UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

ENLAZANDO TWITTER CON WIKIDATA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL EN COMPUTACIÓN

JAVIERA PAOLA DÍAZ MIZUNUMA

PROFESOR GUÍA:
AIDAN HOGAN

MIEMBROS DE LA COMISIÓN:
ANDRÉS ABELIUK KIMELMAN
HERNÁN SARMIENTO ALBORNOZ

Este trabajo ha sido parcialmente financiado por Departamento de Ciencias de la Computación, FCFM, Centro de Costos 1618

SANTIAGO DE CHILE
2024

Resumen

TelarKG es una base de datos orientada a grafos que contiene información referente al proceso constituyente en Chile del año 2021. Entre los datos que almacena se encuentran publicaciones de *Twitter* (tuits) y entidades de convencionales constituyentes de ese periodo, entre otros.

Al explorar la magnitud de 20 millones de tuits almacenados, se revela un desafío: únicamente un pequeño porcentaje establece conexiones dentro de la base de datos. Los tuits enlazados son aquellos publicados por miembros de la convención. En cuanto a los tuits restantes, aunque compartan una temática común, no poseen enlaces identificables. Esta falta de vinculación plantea un desafío significativo para el análisis de los datos almacenados, limitando la capacidad de investigación.

La propuesta de solución consiste en enlazar los tuits con entidades reconocidas procedentes de otra fuente de datos, como *Wikidata*. Esta última dispone de datos y metadatos relevantes para el caso de estudio, incluyendo información sobre los miembros de la convención constitucional, partidos políticos chilenos y lugares geográficos.

Con el propósito de enriquecer *TelarKG* mediante la generación de enlaces en los tuits, se aplica sobre estos *entity linking*, una tarea del Procesamiento de Lenguaje Natural, que genera enlaces entre texto y alguna fuente de datos, como lo es *Wikidata*.

Esta tarea enfrenta dos desafíos fundamentales: la considerable magnitud de datos a procesar y la escasa contextualización proporcionada en los tuits, generando casos desafiantes para el proceso de *entity linking*. Se investiga la viabilidad de etiquetar la totalidad de los datos y se realiza una evaluación de la calidad de las etiquetas identificadas.

Estos resultados se incorporan a *TelarKG* en forma de 71.590 nuevas entidades y 29.311.087 enlaces con tuits. Esta adición posibilita la ejecución de consultas específicas sobre los tuits almacenados en *TelarKG*, ofreciendo la capacidad de realizar consultas más complejas en comparación con el buscador de *Twitter* o con la base de datos de *TelarKG* por sí solos.

Dedicado a la niña que ama explorar.

Agradecimientos

Quisiera expresar mi profundo agradecimiento a Aidan Hogan, cuyo excepcional liderazgo y calidad humana fueron fundamentales para la realización de este proyecto. Su dedicación, guía y paciencia volvieron posible este trabajo.

Mi reconocimiento destacado al Instituto Milenio Fundamento de los Datos, y a todas las personas que han contribuido desde allí. Además de proporcionar el caso de uso para este proyecto, su disposición sobresaliente para brindar ayuda y aportar información crucial ha sido invaluable.

Agradezco sinceramente al cuerpo docente con el que tuve el placer de recibir clases, por compartir sus valiosos conocimientos. Asimismo, quiero expresar mi profundo agradecimiento a mis amigos, quienes convirtieron cada desafío de aprendizaje en una experiencia divertida.

Agradezco de manera especial a mi madre, Javiera, Diego, Daniela y a todos aquellos que me ayudaron a comprenderme, desafiarme y evolucionar con amor y paciencia. Su influencia personal y orientación han sido esenciales para la finalización de este trabajo.

Este proyecto no habría sido posible sin el respaldo y contribuciones de todas estas personas y entidades. Gracias por ser parte fundamental en este proceso.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

Las redes sociales han tenido un alcance masivo durante aproximadamente dos décadas. El contenido generado en estas plataformas, que mayormente consiste en opiniones de individuos comunes, ofrece una oportunidad única para estudiar tanto a la población en general como fenómenos que puedan afectarla. La capacidad de comprender las tendencias y opiniones de una población sobre determinados temas se vuelve factible cuando se cuenta con una cantidad suficiente de datos y las herramientas adecuadas para analizarlos. Si bien la recopilación de datos es una tarea alcanzable, el análisis manual de grandes volúmenes de información resulta poco práctico. Por ende, el uso de herramientas computacionales para clasificar el contenido de las redes sociales se presenta como una solución lógica para facilitar su análisis. Este enfoque es precisamente el que inspira el presente trabajo.

En esta memoria se busca implementar una estructura de datos, enriquecida con publicaciones de *Twitter*[1] e información relevante obtenida desde *Wikidata*, que permita realizar consultas avanzadas. En particular se busca aplicar los resultados de este trabajo en el contexto del proyecto *TelarKG* del Instituto Milenio Fundamentos de los Datos [2].

Twitter es un servicio de microblogueo con 528,3 millones de usuarios activos al mes durante el año 2023.[2] En Chile se contabilizaron alrededor de 3,6 millones de usuarios al inicio del 2023.[3] En este servicio cada usuario tiene la posibilidad de realizar publicaciones de texto plano con un máximo de 280 caracteres, desde ahora tuits (en inglés, *tweets*). En *Twitter*, los *hashtags* sirven para agrupar contenido relacionado, permitiendo la organización de tuits que abordan el mismo tema. Un *hashtag* consiste en una serie de palabras sin espacios, precedida por el signo numeral.

Durante la última década se ha observado el alcance social de *Twitter*. Se utilizó para organizar protestas durante 2011 en el marco de la revolución egipcia y la revolución tunecina. Durante el estallido social en Chile (2019) *Twitter* fue una de las plataformas donde se organizaron diversas manifestaciones, fue utilizada como fuente de información alternativa,

[1] El 23 de Julio del 2023 *Twitter* fue renombrado a X. Como la mayor parte del tiempo de realización de este trabajo se le conoció con el primer nombre, nos referiremos en esta memoria, con el nombre de *Twitter* a esta red social.

[2] Ver `https://www.demandsage.com/twitter-statistics/`

[3] Ver `https://datareportal.com/reports/digital-2023-chile`

y fue uno de los canales donde la población expresaba su opinión respecto a las medidas adoptadas por el gobierno a través de tuits.[4]

Wikidata es una base de conocimiento alojada por la Fundación *Wikimedia* y es editada de manera colaborativa [24]. *Wikidata* es donde se encuentra el almacenamiento central de los datos utilizados por diversos proyectos de la fundación *Wikimedia* (por ejemplo *Wikipedia*), así como también por sitios y servicios independientes de esta, ya que posee un dominio público. Un ejemplo de entidad de *Wikidata* se muestra en la Figura 1.1. Se observa la entidad cuyo identificador es *Q320* y representa a Michelle Bachelet. Esta posee diferentes atributos como: imagen, género, nombre, apellido, cargo(s) ocupado(s), entre otros.

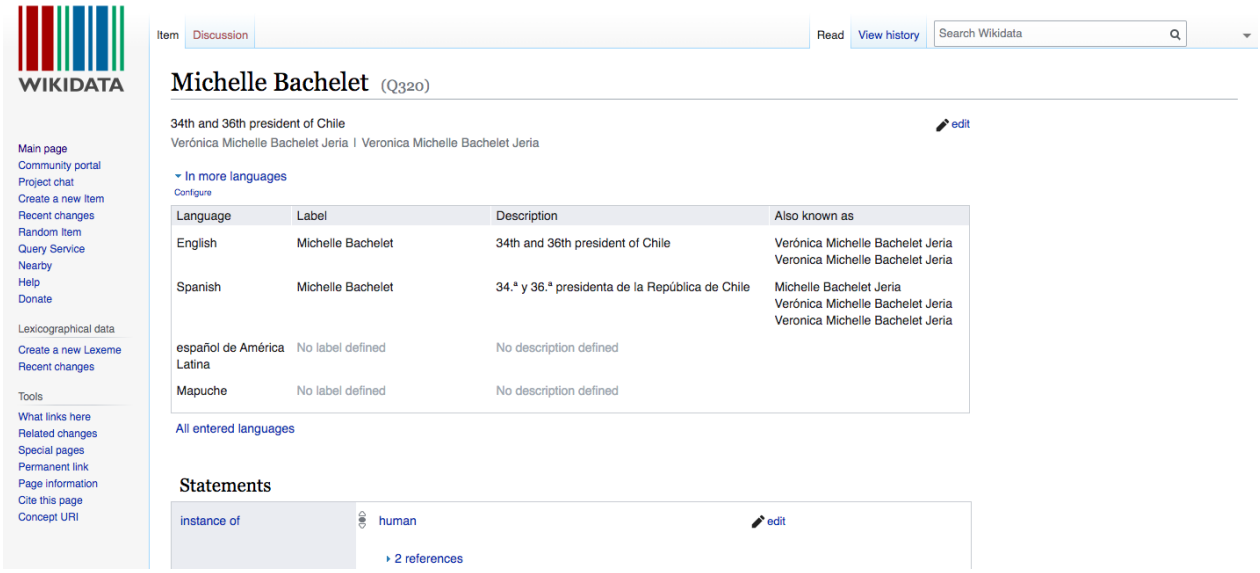

Figura 1.1: Ejemplo de entidad de Michelle Bachelet en *Wikidata*

En el marco del proceso constituyente y aprovechando todas las fuentes de datos disponibles, *TelarKG* es un grafo de conocimiento que contiene información sobre la convención constitucional chilena, como por ejemplo, material discutido, videos y transcripciones. Además, posee información sobre los miembros de la convención, incluyendo sus partidos, pactos, género, entre otros. También posee las cuentas en redes sociales de los convencionales, las cuales, a su vez, están relacionadas con publicaciones en redes sociales, incluyendo tuits de *Twitter*. *TelarKG* es un proyecto del Instituto Milenio Fundamentos de los Datos (desde ahora IMFD), el cual es un centro científico donde se lleva a cabo investigación en materia de datos [2].

Dado que *Twitter* es una basta fuente de información actual y opiniones personales, enlazar estos contenidos con aquellos obtenidos desde *Wikidata* permitirá obtener información apta para el estudio de ciencias sociales, pudiendo consultar sobre el tema particular que se desea estudiar. *Twitter* ha sido la fuente de información de diversos estudios. Por ejemplo, se han medido los sentimientos de un vecindario para predecir la salud de la población que lo reside [9]. Otro ejemplo es un trabajo que busca predecir la geolocalización de usuarios de *Twitter* basada en el texto de tuits [10].

[4]Ver `https://www.elmostrador.cl/agenda-pais/2019/10/21/twitter-como-los-chilenos-han-reaccionado-a-la-situacion-en-chile/`

Un caso de estudio interesante es el que propone *TelarKG*. Considerando que tienen tuits relacionados a la convención constitucional de Chile, estos se pueden procesar para encontrar las entidades de *Wikidata* mencionadas, y estos enlaces agregarlos a *TelarKG* para enriquecer el grafo y de esta forma poder ejecutar consultas más especializadas que antes. Con el trabajo que proponemos, se busca poder llegar a realizar consultas, como por ejemplo, cuántos tuits hablan de un constituyente en particular o cuáles son todos los tuits publicados por convencionales de un partido en específico. Los resultados de ese trabajo podrían adaptarse para otros estudios en el ámbito de ciencias sociales en el futuro. De esta forma se podrán estudiar conversaciones en redes sociales, enriqueciendo las publicaciones con enlaces a *Wikidata*.

1.1. Objetivos

El objetivo general de este trabajo de título es implementar una estructura de datos que enlace tuits archivados y recientes de *Twitter* con *Wikidata*, permitiendo ejecutar consultas más avanzadas sobre los tuits que las que permite la plataforma de *Twitter* actualmente. Se espera que el trabajo a realizar facilite búsquedas sobre grandes volúmenes de contenido de *Twitter* para quién desee estudiar tuits sobre algún tema particular.

