre el conjunto de datos GDS. La Figura 4.8 muestra la comparación de estos métodos con las configuraciones 1K+CNN+R+S+F+DAN, 1K+CNN+R+S+F+TRANSF, 1K+PCNN+ATT+R+S+F+DAN y 1K+PCNN+ATT+R+S+F+TRANSF. Los resultados obtenidos por la implementación de PCNN+ATT realizada por Vashishth et al. (2018) obtiene un área bajo la curva inferior que la utilizada en esta propuesta sobre el conjunto de datos originales. Las configuraciones utilizadas superan los modelos propuestos por Zeng et al. (2015), Lin et al. (2016) y Jat et al. (2018), sin embargo, quedan por debajo del modelo propuesto por Vashishth et al. (2018) y de las redes CNN y PCNN+ATT sobre los datos originales (*baseline*), es decir, sin aplicar filtros. A pesar de esto, este resultado no es concluyente debido que las etiquetas de la partición de evaluación presentan ruido. Sin embargo, muestran que el uso del filtro de los k vecinos más cercanos presenta un buen comportamiento. Otro aspecto es que el uso de una red neuronal que presenta tolerancia al ruido (PCNN+ATT) parece que pudiera presentar un rendimiento superior a una red neuronal sin componentes adicionales con ese objetivo.

Conclusiones del experimento 3

1. El mejor comportamiento se obtiene utilizando uno y dos como valores de k.

2. Eliminar las instancias ruidosas, no incluir las instancias con la etiqueta NA en el cálculo de los k vecinos y el texto entre las dos entidades incluyéndolas presenta un mejor

[7]https://github.com/thunlp/OpenNRE [01/08/2019]

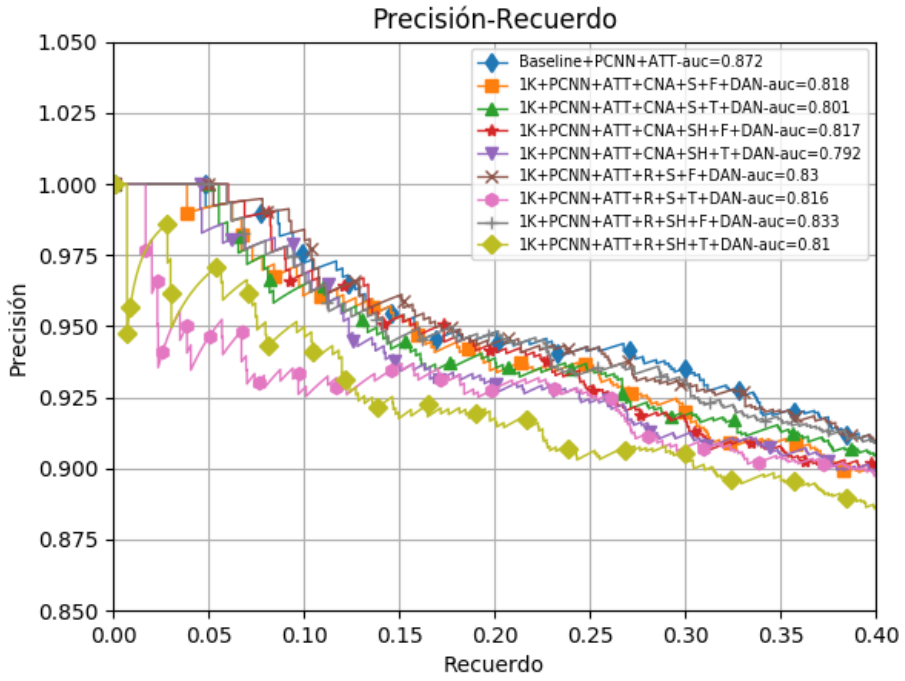


Figura 4.6: Curvas de precisión y recuerdo para diferentes configuraciones y valores de k de la estrategia k vecinos más cercanos utilizando el modelo DAN.

Baseline: se entrenó la red PCNN+ATT con los datos originales. **PCNN+ATT:** red utilizada. **#K:** número de vecinos utilizados para reducir el ruido. **CNA:** acción de cambiar a $\mathcal{NA}$ las etiquetas consideradas ruidosas. **R:** acción de remover las instancias consideradas ruidosas. **S:** texto entre entidades incluyendolas. **SH:** texto entre entidades sin incluirlas. **F:** no incluye las instancias con la relación $\mathcal{NA}$ en el cálculo de los k vecinos. **T:** incluye las instancias con la relación $\mathcal{NA}$ en el cálculo de los k vecinos. **auc:** área bajo la curva.

comportamiento que cambiarle la etiqueta a NA, incluir las instancias con etiqueta NA y tomar el texto entre entidades sin incluirlas respectivamente.

3. Existen clases más ruidosas que otras por lo que pudiera ser útil filtros específicos para cada una de ellas.

4.3. Aplicación de la supervisión distante a un dominio

Para evaluar esta tarea en un dominio de ejemplo se trabaja en la construcción de un conjunto de datos en el área de Espectroscopia funcional del Infrarrojo Cercano (fNIRs). Esto consiste en el etiquetado de las entidades presentes en artículos científicos por parte de especialistas. Luego, se procedería a realizar el etiquetado automático de las relaciones mediante supervisión distante y una ontología del área (base de conocimiento). Lo realizado hasta ahora consiste en