Para llevar a cabo el objetivo general, se busca cumplir los siguientes objetivos específicos:

1. Encontrar relaciones entre entidades mencionadas en los tuits y sus autores, con aquellas entidades existentes en *Wikidata*.
2. Generar una estructura de datos que almacene los tuits y sus entidades relacionadas. Además, debe permitir ejecutar consultas avanzadas sobre sus datos.
3. Realizar diversas consultas sobre una estructura de datos creada a partir de una muestra de todos los datos que se poseen y corroborar manualmente los resultados obtenidos.

1.2. Resumen de la Solución

Utilizando como caso de estudio la convención constitucional realizada en Chile entre los años 2021 y 2022, se comienza por obtener acceso a tuits que hagan referencia a este tema. A continuación, se utilizan diferentes implementaciones de *entity linking* en todos los tuits, encontrando así la aparición de conceptos, o "entidades", que se mencionan en ellos. Además se consideró necesario buscar manualmente y de forma exhaustiva los conceptos mencionados en algunos tuits, para poder evaluar el desempeño de los diferentes sistemas de *entity linking* y elegir cuál o cuáles utilizar. Finalmente se genera una estructura de datos que enlace los tuits con las entidades encontradas, la cual permite generar consultas especializadas sobre los datos.

1.3. Estructura de la Memoria

La estructura de la memoria cuenta con un total de 7 capítulos que exponen la información en un orden que permite un mayor entendimiento. A continuación, se describe brevemente el contenido de cada capítulo.

- Capítulo 2

  Desarrolla los fundamentos teóricos del proyecto, explicando por qué tiene bases para ser una herramienta de utilidad. Aquí se describen los recursos utilizados y el estado del arte de *Twitter*, *Wikidata*, *entity linking*, *TelarKG* y estudios sobre *Twitter*.

- Capítulo 3

  Expone una explicación de alto nivel de la solución utilizada para cumplir con los objetivos.

- Capítulo 4

  Inicio de la solución detallada, comenzando con la forma elegida de descargar los datos, aplicar *entity linking* a los tuits y el criterio de almacenamiento de los resultados.

- Capítulo 5

  Continúa con la exposición de la solución utilizada. Aquí se explica la modelación de los datos y consultas escogidas, además de la integración de los datos.

- Capítulo 6

  Expone a detalle la evaluación y discusión de la solución final, ahondando en la evaluación de *entity linking* y los experimentos realizados.

- Capítulo 7

  Se describen las conclusiones y cumplimiento de objetivos.

Capítulo 2

Estado del Arte

En este capítulo, se presenta el estado del arte de las herramientas utilizadas en la realización de este proyecto. Comienza con una explicación sobre la aplicación *Twitter*, así como su relevancia a nivel mundial y por qué tiene sentido utilizar dicha red social como fuente de información para el presente estudio. Luego, se describen las bases de datos *Wikidata* y *DBpedia*, la clase de información que almacenan, y de qué forma lo hacen. Posteriormente, se detalla qué es *entity linking*, además de características de las cuales este proyecto se sirve para la construcción de su estructura de datos y algunas herramientas existentes que implementan esta tarea. Finalizando el capítulo se explica el caso de estudio que se decidió usar para analizar los resultados del proyecto en cuestión, además de algunos estudios actuales que utilizan *Twitter* como fuente de información.

2.1. *Twitter*

Twitter es una red social de microblogueo, es decir, sus usuarios se comunican a través de publicaciones cuyo principal contenido está en formato de texto. Dichas publicaciones se denominan tuits y poseen un máximo de 280 caracteres. Con respecto al uso que le da la población mundial a esta red social: En 2023 se registraron 95,4 millones de usuarios activos de *Twitter* al mes dentro de los Estados Unidos, siendo este el país con más usuarios. En segundo lugar se encuentra Japón con 67,5 millones. El país de Latinoamérica con más usuarios fue Brasil, con 24,3 millones registrados en esta fecha[1], mientras que Chile tiene 3,4 millones de usuarios registrados.[2] Además, es relevante mencionar que *Twitter* define *mDAU* (*monetizable Daily Active Users*) como los usuarios que ingresan, o se autentican y acceden a *Twitter* durante un día dado, a través del navegador, o de sus aplicaciones que muestran anuncios. El número de *mDAU* ha ido aumentando durante los últimos 6 años, en particular en 2023 se registraron 237,8 millones a nivel mundial. Estos datos fueron obtenidos a través de *demandsage.com* la cual utiliza como fuente los informes cuatrimestrales públicos de *Twitter*.

Twitter ofrece una API que permite la obtención de una muestra, la que cuenta con

[1] Ver `https://www.demandsage.com/twitter-statistics/`

[2] Ver `https://datareportal.com/reports/digital-2023-chile`

aproximadamente el 1 % de los tuits públicos que se encuentren disponibles en tiempo real en dicha plataforma. Esta muestra puede ser filtrada, directamente desde la API, según diferentes parámetros, como la geolocalización, para obtener tuits de un país en específico, como por ejemplo Chile.

Para utilizar la API, se requiere realizar una solicitud cURL, una herramienta de línea de comando que facilita la transmisión de peticiones HTTP y la recepción de respuestas. Esta herramienta, diseñada para facilitar el uso de *Twitter* en diversos proyectos, no es de libre acceso. Para usar la API, se necesita una llave y un token, los cuales solo se pueden crear si se tiene una cuenta de tipo "desarrolladora". En el momento en que se recopilaron los tuits para este trabajo, la obtención de dicha cuenta requería la aprobación de *Twitter*, la cual implicaba presentar una solicitud con información pertinente, como la afiliación institucional y una breve explicación sobre la naturaleza del estudio que se pretende llevar a cabo con los datos provenientes de *Twitter*. Actualmente, se requiere realizar un pago para acceder a la mencionada API.

Para capturar tuits relevantes desde la API, es crucial definir un objetivo claro y determinar qué se desea monitorear en la plataforma. Dado el caso de estudio escogido, se buscan tuits relacionados con el proceso constituyente en Chile, así como comentarios sobre política chilena en general. Además, se busca recopilar todos los tuits emitidos por cada convencional, en la medida de lo posible.

La obtención de tuits de cuentas específicas o sobre temas particulares se logra mediante la adición de reglas a las solicitudes cURL. Por ejemplo, si se desea recibir tuits de la cuenta de la convencional Cristina Dorado Ortiz, cuyo nombre en la plataforma es “criordor”, la petición seguiría el formato que se muestra en el *Listing* 2.1.

```
curl -X POST \
'https://api.twitter.com/2/tweets/search/stream/rules' \
-H "Content-type: application/json" \
-H "Authorization: Bearer $BEARER_TOKEN" -d \
'{
  "add": [{"value": "from:criordor"}]
}'
```

Listing 2.1: Petición de tuits publicados por “criordor” en tiempo real

Actualmente, *Twitter* permite la búsqueda de tuits que contengan alguna palabra clave (*keyword*). De esta manera se pueden encontrar los tuits que mencionen alguna palabra en particular, como por ejemplo “Bachelet”. Esta plataforma también permite la búsqueda por usuario y permite ver sus tuits, como por ejemplo encontrar los publicados desde la cuenta “mbachelet”. Estos son los únicos dos métodos a través de los cuales *Twitter* permite la búsqueda de tuits. Es por esto que si se desean encontrar todos los tuits escritos por, o sobre presidentes y ex-presidentes de Latinoamérica, la única forma de realizarlo es generando una lista con los nombres pertinentes y ejecutar una búsqueda para cada uno. Los tuits arrojados por *Twitter* pueden contener resultados indeseados. Por ejemplo, si se buscan los tuits que contengan la palabra “Bachelet” con la intención de encontrar contenidos relacionados con la ex-presidenta de Chile, se obtendrán todos los tuits que contengan esta palabra, incluidos aquellos que hagan referencia a otra persona con este mismo apellido.

item	propiedad	valor
Q298	P37	Q1321
Chile	idioma oficial	Español

Tabla 2.1: Sentencia existente en *Wikidata*

2.2. *Wikidata*

Wikidata [24] es un repositorio de almacenamiento central al cual se le pueden realizar consultas y ser accedido por otros. Los datos almacenados son libres, es decir, está permitido utilizar estos datos de diferentes maneras: copiar, modificar, distribuir y utilizar los datos, incluso con fines comerciales, sin la necesidad de obtener un permiso. La información que provee *Wikidata* es ingresada y mantenida por editores, quienes deciden las reglas de creación de contenido y administración de este. *Wikidata* es también una "base de datos secundaria". Esto significa que, además de almacenar información, guarda sus fuentes y conexiones a otras bases de datos. *Wikidata* es una de las fuentes de información utilizadas en *Wikipedia*, *Wikibooks* y *Wikiquote*, entre otros.

El repositorio de *Wikidata* está construido en base a ítems. Cada uno tiene una etiqueta (*label*), una descripción y cero o más alias. Cada ítem posee un identificador único que comienza con la letra "Q" seguida de dígitos, por ejemplo Chile (Q298).

Una sentencia (*statement*) describe características detalladas de un ítem y consiste de una propiedad y un valor. Las propiedades en *Wikidata* poseen un identificador único que se define como la letra "P" seguida de dígitos, como con idioma oficial (P37).

Las propiedades pueden enlazar un ítem a otro dentro de *Wikidata*, o enlazarlo a alguna base de datos externa, en este caso la propiedad se denomina identificador (*identifier*). Un ejemplo de sentencia que se encuentra almacenada en *Wikidata* se muestra en la Tabla 2.1.

Actualmente *Wikidata* contiene 105.647.040 ítems y 23.091 usuarios activos (que editan y mantienen la base de datos). Entre julio del 2022 y junio del 2023 se realizaron 21 millones de ediciones. Dentro de los datos que se están almacenando en *Wikidata* se encuentran las cuentas de *Twitter* de algunas personas y organizaciones, pero no posee los tuits publicados por dichas cuentas, ni aquellos donde se mencionen a estas personas u organizaciones.

2.3. *DBpedia*

DBpedia [15] es un proyecto comunitario cuyo esfuerzo es extraer contenido estructurado de la información disponible en varios proyectos de *Wikimedia*. Esta información estructurada se asemeja a un grafo de conocimiento abierto (*open knowledge graph*). Además, se encuentra disponible para cualquier persona en la web.

Un grafo de conocimiento es una base de conocimiento distribuida como un grafo, donde

los nodos representan, generalmente, entidades y atributos y los vértices representan la conexión entre dichas entidades y atributos. A través de esta organización es posible colectar, organizar, compartir, buscar y utilizar los datos.

2.4. *Entity Linking*

PLN (procesamiento de lenguaje natural) es un campo de la ciencia de la computación que busca procesar y analizar grandes cantidades de datos escritos en lenguaje natural (lenguaje hablado por humanos). Dentro de este campo se encuentra la tarea de *entity linking*, cuya función es generar enlaces entre un texto (escrito en lenguaje natural) y alguna entidad que se encuentre dentro de un conjunto de entidades de tamaño finito y determinado. Para los fines de este trabajo se pretende utilizar *entity linking* para generar enlaces entre el texto obtenido desde *Twitter* y las entidades existentes en *Wikidata*.

Esta herramienta busca ser capaz de encontrar, por ejemplo, la entidad "Michelle Bachelet" (con el identificador Q320 en *Wikidata*) en el texto "Michelle obtuvo dos veces la presidencia de Chile", dado que el contexto de la frase permite inferir que "Michelle" hace referencia a la ex-presidenta de Chile y no a otra persona o entidad que posea la palabra "Michelle" en su nombre.

Existen implementaciones de *entity linking* que utilizan particularmente las entidades de *Wikidata* para realizar enlaces a partir de texto. Un ejemplo de esto es *OpenTapioca* [4], cuya particularidad es que fue implementado para ser entrenado y utilizado únicamente con *Wikidata*; además es ligero de entrenar, correr y mantener síncrono con *Wikidata* en tiempo real. Por otro lado existe *HEDWIG* [13], otra implementación de *entity linking* que extrae las menciones desde texto reconociendo el nombre de las entidades y luego genera enlaces con una combinación de reglas estadísticas y *PageRank* [18] (algoritmo que compara la importancia de las diferentes entidades representadas como nodos en un grafo dirigido).