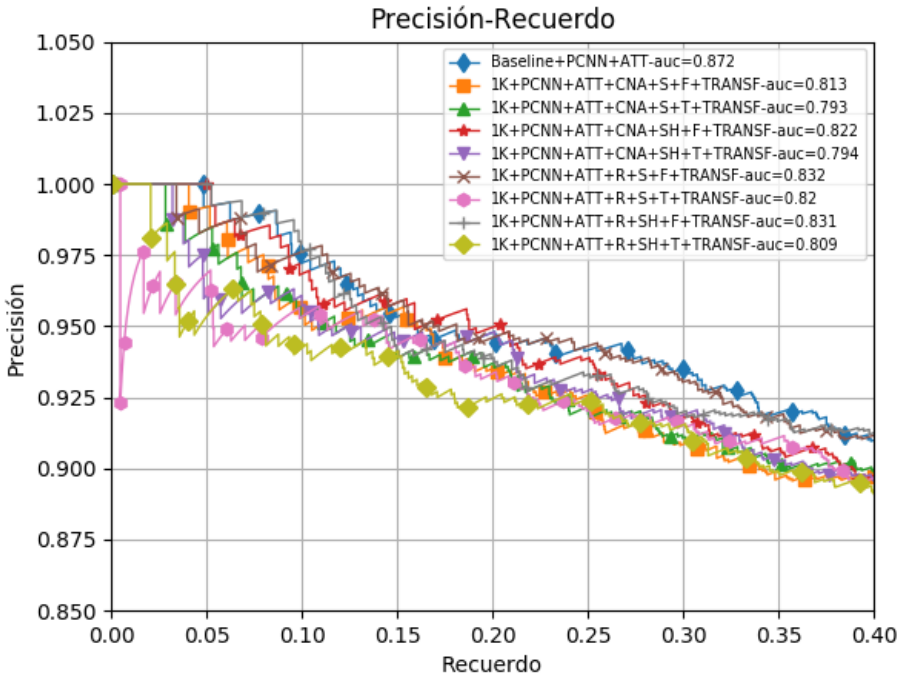


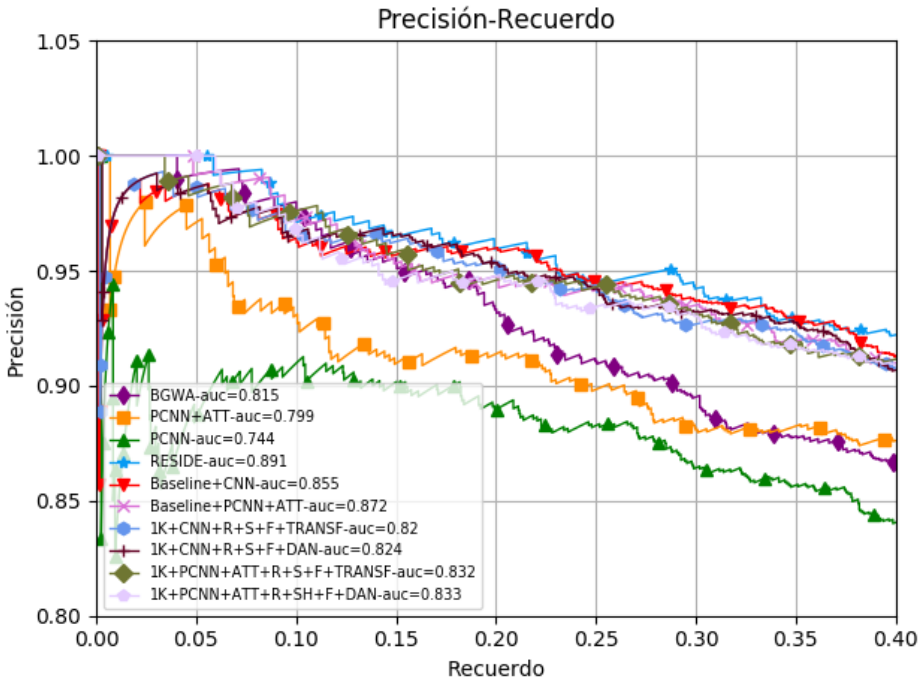
Figura 4.7: Curvas de precisión y recuerdo para diferentes configuraciones y valores de k de la estrategia k vecinos más cercanos utilizando el modelo TRANSF.

Baseline: se entrenó la red PCNN+ATT con los datos originales. **PCNN+ATT:** red utilizada. **#K:** número de vecinos utilizados para reducir el ruido. **CNA:** acción de cambiar a $\mathcal{NA}$ las etiquetas consideradas ruidosas. **R:** acción de remover las instancias consideradas ruidosas. **S:** texto entre entidades incluyendolas. **SH:** texto entre entidades sin incluirlas. **F:** no incluye las instancias con la relación $\mathcal{NA}$ en el cálculo de los k vecinos. **T:** incluye las instancias con la relación $\mathcal{NA}$ en el cálculo de los k vecinos. **auc:** área bajo la curva.

etiquetar entidades de interés en un conjunto de artículos. Esta tarea se encuentra en ejecución con ayuda de un investigador en el área (Dr. Felipe Orihuela Espina). Se escogieron 10 artículos de fNIRS de manera aleatoria relacionados con la parte de instrumentación. A partir de estos artículos se utilizó la herramienta Grobid[8] (Lopez, 2009) para la extracción del texto a partir de archivos PDF. Se probaron otras herramientas como iText[9] y Apache Tika[10] pero la que mejor conservó la estructura del texto original en cuanto a su formato fue Grobid, la cuál se enfoca en publicaciones técnicas y científicas.

A partir del texto extraído, se utilizó TermSuite[11] (Rocheteau and Daille, 2011; Cram and

[8]https://grobid.readthedocs.io/en/latest/ [01/08/2019]
[9]https://itextpdf.com/en [01/08/2019]
[10]https://tika.apache.org/ [01/08/2019]
[11]http://termsuite.github.io/ [01/08/2019]



Baseline: se entrenó la red con los datos originales. **PCNN:** red propuesta por Zeng et al. (2015). **PCNN+ATT:** red propuesta por Lin et al. (2016). **BGWA:** modelo propuesto por Jat et al. (2018). **RESIDE:** propuesto por Vashishth et al. (2018). **#K:** número de vecinos utilizados para reducir el ruido. **CNA:** acción de cambiar a $\mathcal{NA}$ las etiquetas consideradas ruidosas. **R:** acción de remover las instancias consideradas ruidosas. **S:** texto entre entidades incluyéndolas. **SH:** texto entre entidades sin incluirlas. **F:** no incluye las instancias con la relación $\mathcal{NA}$ en el cálculo de los k vecinos. **T:** incluye las instancias con la relación $\mathcal{NA}$ en el calculo de los k vecinos. **auc:** área bajo la curva.