Tanto reconocer entidades en tuits, como desambiguar a cuál entidad se refiere, son tareas desafiantes. Implementaciones con resultados en el estado del arte (donde se utilizan textos escritos cuidadosamente y con una mayor extensión de la que un tuit permite) presentan un desempeño fuertemente disminuido al ser utilizadas con tuits [6]. Esto ocurre porque los tuits están escritos con ortografía y gramática poco rígida, lo que dificulta la identificación de palabras y contenido. Además, poseen poco contexto porque son cortos. Son ruidosos, pues poseen muchos caracteres irrelevantes con respecto al mensaje. Dependen del contexto, dado que pueden estar respondiendo a otro tuit o expresando algo sobre una noticia en desarrollo.

Dado lo anterior, existen implementaciones de *entity linking* diseñadas específicamente para realizar esta tarea dentro de tuits. Feng et al. [7] proponen resolver el problema de la falta de contexto utilizando una hipótesis: Si bien existen muchas entidades candidatas para una mención ambigua en *Twitter* (por ejemplo "New York" puede referirse a la ciudad o al estado de EEUU), solo algunas son más probables. En este estudio se utilizaron diferentes métodos para determinar cuáles serían dichas entidades más probables. Los resultados de este trabajo mostraron un mejor desempeño comparado con otras técnicas del estado del arte, además de un mejor tiempo de ejecución, lo cual es relevante para este trabajo debido a que

se utilizará *entity linking* en contenido obtenido desde un *stream*. Otro desarrollo interesante se realizó en 2020, en el cual se utilizan tanto el texto como elementos visuales de los tuits para obtener un mejor contexto sobre este [1].

El trabajo de memoria de Miguel Zúñiga [27] consiste en un análisis de los políticos latinoamericanos en *Twitter*, obtenido del cruce de información entre una colección histórica de tuits con *Wikidata*. Logró probar que es factible enriquecer tuits con datos de *Wikidata* usando técnicas como *entity linking*. Sin embargo, su trabajo se enfoca exclusivamente en políticos latinoamericanos y considera datos desactualizados.

2.4.1. Sistemas de *Entity Linking*

Esta sección se basa en la disponibilidad de una API pública para los sistemas, lo cual facilita su implementación en este trabajo, evitando la necesidad de entrenar o instalar el sistema localmente. Este criterio es crucial dentro del contexto de este proyecto dada la magnitud de las fuentes de datos, tales como *Wikidata*, *DBpedia*, entre otras. Además, se establecieron otros dos criterios que los sistemas de *entity linking* deben cumplir para ser considerados en este proyecto: deben emplear una fuente de conocimiento con enlaces que apunten a *Wikidata* y, por último, deben estar diseñados para etiquetar texto en español u otro idioma similar, como por ejemplo italiano (en el caso de *TagMe*).

Los últimos años han visto muchos desarrollos en los sistemas de *entity linking* impulsados por avances en el área de *Deep Learning*. Para una discusión más profunda de estos avances, se hace referencia a la encuesta reciente hecha por Sevgili et al. [22]. Entre los sistemas de *entity linking* de vanguardia, se destacan tres debido a la calidad de sus resultados: *REL* [23], *CHOLAN* [11] y aquel propuesto por Kolitsas et al. [14]. A pesar de su alto rendimiento, ninguno de estos sistemas es adecuado para su uso en este proyecto. Los últimos dos carecen de una API. Por otro lado, *REL* es la mejor opción entre los tres, pero no cumple uno de los criterios propuestos, ya que no esta disponible en español u otro idioma similar.

Se estudiaron tres sistemas de *entity linking* diferentes que cumplen los criterios mencionados anteriormente.

- *TagMe* [8] es una herramienta de identificación de entidades que utiliza *Wikipedia* como su fuente de datos. *TagMe* además es capaz de identificar frases cortas significativas en la marcha, en un texto sin estructura y enlazarlas a la página de *Wikipedia* pertinente. Cada entidad de *Wikipedia* ofrece un enlace a una entidad de *Wikidata*, pero no cada entidad de *Wikidata* tiene una entidad de *Wikipedia* asociada.
- *DBPedia Spotlight* [17] es una herramienta para anotar automáticamente las menciones de entidades en *DBpedia* que se encuentren en un texto. Cada entidad de *DBpedia* está asociada con una entidad de *Wikipedia* y así con *Wikidata*. Pero de nuevo, no todas las entidades de *Wikidata* tienen una entidad en *DBpedia* o *Wikipedia*.
- *OpenTapioca* [5] es un sistema rápido de *entity linking* para *Wikidata*. Está síncrono con *Wikidata* en tiempo real lo que propone una característica útil para casos de estudio que se beneficien de tener la mayor cantidad de entidades posibles.

2.4.1.1. Librerías Asociadas

En el contexto de este proyecto, se exploran dos librerías de *Python* para el procesamiento de lenguaje natural y la identificación de entidades: *Spacy* y *tagme-python*.

- *Spacy*[3] es una librería implementada para *Python*, reconocida por sus creadores (la compañía de software *Explosion*) como una herramienta de procesamiento de lenguaje natural con potencia industrial. Sus comandos permiten encontrar etiquetas utilizando directamente *OpenTapioca* y *DBPedia Spotlight*, sistemas de *entity linking* previamente mencionados.

  Esta librería presenta tres características clave que la convierten en una elección adecuada para el proyecto. En primer lugar, se destaca por su facilidad de uso con una instalación rápida y una API simple. Además, demuestra eficacia en tareas de extracción de entidades con información de gran escala y es compatible con más de 73 lenguajes, incluido el español. Dado que el proyecto implica procesar alrededor de 20 millones de tuits en español, no es necesario gestionar manualmente la comunicación con las APIs de *OpenTapioca* o *DBPedia Spotlight*, lo que requeriría más recursos sin aportar valor adicional.

  Spacy permite etiquetar un texto de hasta 1.000.000 caracteres con entidades de *Wikidata* o *DBPedia*. Con unas pocas líneas de código, la librería establece la comunicación con los servidores de *OpenTapioca* o *DBPedia Spotlight*, devolviendo todas las entidades vinculadas. Por ejemplo, al etiquetar con *Opentapioca* el texto "Chile declara a América Latina y el Caribe como zona prioritaria en sus relaciones internacionales", se obtienen las entidades "Chile", "América", "Latina" y "Como", junto con datos adicionales como los ids en *Wikidata*, descripciones y tipos de entidad.

- *tagme-python*[4] es el *wrapper* oficial de la API de *TagMe*, diseñado específicamente para *Python*. Esta librería liviana tiene como único propósito facilitar y agilizar la interacción con la API de *TagMe*. La elección de utilizar esta librería se basa en la intención de mejorar la eficiencia del proceso, ya que manejar la comunicación manual con la API no aportaría un valor significativo al proyecto. Además, proporciona flexibilidad al configurar el idioma en el que se desea realizar el etiquetado, con un límite máximo de 2.000.000 de caracteres por texto a etiquetar. Al procesar con esta librería el ejemplo anterior ("Chile declara a América Latina y el Caribe como zona prioritaria en sus relaciones internacionales"), se obtienen las entidades "Chile", "Declara", "American English", "Hispanic and Latino Americans" y "Lake Como", además de la posición en el texto en el que se encontraron estas entidades y la cadena de caracteres que las representa.

[3] `https://spacy.io/`

[4] `https://github.com/marcocor/tagme-python`

2.5. Caso de Estudio

El caso de estudio seleccionado para el proyecto descrito en este documento es la convención constitucional, elegida en Chile en el año 2021. Existe una colección de datos llamada *TelarKG*; esta es un grafo de conocimiento (*knowledge graph*) que contiene datos sobre la convención constitucional de Chile. Entre los datos incluidos, se destacan alrededor de 20 millones de tuits relevantes al mencionado proceso. *TelarKG* es un proyecto lanzado por el Instituto Milenio Fundamentos de los Datos, el cual es un centro científico multidisciplinario que investiga problemas relacionados con datos, desde el origen de estos, hasta aplicaciones que se puedan obtener desde ellos y el impacto social que producen.

La implementación del proyecto descrito en este documento se logra utilizando los datos de *TelarKG*, así se puede evaluar el resultado de esta generando consultas sobre la convención constitucional y partidos políticos.

2.6. Estudios sobre *Twitter*

El contenido de *Twitter* se ha utilizado en diversos estudios durante los últimos años. Un ejemplo de esto es *"Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach"* [26], un trabajo realizado durante el año 2020 con el objetivo de estudiar y examinar discusiones, preocupaciones y sentimientos relacionados con el virus COVID-19. Para esto se utilizaron tuits publicados por usuarios de *Twitter*. Cabe destacar que para encontrar tuits relevantes para el estudio, éstos se filtraron únicamente a través de una lista de 20 hashtags (por ejemplo, "#coronavirus", "#COVID-19", "#quarantine").

Otro ejemplo reciente del uso de tuits como datos de estudio es el trabajo realizado por Sabuncu et al. [21], también durante el año 2020. El objetivo de este estudio fue crear y poner a prueba un modelo de predicción electoral, que tome en cuenta diferentes factores y técnicas utilizadas en el pasado por otros autores. Este método se puso a prueba con la elección presidencial del año 2020 en Estados Unidos. Los datos utilizados se extrajeron desde *Twitter*, particularmente se utilizan tuits que posean al menos una de las 29 palabras claves que sus autores consideraron pertinentes.

Dado que en ambos estudios solo utilizan búsquedas por palabras explícitas en *Twitter*, creemos que el desarrollo de este trabajo propuesto podría entregar una herramienta diferente y complementaria para la búsqueda de contenido relevante para su posterior análisis.

Capítulo 3

Resumen de la Solución

En este capítulo, se presenta una visión general de la solución propuesta para enriquecer los tuits de *TelarKG* mediante el enlazamiento de menciones de entidades con *Wikidata*. El proceso de desarrollo se inicia con la recopilación de datos relevantes para el caso de estudio, se pueden encontrar más detalles de eso en la Sección 3.1.

La siguiente etapa explicada en la Sección 3.2 implica el uso de técnicas de *entity linking* para enriquecer la estructura de datos. Se procesan los tuits con tres herramientas distintas: *TagMe*, *DBPedia Spotlight*, y *Opentapioca*.

La Sección 3.3 aborda el almacenamiento de información, donde se establece la conexión entre las entidades y sus menciones en los tuits. Esto permite la recopilación de tuits que cumplen con criterios de búsqueda específicos asociados a las entidades encontradas. Se proporcionarán más detalles sobre los componentes de este capítulo en los siguientes.

3.1. Recopilación de Datos

Para comenzar a desarrollar el proyecto descrito en este documento se obtienen algunos tuits relevantes al caso de estudio. El proceso de recopilación de tuits fue hecho por el Instituto Milenio Fundamento de los Datos para el proyecto *TelarKG*. Se logra este objetivo al aplicar un filtro a todos los tuits proporcionados por la API de *Twitter*, un proceso detallado en el Capítulo 2.1. Este filtro se implementa utilizando palabras clave pertinentes al caso de estudio, como “convención”, “constituyente” y “constitución”. También se obtuvieron los tuits hechos por cuentas oficiales de convencionales electos para escribir la constitución, dada su relevancia en el caso de estudio. Además de recopilar tuits, se llevaron a cabo búsquedas de entidades relevantes al proceso constitucional de Chile en *Wikidata*, incluyendo convencionales constituyentes (identificados por el IMFD) y partidos políticos. A continuación, se presentan en la Tabla 3.1 algunos ejemplos de convencionales y sus correspondientes entidades en *Wikidata*.

Convencional	**Identificador de Wikidata**
Damaris Abarca González	Q28480029
Alondra Carrillo Vidal	Q108679269
Eric Chinga Ferreira	Q108220596

Tabla 3.1: Entidades en *Wikidata* para algunos convencionales

3.2. Utilización de *Entity Linking*

Para enriquecer la estructura de datos que se quiere lograr, es necesario encontrar entidades que se mencionen en los tuits obtenidos con anterioridad. Para encontrar las diferentes entidades que se mencionan en los tuits relevantes, estos son procesados con tres herramientas de *entity linking* diferentes, las cuales son *OpenTapioca*, *TagMe* y *DBPedia Spotlight*.

Las tres herramientas de *entity linking* encuentran entidades existentes en tres bases de datos diferentes. *OpenTapioca* usa *Wikidata* como fuente de entidades, *TagMe* utiliza *Wikipedia* con el mismo fin, y *DBPedia Spotlight* encuentra entidades de *DBpedia*.

Cabe mencionar, que existen enlaces entre *Wikipedia*, *Wikidata* y *DBpedia*. Por esto los resultados de los diferentes sistemas de *entity linking* pueden ser comparados, enlazados, o incluso complementarse entre sí. Lo único que se debe tener en consideración es que estos tres no poseen la misma cantidad de entidades, ya que *Wikidata* tiene muchas más, lo que imposibilita enlazar todas las entidades de *Wikidata* con las pertenecientes a *Wikipedia* o *DBpedia*. Sin embargo, en sentido contrario, cada entidad de *DBpedia* y *Wikipedia* tiene una correspondencia con una entidad en *Wikidata*.

3.3. Almacenamiento de Información

La conexión entre las entidades y sus menciones en los tuits cumple la función de permitir recopilar aquellos que cumplan con algún criterio de búsqueda específico, el cual se asocia a las entidades encontradas en ellos. Por ejemplo, se tienen tres tuits donde cada uno de ellos menciona a algún convencional. Al almacenarse las entidades de estos convencionales con un enlace a estos tuits, éstos últimos se pueden obtener desde la base de datos si se piden todos los tuits que hagan menciones a convencionales constituyentes.

A modo de ejemplo se utiliza el tuit que se muestra en la Figura 3.1. Este fue publicado por la constitucional Alejandra Flores Carlos, la cual posee una entidad en *Wikidata* (Figura 3.2). Por lo tanto podemos acceder a cierta información sobre ella que se encuentra almacenada en *Wikidata*, como por ejemplo, género, ciudad y fecha de nacimiento, entre otros. Un ejemplo de parte de la información almacenada en *Wikidata* de la convencional Alejandra Flores Carlos se encuentra en la Figura 3.3.

En la Figura 3.1 se muestra destacada la sección del texto que dice “Universidad Arturo Prat”, caracteres que la herramienta de *entity linking* detectó como la mención de una entidad de *Wikidata*. Dicha entidad corresponde a la Universidad Arturo Prat (Figura 3.4).

Figura 3.1: Ejemplo de tuit publicado por la convencional Alejandra Flores Carlos

Para el ejemplo de la Figura 3.1 se almacenaría el tuit, con enlaces a la entidad encontrada en el texto y a la entidad de su autora, la convencional Alejandra Flores Carlos. Luego la forma de acceder a este tuit sería a través de consultas. Estas consultas se hacen respecto a criterios que deban cumplir las entidades enlazadas. Por ejemplo, se pueden pedir todos los tuits que mencionen alguna ciudad; de esta manera se buscan todas las entidades que cumplan con el criterio de ser una ciudad, y luego se retornan todos los tuits enlazados a ellas. El tuit del ejemplo podría ser retornado al hacer una consulta como "Menciones a instituciones educacionales hechas por convencionales".

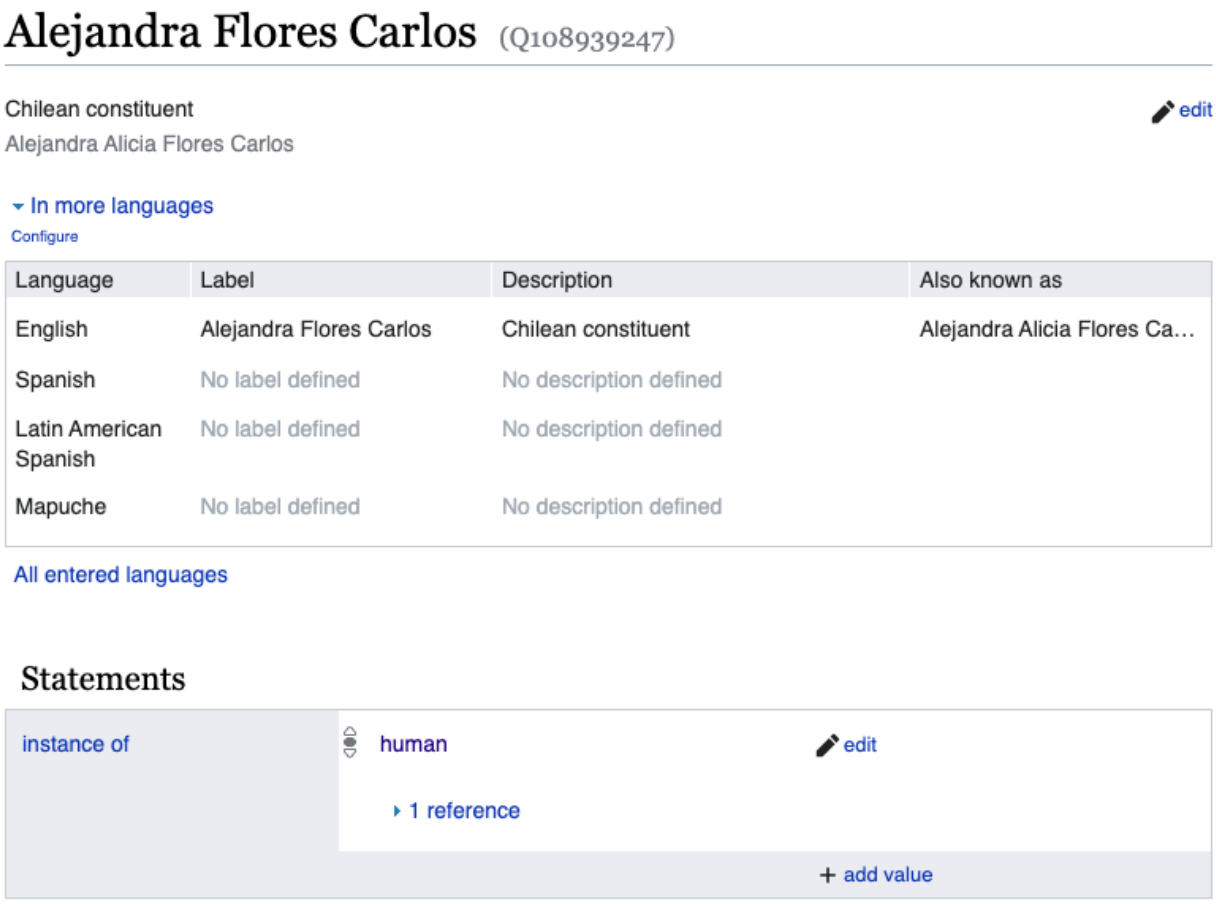

Figura 3.2: Entidad de *Wikidata* de la convencional Alejandra Flores Carlos

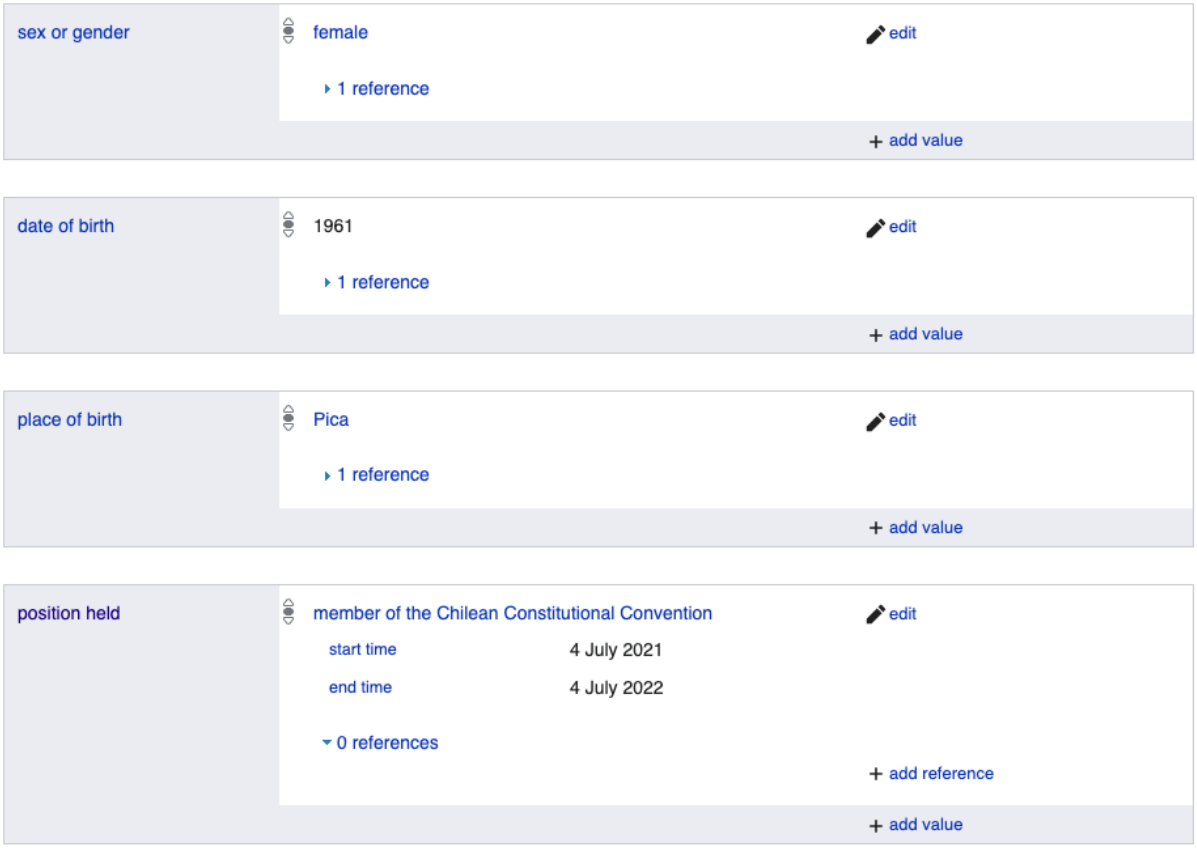

Figura 3.3: Información en *Wikidata* de la convencional Alejandra Flores Carlos

Figura 3.4: Entidad de *Wikidata* de la Universidad Arturo Prat

Capítulo 4

Procesamiento de Datos: De la Recopilación a la Implementación de *Entity Linking*

En este capítulo se proporciona una visión detallada del proceso de recopilación y almacenamiento de datos facilitados por el Instituto Milenio Fundamento de los Datos. El tamaño total de los datos es de aproximadamente 20 millones de tuits, ocupando 18,82 GB. Además, se aborda la implementación de herramientas de *entity linking* como *TagMe*, *DBPedia Spotlight* y *OpenTapioca* para analizar y etiquetar entidades en los tuits. Se describe el procesamiento de datos, la decisión de agrupar tuits en lotes para optimizar el tiempo y la estrategia de descarte de tuits sin información relevante. La cantidad de tuits descartados representa alrededor del 15,4 % de los datos totales.

4.1. Recopilación y Almacenamiento de Datos

El Instituto Milenio Fundamento de los Datos almacena la información esencial para este proyecto en *Google Cloud*, específicamente los tuits relevantes al caso de estudio. La obtención de estos datos se realizó mediante la API de *Twitter*, la cual recibe como entrada una o más palabras clave y proporciona todos los tuits que contienen una o más de esas palabras clave. Para este propósito, se eligieron términos pertinentes al caso de estudio como palabras clave, entre ellos: "convención", "constituyente" y "constitución".