Figura 4.8: Curvas de precisión y recuerdo de diferentes modelos.

Daille, 2016; Daille, 2017) para identificar y extraer términos del mismo. Se filtraron estos términos de conjunto con el especialista en fNIRS y se realizó un preetiquetado de estos términos como posibles entidades de interés. Luego de un primer etiquetado surgió la necesidad de etiquetar solo aquellas entidades de interés para las relaciones predefinidas. En estos momentos se están reetiquetando los documentos..

Capítulo 5

Conclusiones preliminares

A partir de la revisión del estado del arte y de los experimentos realizados durante este periodo, se tienen las siguientes conclusiones preliminares.

- Teniendo en cuenta lo visto hasta ahora, puede darse el caso de que ninguna de las instancias que presente el mismo par de entidades exprese una relación. De ahí que pudiera ser útil relajar aún más la afirmación realizada por Riedel et al. (2010). Varios de los métodos relacionados con la supervisión distante presentan esta desventaja al asumir que al menos una instancia expresa la relación.

- Las representaciones de las sentencias que se han evaluado no logran separar correctamente aquellas con la misma relación de las demás. Esto indica la necesidad de obtener una representación orientada a la extracción de relaciones.

- Otro elemento es que la existencia de clases más ruidosas que otras refleja la utilidad de aplicar filtros independientes para cada unas de ellas.

- El uso de métodos tolerantes al ruido después de aplicar un filtrado a los datos pudiera presentar mayor rendimiento que la aplicación del método sin aplicar filtros o aplicar un algoritmo sin tolerancia.

Por último, los conjuntos de datos que existen para la tarea de supervisión distante presentan algunas deficiencias como instancias duplicadas, errores gramaticales, presencia de varias relaciones en una misma instancia y solo se etiqueta una, el desconocimiento de cuáles son las instancias ruidosas, entre otras. Debido a esto se hace necesario la construcción de un nuevo conjunto de datos que corrija algunas de estas deficiencias priorizando el conocimiento de cuáles instancias son las ruidosas. Esto está motivado porque las curvas de precisión y recuerdo se basan en etiquetas que pueden contener ruido y la precisión en K es muy costosa en conjuntos de datos grandes y depende de quien realice la evaluación. El conocimiento de las instancias ruidosas permitiría medir la tasa de reducción de ruido que presentan los filtros así como el rendimiento del clasificador y una mejor comparación con el estado del arte.

Referencias

Aggarwal, C. C. (2018). *Machine Learning for Text*. Springer International Publishing AG.

Agichtein, E. and Gravano, L. (2000). Snowball: Extracting Relations from large Plain-Text Collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM.

Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M., and Cremers, D. (2018). Clustering with Deep Learning: Taxonomy and New Methods. *arXiv:1801.07648v2 [cs.LG]*, pages 1–12.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. In *Proceedings of KDD*, page 13, Halifax, Canada.

Banko, M., Cafarella, M., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open Information Extraction from the Web. In *International Joint Conferences on Artificial Intelligence*, pages 2670–2676.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, Vancouver, Canada. ACM.

Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., and Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-relational Data. In *Advances in neural information processing systems*, pages 2787–2795.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases*, pages 172–183. Springer, Berlin.

Brodley, C. E. and Friedl, M. A. (1996). Identifying and eliminating mislabeled training instances. In *Proceedings of the National Conference on Artificial Intelligence*, pages 799–805.

Brodley, C. E. and Friedl, M. A. (1999). Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research*, 11:131–167.

Cai, R., Zhang, X., and Wang, H. (2016). Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 756–765, Berlin, Germany. Association for Computational Linguistics.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal Sentence Encoder. *arXiv:1803.11175v2 [cs.CL]*, page 7.

Cram, D. and Daille, B. (2016). Terminology Extraction with Term Variant Detection. In *Proceedings ofthe 54th Annual Meeting ofthe Association for Computational Linguistics—System Demonstrations*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

Daille, B. (2017). *Term Variation in Specialised Corpora*, volume 19 of *Terminology and Lexicography Research and Practice*. John Benjamins Publishing Company, Amsterdam.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, pages 233–240, Pittsburgh, Pennsylvania, USA. ACM Press.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Advances in neural information processing systems*, pages 3844–3852.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805v1 [cs.CL]*.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 165(1):91–134.

Etzioni, O., Fader, A., Christensen, J., Soderland, S., and Mausam (2011). Open information extraction: The second generation. *International Joint Conferences on Artificial Intelligence*, 11:3–10.

Frénay, B. and Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869.

Gábor, K., Buscaldi, D., Schumann, A.-K., QasemiZadeh, B., Zargayouna, H., and Charnois, T. (2018). Semeval-2018 Task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Goyal, A., Gupta, V., and Kumar, M. (2018). Recent Named Entity Recognition and Classification techniques: A systematic review. *Computer Science Review*, 29:21–43.

Grishman, R. (2015). Information Extraction. *IEEE Intelligent Systems*, 30(5):8–15.

Guo, X., Zhang, H., Liu, R., Ding, X., Tian, R., and Wang, B. (2018). Attention-Based Combination of CNN and RNN for Relation Classificatio. In *25th International Conference Neural Information Processing, ICONIP 2018*, pages 244–255. Springer Nature Switzerland AG.

Guo, X., Zhang, H., Yang, H., Xu, L., and Ye, Z. (2019). A Single Attention-Based Combination of CNN and RNN for Relation Classification. *IEEE Access*, 7:12467–12475.