El instituto ha implementado un proceso automatizado mediante el cual una máquina virtual se activa cada noche. Esta máquina virtual solicita a la API los tuits disponibles publicados en las últimas 24 horas que cumplen con las condiciones establecidas, ya sea al contener una palabra clave o ser publicados por un convencional.

Cabe destacar que *Google Cloud* es una plataforma de servicios en la nube proporcionada por *Google*, ofreciendo diversas herramientas y recursos para el almacenamiento, gestión y análisis de datos. Para iniciar el proceso de obtención y extracción de datos, es fundamental

comprender las especificidades de la descarga en el entorno de *Google Cloud*.[1]

En primer lugar, se destaca la limitación de exportar datos directamente de una tabla a un archivo local, Hojas de cálculo de *Google* o *Google Drive*. La única ubicación de exportación permitida es *Cloud Storage*. Adicionalmente, existe una restricción en la cantidad de datos exportables en un solo archivo, teniendo 1GB como máximo tamaño permitido. Si se necesita exportar datos que superan este límite, estos se pueden distribuir en varios archivos anidados.

Google Cloud permite la exportación de archivos anidados en los formatos *Avro*, *JSON* y *Parquet*. Este aspecto cobra relevancia al considerar que los tuits utilizados en este proyecto están divididos en dos tablas de datos ubicadas en *Google Cloud*, donde una de ellas excede el límite de 1GB. Todos los tuits proporcionados por el Instituto Milenio Fundamento de los Datos fueron recopilados utilizando la API de *Twitter* (Sección 2.1) y se encuentran almacenados en dos tablas: *tw-streaming* y *tw-convencionales*.

tw-streaming, con formato *Avro*, contiene 19.545.968 filas y ocupa 18,22 GB. Incluye diversas columnas como *timestamp*, identificadores, información de usuario, detalles sobre el contenido del tuit, entre otros. Las columnas comunes entre ambas tablas abarcan aspectos como la marca de tiempo de creación (created_at), identificadores (id), información sobre el usuario (user), idioma (lang), contador de retuits (retweet_count), y muchos otros.

La tabla *tw-convencionales*, en formato *JSON*, comprende 616.486 filas con un tamaño de 1,1 GB. Recopila tuits realizados por los primeros convencionales electos, desde julio de 2021 hasta septiembre de 2022 (periodo en el ejercían el rol de convencionales constituyentes). Al igual que la tabla *tw-streaming*, incluye columnas compartidas que abarcan información sobre el tuit, usuario y metadatos relevantes. Entre estas columnas comunes se encuentran *timestamp*, identificadores, información del usuario, idioma, contador de retuits, y más.

Con respecto a la información relevante que se puede obtener directamente de los datos: El promedio de tuits publicados por convencional es de 4.532,99, con una longitud promedio de 124,79 caracteres. La convencional con más tuits publicados es Cristina Dorador Ortiz, quien, con el nombre de usuario "criordor" acumula un total de 35.299 publicaciones. En cambio, de Paola Grandón González, con el nombre de usuario "D17Grandon", solo se dispone de un tuit.

Para la ejecución de este proyecto, se cuenta con un conjunto total de aproximadamente 20 millones de tuits, los cuales ocupan un espacio aproximado de 18,82 GB. La considerable magnitud de los datos constituye un desafío significativo para su gestión, procesamiento y análisis.

4.2. Implementación de *Entity Linking*

La existencia de una API fue el factor determinante en la elección de *TagMe*, *DBPedia Spotlight* y *OpenTapioca* como herramientas de *entity linking*. Esto se debe a la complejidad de operar con ellas localmente, ya que implica un esfuerzo significativo en términos de tiempo y

[1] Ver `https://cloud.google.com/bigquery/docs/exporting-data?hl=es-419`

recursos. En particular, al intentar correr *OpenTapioca* localmente, por ejemplo, se enfrentaría al desafío de parsear el *dump* de *Wikidata*, el cual tiene un tamaño aproximado de 1,4TB. Este proceso requiere de recursos considerables y puede ser impracticable para entornos con limitaciones de recursos.

Este trabajo requiere un proceso de entrenamiento para identificar etiquetas adecuadas, lo que implica horas considerables de procesamiento de información para obtener resultados relativamente confiables. Sin embargo, es importante destacar que este enfoque va más allá del alcance de este proyecto, ya que el proceso de etiquetado es solo una parte de la propuesta global.

Cada una de las herramientas de *entity linking* seleccionadas produce resultados ligeramente distintos, debido a que cada una utiliza diferentes fuentes de entidades (*TagMe* usa *Wikipedia*, *OpenTapioca* usa *Wikidata* y *DBPedia Spotlight* usa *DBPedia*). A pesar de estas diferencias, es posible unificar los resultados al encontrar la entidad en *Wikidata* a la que hacen referencia. Esto dado que *Wikidata* contiene todas las entidades de *Wikipedia* y *DBPedia*.

Aunque cada API posee sus particularidades, comparten algunas características comunes. Por ejemplo, permiten la selección del idioma para la búsqueda de etiquetas y establecen un límite en el tamaño del texto a etiquetar. Con el objetivo de analizar de manera más detallada las similitudes y diferencias entre los distintos sistemas, se utilizó cada uno de ellos para etiquetar el tuit que se muestra en la Figura 4.1. Los resultados de estas aplicaciones se presentan en los *Listings* 4.2, 4.4 y 4.6. A continuación, se detallan las especificaciones de cada sistema.

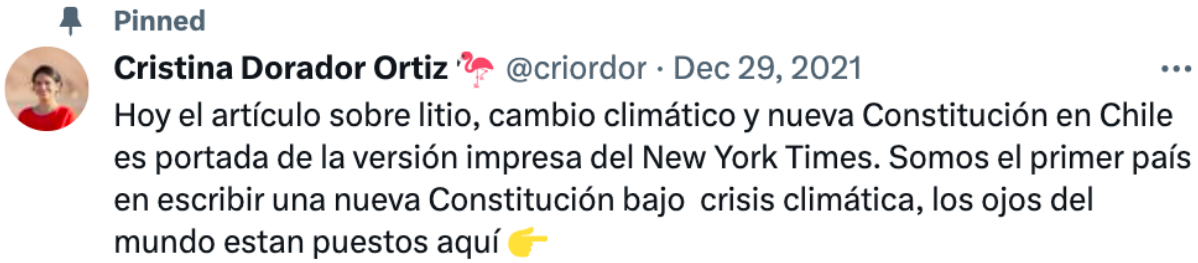

Figura 4.1: Ejemplo de tuit para ilustrar el funcionamiento de sistemas de *entity linking*.

4.2.1. *TagMe*

Se utiliza la librería *tagme-python* para solicitar a la API de *TagMe* que etiquete tuits, reciba y gestione los datos de la respuesta. Para establecer esta conexión, es necesario contar con un token, el cual se obtiene fácilmente creando una cuenta en *D4science*, la infraestructura que aloja a *TagMe*. El tamaño máximo permitido para el texto a etiquetar es de 2 MB, equivalente a aproximadamente 2.000.000 de caracteres. *TagMe* tiene la capacidad de etiquetar textos en inglés, italiano y alemán.

Cuando se le solicita a este sistema etiquetar el tuit presentado en la Figura 4.1, este

retorna lo reflejado en el *Listing* 4.2. En este, se detallan todos los campos vinculados a una entidad identificada por este sistema de *entity linking*, los cuales se explican a continuación:

- mention: Secuencia de caracteres en el texto que hace referencia a la entidad.
- start: Posición de inicio en el texto enviado donde se identificó la referencia a la entidad.
- score: Medida de confiabilidad de esta subcadena como una mención significativa.
- end: Posición final en el texto enviado donde se encontró la referencia a la entidad.
- id: Identificación de la entidad en *Wikipedia*.
- entity_title: Nombre de la página en *Wikipedia* correspondiente a la entidad identificada.

La solicitud de etiquetado del tuit mostrado en la Figura 4.1 mediante la librería *tagme-python* está detallada en el *Listing* 4.1.

```
import tagme

tagme.GCUBE_TOKEN = "<Token generado>"
tweet = "..."

#Configuracion de idioma
lunch_annotations = tagme.annotate(tweet, lang='en')

#Entidades con una score significativo
for ann in lunch_annotations.get_annotations(0.1):
    info = [ann.begin,  ann.end, ann.entity_id, ann.
   entity_title, ann.score]
    print (* info , sep = '\t')
```

Listing 4.1: Campos disponibles de cada entidad encontrada por *TagMe*

```
0       3       14052346        Peter Hoy
   0.14891961216926575
54      66      5274546 Constitución, Chile
   0.21982082724571228
67      69      8569916 English language
   0.10053771734237671
70      75      5489    Chile   0.4864366352558136
76      78      26825   Spanish language
   0.11912304908037186
101     108     3496605 Impresa 0.11152204871177673
113     127     30680   The New York Times
   0.38288381695747375
172     184     5274546 Constitución, Chile
   0.1454741358757019
213     217     44589655        Ojós    0.4615384638309479
```

Listing 4.2: Resultados para consulta a *TagMe*

4.2.2. *DBPedia Spotlight*

Como se mencionó previamente, existe una API para realizar solicitudes a *DBPedia Spotlight*. Además la librería *Spacy* posee una configuración para ejecutar consultas directamente a la API de *DBPedia Spotlight*. En consecuencia, se llevan a cabo las solicitudes utilizando esta librería de *Python*. Hay que especificar el idioma del texto que se desea etiquetar, teniendo como opciones alemán, inglés, español e italiano. La API acepta solicitudes con textos de hasta 1.000.000 caracteres de longitud.

El código utilizando *Spacy* para envíar a la API de *DBPedia Spotlight* la solicitud para etiquetar el tuit representado en la Figura 4.1 se encuentra en el *Listing* 4.3. La respuesta se encuentra presentada en el *Listing* 4.4. A continuación, se proporciona un desglose de cada campo que abarca una entidad identificada por el sistema:

La solicitud en *Python* para etiquetar el tuit de la Figura 4.1, utilizando *Spacy* con la configuración de *DBPedia Spotlight*, se encuentra en el *Listing* 4.3. La respuesta a esta solicitud se detalla en el *Listing* 4.4. A continuación, se desglosan los campos de una entidad identificada por el sistema:

- @URI: Identificador de *DBPedia* asociado a la entidad identificada.
- @support: Valor numérico que indica el nivel de confianza o certeza con el que se ha identificado la entidad.
- @types: Categorías o clases a las que pertenece la entidad según la ontología de *DBpedia*.
- @surfaceForm: Caracteres del texto que hacen referencia a la entidad.
- @offset: Posición en el texto donde comienza la mención a la entidad.
- @similarityScore: Puntuación de similitud o correspondencia semántica entre la mención de texto y la entidad identificada.
- @percentageOfSecondRank: Porcentaje de confianza asignado a la segunda entidad candidata más probable.

```
import spacy

tweet = "..."
#Configuracion de idioma
nlp = spacy.blank('es')
nlp.add_pipe('dbpedia_spotlight', config={'language_code': 'es'
    })

doc = nlp(tweet)
for ent in doc.ents:
    info = [ent._.dbpedia_raw_result['@URI'], ent._.
    dbpedia_raw_result['@support'], ent._.dbpedia_raw_result['
    @types'], ent._.dbpedia_raw_result['@surfaceForm'], ent._.
```

```
    dbpedia_raw_result['@offset'], ent._.dbpedia_raw_result['
    @similarityScore'], ent._.dbpedia_raw_result['
    @percentageOfSecondRank']]
11      print(*info, sep = '\t')
12      print()
```

Listing 4.3: Campos disponibles de cada entidad encontrada por *DBPedia Spotlight*

```
1 http://es.dbpedia.org/resource/Litio    1120            litio
      22      0.9999087767849733      5.979892220929704E-5
2
3 http://es.dbpedia.org/resource/Chile    119751  Wikidata:Q6256,
    Schema:Place,Schema:Country,DBpedia:PopulatedPlace,DBpedia:
    Place,DBpedia:Location,DBpedia:Country    Chile    70
    0.9999957226835892      2.999823076286813E-6
4
5 http://es.dbpedia.org/resource/Nueva_York       80981
    Wikidata:Q3455524,Schema:Place,Schema:AdministrativeArea,
    DBpedia:Region,DBpedia:PopulatedPlace,DBpedia:Place,DBpedia:
    Location,DBpedia:AdministrativeRegion New York        113
      0.9998201779831786      1.5228925998135275E-4
```

Listing 4.4: Resultados para consulta a *DBPedia Spotlight*

4.2.3. *OpenTapioca*

OpenTapioca no solo cuenta con una API, si no que también existe una librería de *Python*, diseñada para hacer peticiones HTTP a *OpenTapioca*. Esta característica facilita la integración de *OpenTapioca* en flujos de trabajo basados en *Python*. *OpenTapioca* tiene la capacidad de procesar textos en varios idiomas, entre ellos, inglés y español, que son relevantes para el alcance de este trabajo, ya que los tuits están en su gran mayoría escritos en español.