He, Z., Chen, W., Li, Z., Zhang, M., Zhang, W., and Zhang, M. (2018). SEE: Syntax-Aware Entity Embedding for Neural Relation Extraction. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5795–5802. Association for the Advancement of Artificial Intelligence.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational Linguistics (COLING-92)*, pages 539–545, Nantes.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Séaghdha, D. Ó., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). SemEval-2010 Task 8 : Multi-Way Classification of Semantic Relations Between Pairs of Nominals. In *Proceedings ofthe 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting ofthe Association for Computational Linguistics*, pages 541–550, Portland, Oregon. Association for Computational Linguistics.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. (2015). Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Proceedings of the 53rd Annual Meeting ofthe Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Jat, S., Khandelwal, S., and Talukdar, P. (2018). Improving Distantly Supervised Relation Extraction using Word and Entity Based Attention. *arXiv:1804.06987v1 [cs.CL]*.

Ji, G., Liu, K., He, S., and Zhao, J. (2017). Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 3060–3066.

Keilwagen, J., Grosse, I., and Grau, J. (2014). Area under Precision-Recall Curves for Weighted and Unweighted Data. *PLoS ONE*, 9(3):e92209.

Kim, J.-T. and Moldovan, D. I. (1993). Acquisition of semantic patterns for information extraction from corpora. In *Proceedings of 9th IEEE Conference on Artificial Intelligence for Applications*, pages 171–176. IEEE.

Lee, J., Seo, S., and Choi, Y. S. (2019). Semantic Relation Classification via Bidirectional LSTM Networks with Entity-aware Attention using Latent Entity Typing. *arXiv:1901.081163v1 [cs.CL]*.

Lin, Y., Shen, S., Liu, Z., Luan, H., and Sun, M. (2016). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting ofthe Association for Computational Linguistics*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

Liu, K. and El-Gohary, N. (2017). Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. *Automation in Construction*, 81:313–327.

Liu, T., Wang, K., Chang, B., and Sui, Z. (2017). A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1790–1795, Copenhagen, Denmark.

Liu, Y., Liu, K., Xu, L., and Zhao, J. (2014). Exploring fine-grained entity type constraints for distantly supervised relation extraction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2107–2116, Dublin, Ireland.

Liu, Y., Wei, F., Li, S., Ji, H., Zhou, M., and Wang, H. (2015). A Dependency-Based Neural Network for Relation Classification. *arXiv:1507.04646v1 [cs.CL]*.

Lopez, P. (2009). GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *International Conference on Theory and Practice of Digital Libraries (ECDL)*, pages 473–474. Springer-Verlag Berlin Heidelberg.

Mahdisoltani, F., Biega, J., and Suchanek, F. M. (2015). Yago3: A Knowledge Base from Multilingual Wikipedias. In *7th Biennial Conference on Innovative Data Systems Research (CIDR 2015)*, Asilomar, California, USA.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Matic, N., Guyon, I., Bottou, L., Denker, J., and Vapnik, V. (1992). Computer aided cleaning of large databases for character recognition. In *11th IAPR International Conference on Pattern Recognition. Vol. II. Conference B: Pattern Recognition Methodology and Systems*, pages 330–333. IEEE.

Mausam, Schmitz, M., Bart, R., Soderland, S., and Etzioni, O. (2012). Open Language Learning for Information Extraction. In *Proceedings ofthe 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, Korea. Association for Computational Linguistics.

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., and Long, J. (2018). A Survey of Clustering with Deep Learning: From the Perspective of Network Architecture. *IEEE Access*, 6:39501–39514.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the ACL*, pages 1003–1011, Suntec, Singapore.

Miranda, A. L. B., Garcia, L. P. F., Carvalho, A. C. P. L. F., and Lorena, A. C. (2009). Use of Classification Algorithms in Noise Detection and Elimination. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 417–424. Springer-Verlag Berlin Heidelberg.

Miwa, M. and Bansal, M. (2016). End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. *arXiv:1601.00770v3 [cs.CL]*.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nguyen, T. H. and Grishman, R. (2015). Relation extraction: Perspective from convolutional neural networks. In *Proceedings of NAACL-HLT 2015*, pages 39–48, Denver, Colorado. Association for Computational Linguistics.

Nogueira dos Santos, C., Xiang, B., and Zhou, B. (2015). Classifying Relations by Ranking with Convolutional Neural Networks. *arXiv:1504.06580v2 [cs.CL]*.

Pal, H. and Mausam (2016). Demonyms and Compound Relational Nouns in Nominal Open IE. In *Proceedings of AKBC 2016*, pages 35–39, San Diego, California. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365v2 [cs.CL]*.

Piskorski, J. and Yangarber, R. (2013). Information extraction: Past, Present and Future. In *Multi-source, Multilingual Information Extraction and Summarization 11*, pages 23–49. Springer-Verlag Berlin Heidelberg.

Qin, P., Xu, W., and Guo, J. (2016). An empirical convolutional neural network approach for semantic relation classification. *Neurocomputing*, 190:1–9.

Qin, P., Xu, W., and Wang, W. Y. (2018a). DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction. *arXiv:1805.09929v1 [cs.CL]*.

Qin, P., Xu, W., and Wang, W. Y. (2018b). Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2137–2147.

Riedel, S., Yao, L., and McCallum, A. (2010). Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin. Springer.

Rillof, E. (1996). Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the national conference on artificial intelligence*, pages 1044–1049.

Rink, B. and Harabagiu, S. (2010). UTD: Classifying Semantic Relations by Combining Lexical and Semantic Resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 256–259, Uppsala, Sweden. Association for Computational Linguistics.

Rocheteau, J. and Daille, B. (2011). TTC TermSuite: A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 9–12, Chiang Mai, Thailand. Asian Federation of Natural Language Proceesing.

Rotsztejn, J., Hollenstein, N., and Zhang, C. (2018). ETH-DS3Lab at SemEval-2018 Task 7: Effectively Combining Recurrent and Convolutional Neural Networks for Relation Classification and Extraction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 689–696.

Ru, C., Tang, J., Li, S., Xie, S., and Wang, T. (2018). Using semantic similarity to reduce wrong labels in distant supervision for relation extraction. *Information Processing & Management*, 54(4):593–608.

Saha, S., Pal, H., and Mausam (2017). Bootstrapping for Numerical Open IE. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 317–323, Vancouver, Canada. Association for Computational Linguistics.

Sarawagi, S. (2007). Information extraction. *Foundation and Trends in Databases*, 1(3):261–377.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47.

Shen, Y. and Huang, X. (2016). Attention-Based Convolutional Neural Network for Semantic Relation Extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2526–2536, Osaka, japan.

Smirnova, A. and Cudré-Mauroux, P. (2018). Relation Extraction Using Distant Supervision: A Survey. *ACM Computing Surveys*, 51(5):1–35.

Snow, R., Jurafsky, D., and Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In *Advances in neural information processing systems*, pages 1297–1304.

Socher, R., Huval, B., Manning, C. D., and Ng, A. Y. (2012). Semantic Compositionality through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211, Jeju Island, Korea. Association for Computational Linguistics.

Soderland, S. (1999). Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34:233–272.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.

Surdeanu, M., Tibshirani, J., Nallapati, R., and Manning, C. D. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea. Association for Computational Linguistics.

Takamatsu, S., Sato, I., and Nakagawa, H. (2012). Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 721–729, Jeju, Republic of Korea. Association for Computational Linguistics.

Tan, P.-N., Steinbach, M., and Kumar, V. (2014). *Data mining cluster analysis: basic concepts and algorithms*. Pearson.

Vashishth, S., Joshi, R., Prayaga, S. S., Bhattacharyya, C., and Talukdar, P. (2018). Reside: Improving Distantly-Supervised Neural Relation Extraction using Side Information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and Kaiser, L. (2017). Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008, Long Beach, CA, USA.

Vo, D. T. and Bagheri, E. (2018). Self-training on refined clause patterns for relation extraction. *Information Processing and Management*, 54(4):686–706.

Wang, G., Zhang, W., Wang, R., Zhou, Y., Chen, L., Zhang, W., Zhu, H., and Chen, H. (2018). Label-free distant supervision for relation extraction via knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2246–2255.

Wang, L., Cao, Z., de Melo, G., and Liu, Z. (2016). Relation Classification via Multi-Level Attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1298–1307, Berlin, Germany. Association for Computational Linguistics.

Wu, S., Fan, K., and Zhang, Q. (2018). Improving Distantly Supervised Relation Extraction with Neural Noise Converter and Conditional Optimal Selector. *arXiv e-prints*.

Wu, S. and He, Y. (2019). Enriching Pre-trained Language Model with Entity Information for Relation Classification. *arXiv:1905.082844v1 [cs.CL]*.

Xiao, M. and Liu, C. (2016). Semantic relation classification via hierarchical recurrent neural network with attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1254–1263, Osaka, Japan.

Xu, P. and Barbosa, D. (2019). Connecting Language and Knowledge with Heterogeneous Representations for Neural Relation Extraction. *arXiv:1903.10126v3 [cs.CL]*.

Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., and Jin, Z. (2016). Improved Relation Classification by Deep Recurrent Neural Networks with Data Augmentation. *arXiv:1601.03651v2*.

Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., and Jin (2015). Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1785–1794, Lisbon, Portugal. Association for Computational Linguistics.

Ye, Z.-X. and Ling, Z.-H. (2019). Distant Supervision Relation Extraction with Intra-Bag and Inter-Bag Attentions. *arXiv:1904.00143v1 [cs.CL]*.

Yu, M., Gormley, M., and Dredze, M. (2014). Factor-based Compositional Embedding Models. In *NIPS Workshop on Learning Semantics*, pages 95–101.

Zeng, D., Liu, K., Chen, Y., and Zhao, J. (2015). Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.

Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J. (2014). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland.

Zhang, S., Zheng, D., Hu, X., and Yang, M. (2015). Bidirectional Long Short-Term Memory Networks for Relation Classification. In *29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78, Shanghai, China.

Zhang, X., Chen, F., and Huang, R. (2018). A Combination of RNN and CNN for Attention-based Relation Classification. *Procedia Computer Science*, 131:911–917.

Zheng, S., Xu, J., Zhou, P., Bao, H., Qi, Z., and Xu, B. (2016). A neural network framework for relation extraction: Learning entity semantic and relation pattern. *Knowledge-Based Systems*, 114:12–23.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

Zhou, P., Xu, J., Qi, Z., Bao, H., Chen, Z., and Xu, B. (2018). Distant supervision for relation extraction with hierarchical selective attention. *Neural Networks*, 108:240–247.

Zhu, J., Qiao, J., Dai, X., and Cheng, X. (2017). Relation Classification via Target-Concentrated Attention CNNs. In *24th International Conference Neural Information Processing, ICONIP 2017*, pages 137–146. Springer International Publishing AG.

Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Anexos A

Aprendizaje automático multi-instancia

En el aprendizaje multi-instancia (MIL) asociado a la supervisión distante, todas las sentencias con el mismo par de entidades constituyen una bolsa y cada sentencia o texto formado por varias sentencias dentro de esta es una instancia (Liu and El-Gohary, 2017; Ji et al., 2017).
Sea el conjunto de entrenamiento de N bolsas $B_1, B_2, \ldots, B_N$ y la i-ésima bolsa contiene q_i instancias $B_i = b_1^i, b_2^i, \ldots, b_{q_i}^i (i = 1, \ldots, N)$. El objetivo de este enfoque es predecir las etiquetas (relaciones) de las bolsas no vistas.