De acuerdo con la documentación, *OpenTapioca* admite solicitudes con textos de hasta 1.000.000 caracteres, lo que equivale aproximadamente a 3.500 tuits. Esto es particularmente relevante en este trabajo, dado que se busca procesar alrededor de 20 millones de tuits. Pese a la información entregada por la documentación, en la práctica *OpenTapioca* admite textos de aproximadamente 112.000 caracteres, que equivalen a 400 tuits.

En el *Listing* 4.5, se proporciona la información disponible para cada entidad identificada mediante *Spacy*. El formato de la información obtenida para cada entidad identificada por *OpenTapioca* se muestra en la Figura 4.6. A continuación, se describen en detalle cada uno de los campos disponibles por entidad:

- text: El texto que está siendo etiquetado.
- kb_id: Identificador de *Wikidata* asociado a la entidad encontrada.
- label: Hace referencia al tipo general de la entidad encontrada.

- description: Breve descripción de la entidad.
- score: Medida cuantitativa de la confianza en la precisión de la entidad extraída.
- types: Especifica el tipo de entidad encontrada.
- aliases: Nombres alternativos asociados a la entidad identificada.
- rank: Indica la posición o importancia relativa de una entidad en comparación con otras entidades extraídas. Un rango más bajo equivale a una entidad más relevante.

```python
import spacy

tweet = '...'
#Configuracion de idioma
nlp = spacy.blank("es")
nlp.add_pipe('opentapioca')

doc = nlp(tweet)

for span in doc.ents:
    info = [span.text, span.kb_id_, span.label_, span._.description, span._.score, span._.types, span._.aliases, span._.rank]
    print(*info, sep = '\t')
    print()
```

Listing 4.5: Campos disponibles de cada entidad encontrada por *OpenTapioca*

```
Chile   Q298    LOC     ['country in South America']    0.6655630640180009      {'Q43229': True, 'Q618123': True, 'Q5': False, 'P2427': False, 'P1566': True, 'P496': False}     ['Chile', 'Republic of Chile', '****', 'CHI', 'cl', 'Republica de Chile', 'República de Chile']    14.407348194896992

New York Times  Q9684   ORG     ['American daily newspaper']    1.0095550863536125      {'Q43229': True, 'Q618123': False, 'Q5': False, 'P2427': False, 'P1566': False, 'P496': False}      ['The New York Times', 'The NYT', 'N Y Times', 'N. Y. Times', 'The Times', 'New York Times (newspaper)', '@nytimes', 'The NY Times', 'The Gray Lady', 'New-York Daily Times', 'New York Times', 'The New York Daily Times', 'The Sunday Times', 'NYT', 'The Paper of Record', 'NYTimes', 'nytimes.com', 'Times']      8.745515277724659

puestos Q2946   LOC     ['palace in Versailles, France and location of the Museum of the History of France']    0.08407655720953688     {'Q43229': True, 'Q618123': True, 'Q5
```

```
': False, 'P2427': False, 'P1566': True, 'P496': False}
 ['Palace of Versailles', 'Puestos', 'Versailles. Palais', '
Versailles. Chateau', 'Château', 'Versailles', 'Château de
Versailles', 'Versailles. Château', 'Chateau', 'Chateau de
Versailles', 'Palais du Versailles', 'Versailles Palace', '
France. Chateau Versailles']      10.214684448908935
```

Listing 4.6: Resultados para consulta a *OpenTapioca*

4.3. Procesamiento de Datos con *Entity Linking*

Para llevar a cabo el procesamiento de todos los datos suministrados por el Instituto Milenio Fundamento de los Datos en el contexto del proyecto *TelarKG*, se optó por utilizar los sistemas de *entity linking DBPedia Spotlight* y *OpenTapioca*. Esta elección se basó en evaluaciones previas destinadas a comprender la calidad de las etiquetas generadas por los tres sistemas mencionados en la sección anterior. Para obtener detalles adicionales sobre el proceso de evaluación y la selección de estas herramientas, se puede consultar el Capítulo 6.

Se decidió agrupar tuits en lotes para su etiquetado, aprovechando la capacidad máxima que permitiera cada sistema. Esta elección se respalda por dos razones fundamentales. En primer lugar, dada la vasta cantidad de datos, etiquetar los 20 millones de tuits individualmente consumiría un tiempo excesivo, pues generaría una cantidad enorme de peticiones HTTP. (se pueden encontrar más detalles de esto en la Sección 6.1). Además, considerando que la mayoría de los tuits se centran en la política chilena, agruparlos en lotes no debería afectar significativamente el tipo de entidad detectada por los sistemas en los distintos tuits.

Además de organizar los tuits en lotes para su procesamiento, se incorporan pausas entre cada solicitud. Esta práctica tiene como objetivo prevenir la interrupción de la comunicación con las diversas APIs, debido a una posible sobrecarga de peticiones. Se optó por establecer un tiempo de espera de 2 segundos para simular un patrón de tráfico más similar al humano.

Como se detalla en la Sección 4.2, la herramienta de *entity linking DBPedia Spotlight* devuelve entidades de *DBPedia*. Con el objetivo de uniformizar los resultados de ambos sistemas de *entity linking*, se procede a buscar los identificadores de *Wikidata* asociados a las mismas entidades identificadas por *DBPedia Spotlight*. Este proceso se realiza mediante una consulta a *Wikidata*, un ejemplo de código se presenta en el *Listing* 4.7. Las entidades de *DBpedia* son proporcionadas por *DBPedia Spotlight* en forma de URL que redirige a la página correspondiente en *DBPedia* en español.

Para llevar a cabo el emparejamiento de estas entidades con las de *Wikidata*, se utiliza como criterio la coincidencia de nombres entre las páginas de *Wikipedia* y *DBpedia*. La consulta solicita a *Wikidata* que proporcione los identificadores de las entidades asociadas a las diversas páginas de *Wikipedia* (equivalentes a las páginas de *DBPedia* en estricto sentido). Estas consultas a *Wikidata* se realizan utilizando el lenguaje *SPARQL*.

Se identificaron 73.715 entidades mencionadas en los tuits de *TelarKG* mediante *DBPedia Spotlight*. De este conjunto, 70.550 entidades, lo que equivale al 95,7 %, fueron exitosamente

mapeadas como entidades de *Wikidata*. Aquellas menciones a entidades que no pudieron ser localizadas en *Wikidata* fueron excluidas del análisis. Aunque no fue posible determinar las razones detrás de la imposibilidad de encontrar estas entidades en su totalidad, debido al significativo costo temporal que implica, algunas de estas entidades estaban identificadas con nombres alternativos en *Wikidata*. Obtener las entidades restantes requeriría una búsqueda individual de cada una, y dado que representan solo el 4,3 % del total, se tomó la decisión de no incluirlas en el análisis principal.

Es fundamental resaltar que los nombres obtenidos deben ser codificados en formato de codificación porcentual para evitar la pérdida de resultados debido a caracteres especiales. Esta tarea se realiza mediante el uso de la función `urllib.parse.quote(entity_name, safe='~()_-*!.')` en *Python*, asegurando una representación segura de los caracteres en la URL.

```
SELECT ?label ?wiki_id WHERE {
   ?s schema:about ?wiki_id .
   BIND(IRI(CONCAT("https://es.wikipedia.org/wiki/",?label)) AS
    ?s)
   VALUES ?label {
      "Jose_Soto_(actor_espa%C3%B1ol)"
   }
}
```

Listing 4.7: Consulta en SPARQL para obtener identificadores de *Wikidata* de la página de *Wikipedia* `https://es.wikipedia.org/wiki/Jose_Soto_(actor_espa%C3%B1ol)`

4.4. Limpieza de Tuits

Con aproximadamente 20 millones de tuits disponibles para este proyecto, se implementaron estrategias para evitar el almacenamiento de información innecesaria. Como resultado, algunos tuits fueron descartados debido a la falta de información relevante. El proceso detallado de selección y descarte se ilustra en la Figura 4.2.

El proceso de descarte se inicia al recibir un tuit; luego, se procesan sus datos. Se verifica si fue publicado por un miembro de la convención constitucional y se buscan etiquetas mediante el sistema de *entity linking*. Si carece de etiquetas y no fue publicado por un convencional, se descarta el tuit. Por el contrario, si fue publicado por un convencional o tiene etiquetas con el sistema de *entity linking*, se almacena el tuit enlazado a las entidades, indicando su ubicación en el texto y su tipo. Además, se guarda un enlace al creador del tuit si es relevante. Este proceso se repite para todos los tuits, dando lugar a la creación de una estructura que los contiene, junto con toda la información recopilada sobre el autor y las entidades de *Wikidata* identificadas.

Los tuits escritos por individuos no registrados como convencionales constituyentes y que no poseen etiquetas son descartados, ya que no pueden ser encontrados mediante el algoritmo de búsqueda implementado. Esto se debe a que la búsqueda se realiza a través de las entidades de *Wikidata*, y las consultas describen criterios que deben cumplir dichas

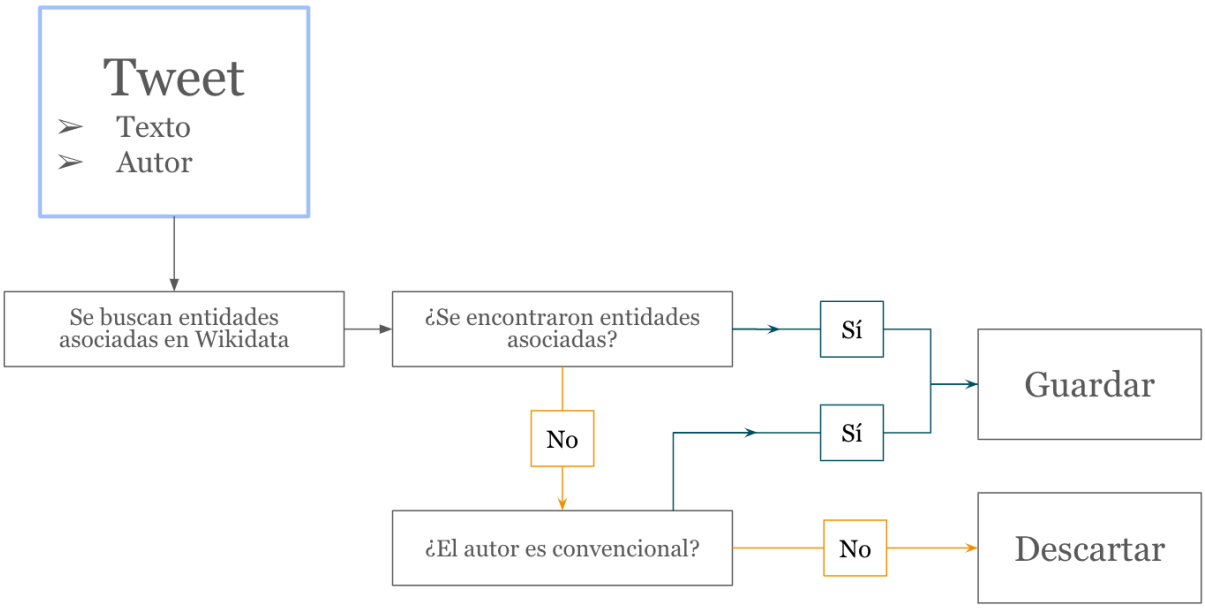

Figura 4.2: Flujo de selección de tuits

entidades. El sistema retorna los tuits que contienen al menos una entidad que cumple con esos criterios. Un tuit sin entidad es inaccesible en este modelo. La cantidad de tuits descartados representa aproximadamente el 15,4 % de todos los datos, equivalente a 3.151.496 tuits.