Anexos B

Coeficiente silhouette

El coeficiente silhouette combina la cohesión y la separación de los grupos (Tan et al., 2014). La cohesión se refiere a la compactación de los grupos, es decir, cuán cerca están los elementos de un mismo grupo. La separación es el aislamiento, es decir, cuán bien está separado un grupo de los demás. Las Ecuaciones B.1 y B.2 indican como calcular el coeficiente silhouette para un elemento y para todos respectivamente.

$$
\begin{aligned}
silhouette_i &= \frac{(b_i - a_i)}{max(a_i, b_i)} \\
-1 \leq silhouette_i &\leq 1 \\
silhouette_i < 0 &\Rightarrow a_i > b_i, \\
silhouette_i > 0 &\Rightarrow a_i < b_i, \\
silhouette_i = 1 &\Rightarrow a_i = 0
\end{aligned}
\tag{B.1}
$$

donde i se refiere al i-ésimo objeto, a_i el promedio de las distancias del i-ésimo objeto a los elementos restantes de su grupo y b_i el promedio de las distancias del i-ésimo objeto a los elementos en otro grupo.

$$
silhouette = \frac{\sum_{i=1}^{N} silhouette_i}{N} \tag{B.2}
$$

donde N es la cantidad de elementos.

Anexos C

Formas de evaluación

C.1. Precisión, recuerdo y medida F

En los problemas de clasificación la decisión hecha por este se puede representar mediante una matriz de confusión o tabla de contingencia (Davis and Goadrich, 2006; Sebastiani, 2002). Esta estructura presenta 4 categorías: verdaderos positivos (TP), son ejemplos positivos etiquetados como positivos, falsos positivos (FP), se refiere a ejemplos negativos etiquetados de manera errónea como positivos, verdaderos negativos (TN), se corresponde a instancias negativas etiquetadas como negativas y falsos negativos (FN) son ejemplos positivos incorrectamente etiquetados como negativos.
La precisión se corresponde con la proporción entre las instancias correctamente etiquetadas como positivas sobre la suma de las instancias etiquetadas como positivas (ver Ecuación C.1) (Davis and Goadrich, 2006; Sebastiani, 2002).

$$Precisión = \frac{TP}{TP + FP} \tag{C.1}$$

De manera similar, el recuerdo consiste en la proporción entre las instancias correctamente etiquetadas como positivas sobre la suma de las instancias que realmente pertenecen a la clase positiva (ver Ecuación C.2) (Davis and Goadrich, 2006; Sebastiani, 2002).

$$Recuerdo = \frac{TP}{TP + FN} \tag{C.2}$$

Una medida que combina, mediante la media armónica, la precisión y el recuerdo es la F_1 (ver Ecuación C.3) (Sebastiani, 2002).

$$F_1 = \frac{Precisión * Recuerdo}{Precisión + Recuerdo} \tag{C.3}$$

C.2. Curvas de precisión y recuerdo

Las curvas de precisión y recuerdo grafican la precisión frente al recuerdo para diferentes umbrales (Keilwagen et al., 2014). Estas curvas se calculan a partir de la etiqueta verdadera y una puntuación dada por el clasificador. Se utilizan generalmente en problemas de clasificación binarios (Davis and Goadrich, 2006) y en recuperación de información (Manning et al., 2008). A pesar de esto, se utilizan con frecuencia en la supervisión distante (Mintz et al., 2009; Riedel et al., 2010; Zeng et al., 2015; Lin et al., 2016; Ji et al., 2017; Liu et al., 2017; He et al., 2018; Wang et al., 2018; Qin et al., 2018b; Ru et al., 2018; Jat et al., 2018; Wu et al., 2018; Ye and Ling, 2019). Estas curvas en la supervisión distante son un intento de medir el comportamiento de cada uno de los métodos. Sin embargo, no son concluyentes en cuanto a cuál método es

mejor debido a que se basan en las etiquetas originales, las cuáles pueden presentar ruido.

C.3. Precisión en K

La precisión en K (Precisión@K) (ver Ecuación C.4) mide la cantidad de elementos correctos en una ventana de K elementos (Manning et al., 2008). Se utiliza con frecuencia en recuperación de información para medir la precisión en K documentos recuperados. Según Manning et al. (2008), tiene la ventaja de no requerir ninguna estimación del tamaño del conjunto de documentos relevantes. Se ha utilizado en la supervisión distante por Zeng et al. (2015), Lin et al. (2016), Ji et al. (2017) He et al. (2018), Wang et al. (2018), Wu et al. (2018) y Ye and Ling (2019).

$$Precisión@K = \frac{|elementos\ correctos \cap K\ elementos|}{|K\ elementos|} \tag{C.4}$$

Anexos D

Instancias por clase detectadas como ruidosas.

Relación	K=0	K=1	K=2	K=3	K=5	K=7	K=9
perGraduatedInstitution	3091	2159 (-932)	2159 (-932)	2106 (-985)	2091 (-1000)	2072 (-1012)	2047 (**-1044**)
perHasDegree	2075	1674 (-401)	1674 (-401)	1654 (**-421**)	1657 (-418)	1669 (-406)	1690 (-385)
perPlaceOfBirth	2324	1472 (-852)	1472 (-852)	1375 (-949)	1381 (-943)	1384 (-940)	1365 (**-959**)
perPlaceOfDeath	2453	1930 (-523)	1930 (-523)	1878 (-575)	1805 (-648)	1813 (-640)	1793 (**-660**)
$\mathcal{NA}$	3218	5926 (+2708)	5926 (+2708)	6148 (+2930)	6227 (+3009)	6216 (+2998)	6266 (**+3048**)
Total	13161	13161	13161	13161	13161	13161	13161

Teniendo en cuenta la representación DAN, tomando el texto entre entidades incluyéndolas y sin incluir las instancias con la clase $\mathcal{NA}$ para el cálculo de los k vecinos.
(-) Instancias detectadas como ruidosas. **(+)** Instancias que se añaden a la clase $\mathcal{NA}$ en caso de utilizar esta estrategia.

Tabla D.1: Instancias por clase luego de aplicar el filtro para diferentes valores de k vecinos más cercanos.