Capítulo 5

Caso de Uso: *TelarKG*

El caso de uso seleccionado para evaluar el desarrollo del proyecto de memoria es *TelarKG*, un grafo de conocimiento que almacena información relacionada con el proceso constituyente en Chile del año 2021. *TelarKG* es una iniciativa impulsada por el Instituto Milenio Fundamentos de los Datos (IMFD), que propone un conjunto de datos e información relevante para su empleo en estudios del evento. Facilita la realización de consultas específicas sobre entidades representadas en esta base de conocimientos. *TelarKG* actúa como un puente entre la información cruda obtenida de internet y aquellos interesados en investigar este fenómeno social.

Aunque *TelarKG* tiene limitaciones en su alcance, como la incapacidad de almacenar todos los tuits escritos en Chile durante el periodo, a través de filtros proporciona una muestra relacionada al tema abordado. En este trabajo, se incorporan técnicas de *entity linking* sobre los tuits almacenados en *TelarKG*, ampliando el modelo para incluir las menciones encontradas en cada tuit y las entidades a las que hacen referencia. Esta expansión permitirá buscar, por ejemplo, todos los tuits que hagan referencia a algún político o país en particular.

TelarKG se presenta en formato *QuadModel*. Para construir una base de datos a partir de esta información, se utiliza *MillenniumDB* [25], un sistema de gestión de bases de datos orientado a grafos, desarrollado por el IMFD. Las consultas que se pueden realizar en las bases de datos, creadas a partir de datos en formato *QuadModel*, se hacen usando el lenguaje de consultas *MQL*.

MillenniumDB es un motor de base de datos modular, persistente y de código abierto. Se basa en un modelo de datos gráfico llamado grafos de dominio, que proporciona una abstracción sencilla sobre la cual se pueden admitir una variedad de modelos de grafo populares. Esto ofrece un motor de gestión de datos flexible para diversos tipos de grafos de conocimiento. El motor se fundamenta en una combinación de técnicas de gestión de datos relacionales, algoritmos de vanguardia para uniones óptimas en el peor de los casos, así como algoritmos específicos para grafos para evaluar consultas de trayectorias [25].

El presente capítulo aborda la integración de los resultados de *entity linking* en *TelarKG*, comenzando con la sintaxis del formato *QuadModel* en el que están estructurados los datos de *TelarKG* en la Sección 5.1, seguido por la presentación del modelo del grafo abstracto inicial

en la Sección 5.2. Luego se detallan las modificaciones realizadas y se ofrece una justificación para las decisiones tomadas en la Sección 5.3. Concluyendo, se explora la sintaxis del lenguaje de consultas utilizado, acompañada de un ejemplo ilustrativo en la Sección 5.4.

5.1. Sintaxis de *QuadModel*

Los datos de entrada que recibe *MillenniumDB* pueden estar en formato *QuadModel* o *RDF*. Este último es un marco general para representar datos interconectados en la web, mientras que *QuadModel* es una extensión de este, lo que le facilita al último la administración de información de manera más eficiente [25].

Las declaraciones en *RDF* se conforman de tripletas que contienen:

- Sujeto: Representa el recurso principal, sobre el cual se está haciendo una afirmación.
- Predicado: Indica la relación o propiedad entre sujeto y objeto.
- Objeto: Es el objeto o el valor asociado al sujeto con la propiedad expresada por el predicado.

Por ejemplo, "Chile tiene como capital a Santiago" se podría expresar en tripletas donde el objeto es a su vez otro sujeto. Chile es el sujeto de la oración, "tiene como capital" el predicado y Santiago sería el objeto.

QuadModel es una extensión del modelo de tripletas, en el que se incorpora un cuarto elemento conocido como "identificador". Este componente adicional se utiliza para identificar cada una de las tripletas y proporciona la posibilidad de asociar información adicional. Por ejemplo, consideremos la tripleta `(Chile, tiene como capital, Santiago)`; al introducir el cuarto elemento como identificador, obtenemos un cuarteto como `(Chile, tiene como capital, Santiago, t1)`, donde "`t1`" corresponde al identificador de la tripleta inicial.

Este identificador se puede utilizar para asignar más información al cuarteto. Por ejemplo, podemos agregar un nuevo cuarteto como `(t1, desde el año, 1810, t2)`, donde la tripleta `(t1, desde el año, 1810)` ahora tiene el identificador "`t2`".

Existen dos tipos de sentencias permitidas en la sintaxis de *QuadModel*: La primera indica la existencia de una entidad dentro del modelo. Comienza con un identificador, seguido por cero o más etiquetas y luego se pueden agregar atributos de la entidad (metadatos) que pueden tener valores de enteros, flotantes, texto o booleano.

El segundo tipo de sentencia señala la relación entre dos entidades ya presentes en la base de datos. Inicialmente, se especifica el identificador de la entidad que actúa como el sujeto en esta conexión, seguido por `->` o `<-`, indicando la dirección de la relación. Posteriormente, se incluye el identificador de la entidad que funciona como objeto en este contexto, seguido por el predicado, que representa la naturaleza de la relación entre ambas.

A continuación se presenta un ejemplo de cada tipo. En el primer caso, se ilustra la existencia de dos entidades, las cuales se encuentran en los *Listings* 5.1 y 5.2. Posteriormente,

en el segundo caso (*Listing* 5.3), se muestra la relación entre estas dos entidades. Es importante señalar que cada entidad puede carecer de etiquetas y atributos, o bien, puede poseer múltiples etiquetas y atributos.

```
pact_YQ :Pact name:"YQ"
```

Listing 5.1: Definición de nodo *pact_ YQ*

```
part_CS :Party name:"CS"
```

Listing 5.2: Definición de nodo *pact_ CS*

```
part_CS->pact_YQ :pact
```

Listing 5.3: Definición de relación *:pact* entre nodo *pact_ YQ* y nodo *pact_ CS*

En los *Listings* 5.1 y 5.2 se definen dos entidades. La primera posee el identificador "`pact_YQ`", con la etiqueta "`Pact`" y un atributo llamado "nombre" con el valor "YQ". La segunda entidad, identificada como "`part_CS`", está etiquetada como "`Party`" y tiene un único atributo llamado "nombre" con el valor "CS". Luego, en el *Listing* 5.3, se establece que la entidad identificada como "`part_CS`" pertenece al pacto "`pact_YQ`". Una representación gráfica se muestra en la Figura 5.1, donde cada entidad (`part_CS` y `pact_YQ`) se representa como un nodo en el grafo. La conexión entre las entidades se indica con un arco, en este caso, de tipo `:pact`.

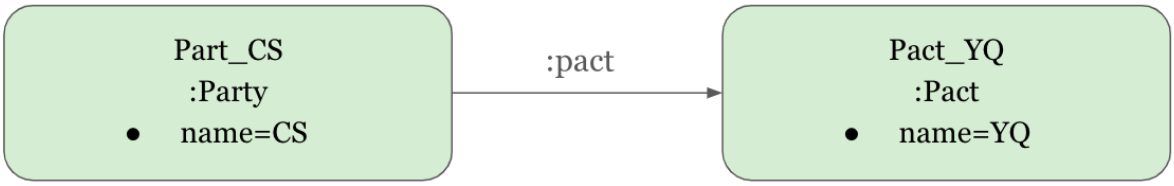

Figura 5.1: Grafo de relación entre *Pact_ YQ* y *Party_ CS*

5.2. Estado Inicial de *TelarKG*

TelarKG contiene datos vinculados al proceso constituyente del 2021 en Chile, ofreciendo información sobre los convencionales electos, sus partidos y pactos políticos, incluyendo las cuentas y algunas publicaciones de los convencionales en tres redes sociales: *Facebook*, *Instagram* y *Twitter*. Además, proporciona detalles sobre las votaciones realizadas, así como videos y transcripciones de las sesiones durante este proceso.

La estructura inicial del modelo de grafo abstracto de *TelarKG* se presenta en la Figura 5.2. En esta representación, cada cuadrado verde simboliza un tipo de nodo (entidad), con la

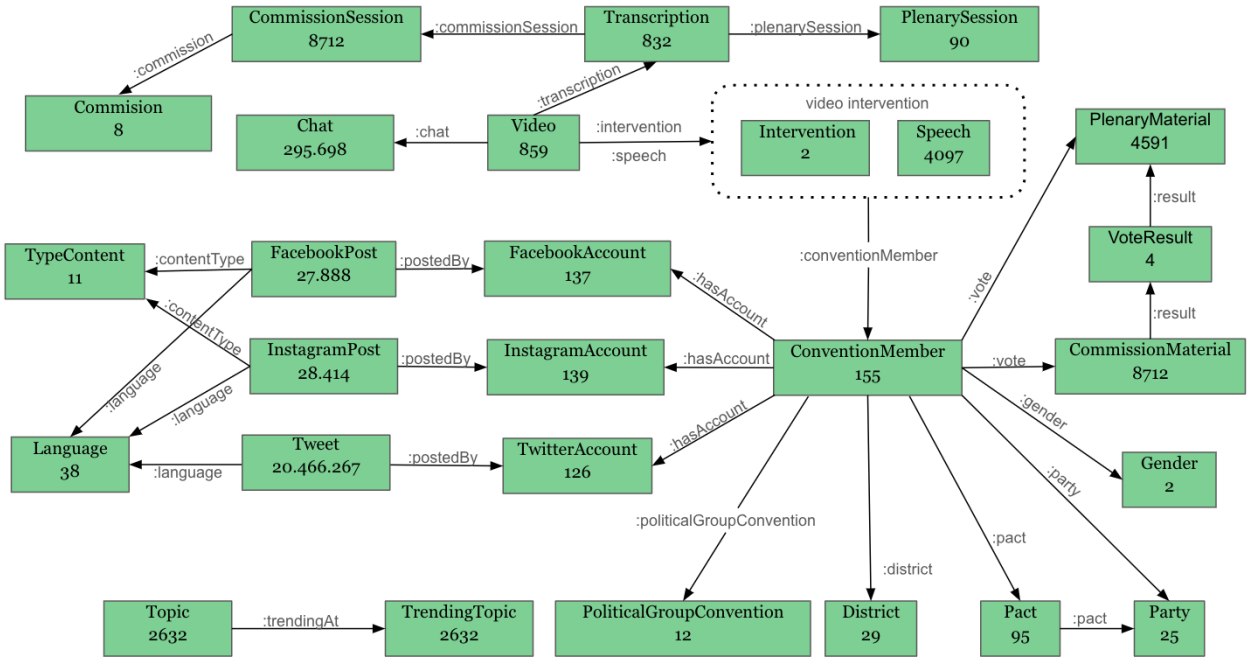

Figura 5.2: Esquema del grafo original de *TelarKG*
(Basado en el esquema disponible en `https://telarkg.imfd.cl/docs/` por Henry Rosales)

etiqueta correspondiente en la parte superior y la cantidad de nodos de ese tipo en *TelarKG* en la parte inferior. Los nodos de interés para nuestro análisis son: *ConventionMember*, *TwitterAccount*, *Tweet*, *Language*, *Gender*, *Pact* y *Party*. A continuación, se describen en detalle los atributos existentes en cada uno de estos tipos de nodos.