Anexos E

Curvas de precisión y recuerdo de la red CNN

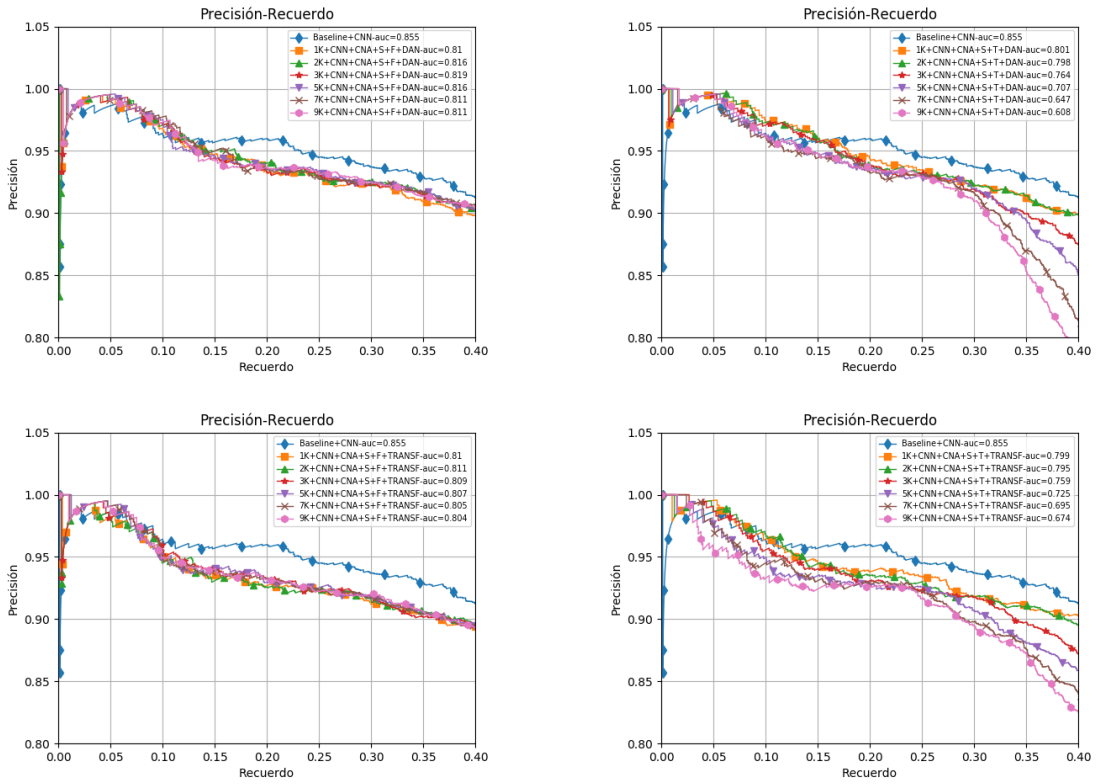


Figura E.1: Curvas de precisión y recuerdo de la red CNN sobre GDS utilizando la acción de cambiar a $\mathcal{NA}$ las etiquetas ruidosas.

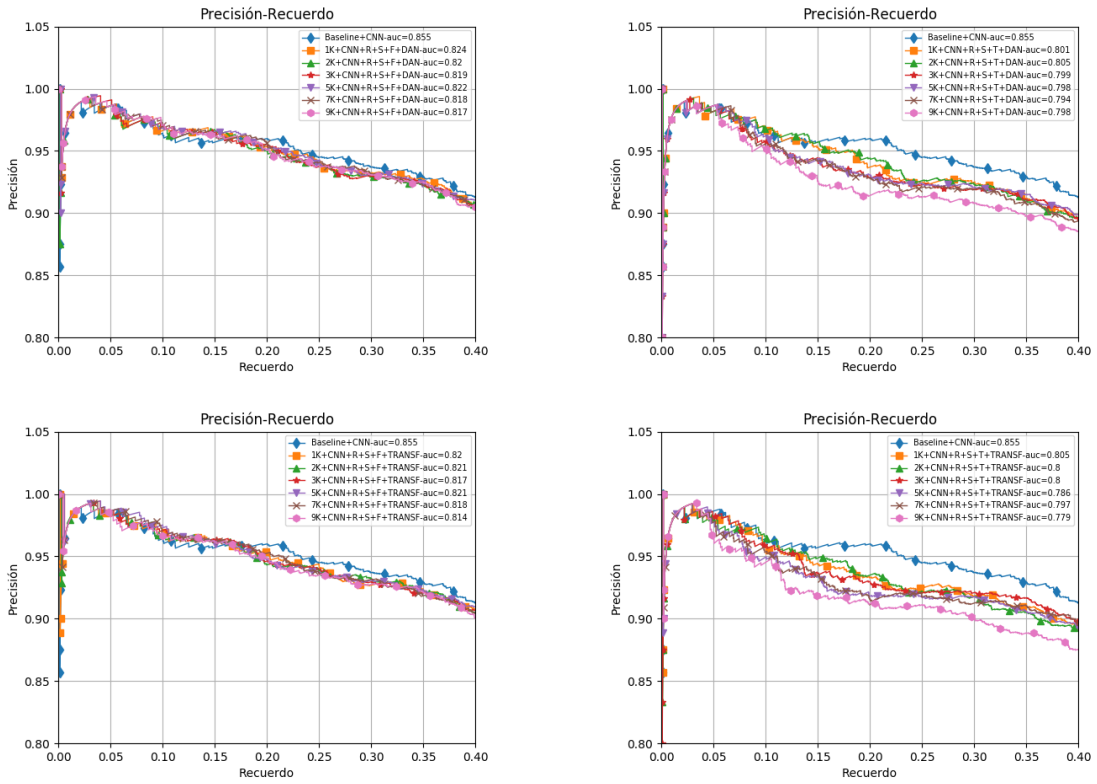


Figura E.2: Curvas de precisión y recuerdo de la red CNN sobre GDS utilizando la acción de eliminar las etiquetas ruidosas.

Anexos F

Curvas de precisión y recuerdo de la red PCNN+ATT

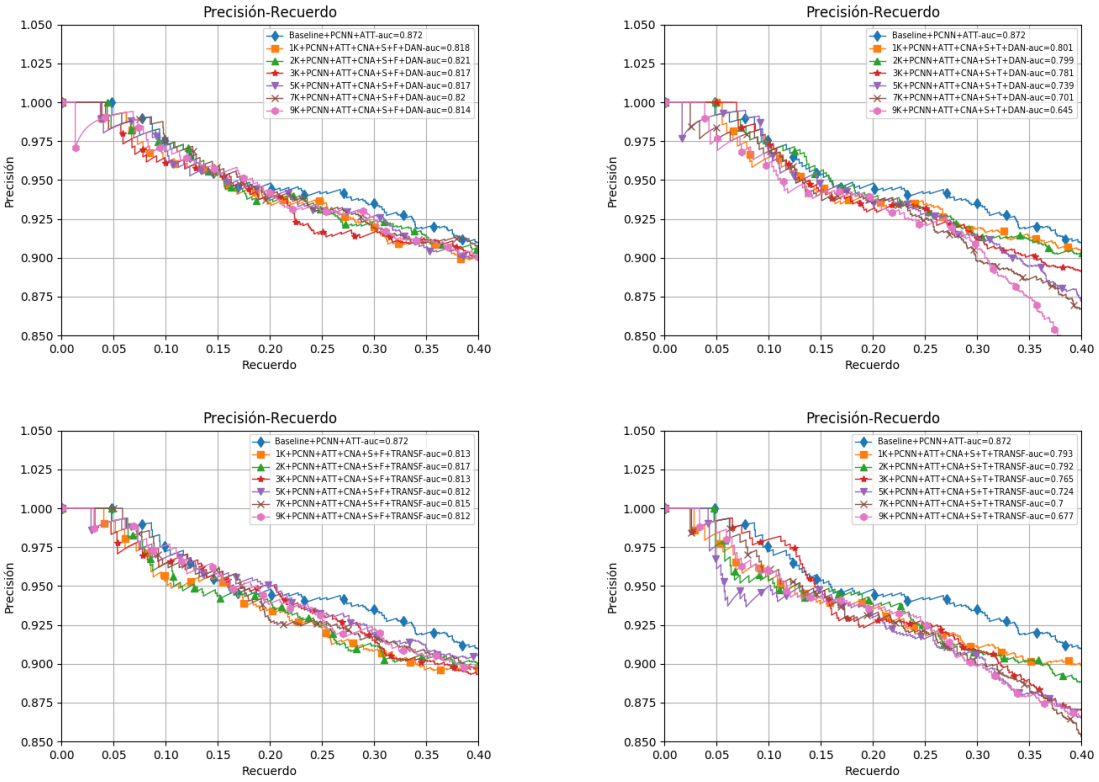


Figura F.1: Curvas de precisión y recuerdo de la red PCNN+ATT sobre GDS utilizando la acción de cambiar a $\mathcal{NA}$ las etiquetas ruidosas.

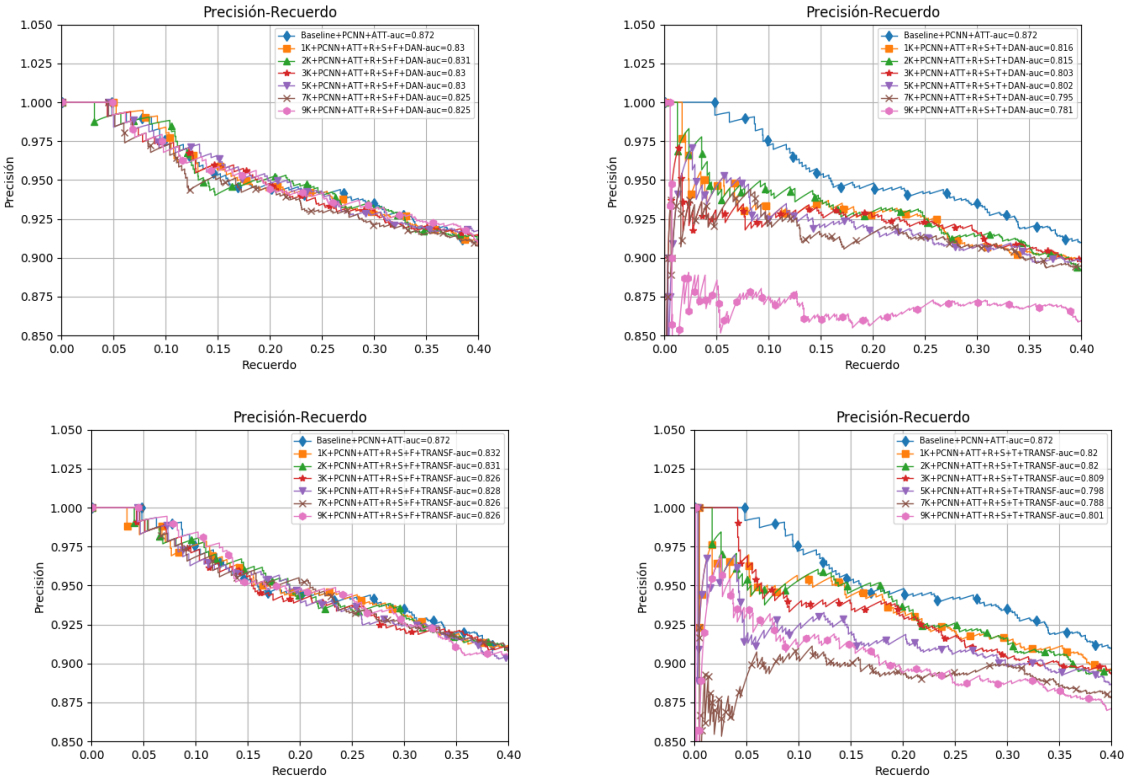


Figura F.2: Curvas de precisión y recuerdo de la red PCNN+ATT sobre GDS utilizando la acción de eliminar las etiquetas ruidosas.