- **ConventionMember**: Nodo en el grafo que representa a los miembros de la convención constitucional. Los atributos de este tipo de nodos incluyen los siguientes:
  - inVoteName: Nombre completo del convencional constituyente.
  - Pact: Pacto al que pertenece.
  - name: Nombre y apellido del convencional constituyente.
  - age: Edad.
  - occupation: Ocupación del convencional constituyente.
  - web: Página web del convencional constituyente.
  - description: Pequeña descripción del convencional constituyente, haciendo referencia a su trabajo pasado.
  - photoId: Identificador de una foto del convencional constituyente.
  - photo: Booleano que indica si el convencional constituyente tiene o no una foto en la base de datos.
  - nombreEnPartido: Nombre al que se asocia este convencional constituyente a su partido.

  - wikidata: Página de *Wikidata* del convencional constituyente.

- **TwitterAccount**: Etiqueta asignada a los nodos en el grafo que representan una cuenta de *Twitter*. Posee los siguientes atributos:
  - id: Identificador único de la cuenta de *Twitter*.
  - strId: Identificador en otro formato.
  - location: Ubicación asociada a la cuenta.
  - name: Nombre con el que el propietario de la cuenta se registró en la plataforma.
  - createdAt: Fecha y hora de creación de la cuenta de *Twitter*.
  - description: Descripción que se encuentra en el perfil de *Twitter*.
  - protected: Booleano que indica si la cuenta es privada (solo personas aprobadas por el propietario pueden ver su contenido).
  - screenName: Nombre público de la cuenta de *Twitter*.
  - followersCount: Número de seguidores que tiene la cuenta.
  - url: URL de alguna página web asociada al propietario de la cuenta.
  - friendsCount: Número que indica la cantidad de cuentas que esta cuenta sigue en la plataforma.
  - profileImageUrlHttps: URL de la imagen de perfil en formato seguro (HTTPS).
  - listedCount: Número de listas en las que la cuenta está incluida.
  - verified: Indica si la cuenta está verificada.
  - favouritesCount: Número de tuits marcados como favoritos.
  - statusesCount: Número total de tuits publicados.
  - profileBannerUrl: URL de la imagen utilizada como banner en el perfil de la cuenta.
  - defaultProfile: Indica si la cuenta utiliza el diseño de perfil predeterminado.
  - defaultProfileImage: Indica si la cuenta utiliza la imagen de perfil predeterminada.
- **Tweet**: Etiqueta que poseen los nodos en el grafo que representa un tuit (publicación en *Twitter*). Sus atributos son los que siguen:
  - createdAt: Fecha y hora en la que el tuit fue creado.
  - isQuoteStatus: Indica si el tuit es una cita o respuesta a otro tuit.
  - id: Identificador único del tuit en la base de datos.
  - truncated: Indica si el texto del tuit ha sido truncado debido a su longitud.
  - favorited: Indica si el tuit ha sido marcado como favorito por algún usuario.
  - retweetCount: Número de retuits que ha recibido el tuit.
  - favoriteCount: Número de veces que el tuit ha sido marcado como favorito.
  - text: Texto completo publicado en el tuit.
  - retweeted: Booleano que indica si el tuit se ha sido retuiteado. Un retuit es la republicación de un tuit existente por parte de otro usuario.

- **Language**: Etiqueta que poseen los nodos que representan un lenguaje natural.
  - name: El nombre del idioma en inglés, por ejemplo, "Spanish" o "English".
  - code: Código que identifica al idioma, por ejemplo, "es" para español o "en" para inglés.
- **Gender**: Etiqueta de los nodos que representa un género.
  - name: Representa el nombre del género y solo puede tener uno de los siguientes valores: "female" (femenino) o "male" (masculino).
- **Pact**: Etiqueta que poseen los nodos que representan un pacto político (unión de partidos políticos).
  - name: Corresponde a la sigla con la que se identifica el nombre del pacto en particular.
- **Party**: Etiqueta de los nodos que representa un partido político.
  - name: Sigla con la que se conoce al partido político al que hace referencia.

5.3. Extensión de *TelarKG*

Para incorporar los enlaces a *Wikidata* derivados de la ejecución de *entity linking* a los tuits almacenados en *TelarKG*, fue necesario introducir nodos adicionales en el grafo inicial de *TelarKG*. Esto se debió a que no era viable almacenar toda la nueva información como propiedades de los nodos tipo *Tweet*. La razón radica en que, si bien existen tuits sin entidades mencionadas, otros presentan más de una mención. En el formato *QuadModel* (el formato de datos de *TelarKG*), no se permite el uso de listas como valores de una propiedad. Por lo tanto, fue necesario agregar nodos a la base de datos para representar las menciones a las entidades encontradas en los tuits.

Dado que los enlaces en las bases de datos *MillenniumDB* solo pueden tener una etiqueta y no admiten atributos adicionales, se optó por incluir nodos con etiqueta *"Mention"* como intermediarios entre los nodos de tipo *"Tweet"* y los nodos correspondientes a la entidad mencionada. Esto se debió a que existen metadatos relevantes sobre el enlace a la entidad hallada en el tuit, como la posición exacta en la que aparece la entidad y el sistema de *entity linking* que la identificó. Para esto, el nodo de tipo *"Mention"* posee los atributos *"text"*, *"start"*, *"end"* y *"system"*. Una alternativa sería conectar directamente los nodos de tipo *"Tweet"* a un nodo que represente la entidad correspondiente, pero esta opción elimina la información sobre la ubicación en el texto donde se menciona la entidad, considerada relevante y útil de almacenar. Además, es importante tener en consideración que la misma entidad de *Wikidata* puede ser referenciada en más de un tuit y, siendo una entidad, se representa cada una en un nodo diferente que puede estar conectado a uno o más tuits.

Los tuits que contienen menciones, identificadas tanto por *OpenTapioca* como por *DBPedia Spotlight*, están vinculados mediante la relación *:mention* a un nodo con etiqueta *Mention*. Estos nodos a su vez, están conectados mediante la relación *:entity* a un nodo etiquetado

como *Entity*, *ConventionMember* o *Party*. La elección depende de la entidad encontrada por el sistema de *entity linking*.

Con el objetivo de proporcionar información adicional sobre la entidad identificada, se añaden etiquetas que representan los tipos a los que pertenece dicha entidad. Estas etiquetas se obtienen mediante las herramientas de *entity linking*. Algunas de las etiquetas encontradas incluyen: *:Location*, *:Person*, *:Organization*, *:PoliticalParty*, *:Work*, *:Place*, entre otras.

Aunque la mayoría de los nodos de tipo *Entity* se crean durante el proceso de agregar las entidades mencionadas en los tuits a la base de datos, algunas entidades encontradas ya existen en *TelarKG*. Todos los constituyentes tienen asociado su identificador de *Wikidata* y también representan la entidad dentro de esta base de datos. Por lo tanto, si se encuentra una mención a cualquier convencional, esta mención se vincula a nodos etiquetados como *ConventionMember*. Para lograr lo mismo con los nodos de tipo *Party*, se añaden manualmente sus identificadores de *Wikidata* a los datos originales en formato *QuadModel*. Esta metodología se elige debido a que la cantidad de identificadores es manejable, facilitando así la eficiencia en la ejecución de la tarea. Además, resulta crucial desambiguar las siglas que identifican a las agrupaciones políticas. Es importante señalar que de las 25 agrupaciones políticas, un 20 %, es decir, 5 de ellas, carecen de entidades en *Wikidata*. Es relevante destacar que no se encuentran miembros de la convención constitucional asociados a tres de estos partidos.

En la Figura 5.3, se muestra la actualización del grafo abstracto de *TelarKG*, incluyendo las menciones a entidades de *Wikidata* encontradas en los tuits almacenados desde la versión inicial. Los nodos del grafo están representados por recuadros verdes, con la etiqueta escrita en su interior y el número de nodos de ese tipo debajo. Los nuevos tipos de nodos se destacan en negrita y con un color diferente.

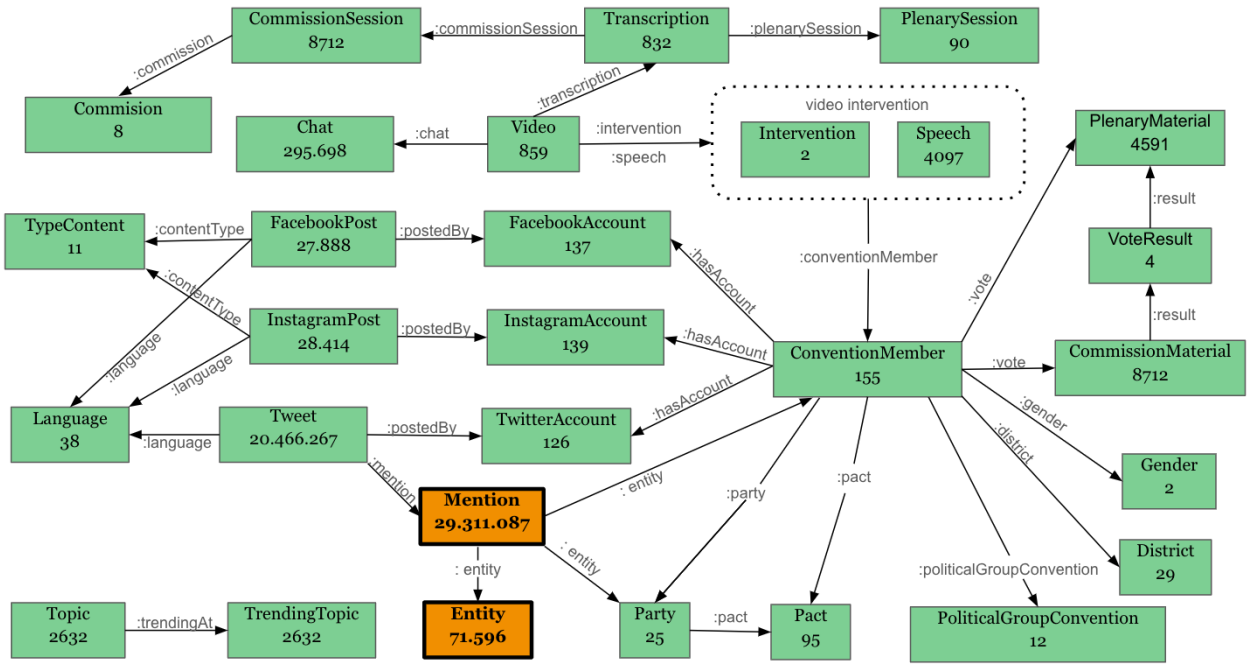

Figura 5.3: Esquema del grafo actualizado de *TelarKG*

Los nuevos tipos de nodo, incluyendo sus atributos, se detallan a continuación:

- ***Mention***: Etiqueta que poseen los nodos que representan la existencia de una mención de una entidad de *Wikidata* en un tuit.
  - text: Caracteres del tuit que hacen mención a la entidad de *Wikidata*.
  - start: Posición en el texto donde inicia el texto que hace referencia a la entidad.
  - end: Posición en el texto donde termina el texto que hace referencia a la entidad.
  - system: Sistema con el que se obtuvo la mención a la entidad. (*OpenTapioca* o *DBPedia Spotlight*).
- ***Entity***: Etiqueta que poseen los nodos que representan una entidad de *Wikidata*. tiene un único atributo llamado *"wikidata"* y corresponde al identificador de *Wikidata* de la entidad.

En el fragmento de código *QuadModel* proporcionado (*Listing* 5.4), se ilustra el proceso de creación y vinculación de entidades utilizando el modelo de cuartetos. Se inicia con la definición de dos entidades: una etiquetada como *Entity* con el identificador *wd_Q298*, que representa un lugar poblado de tipo *Location* y *Country* correspondiente a la entidad de *Wikidata*, con la URL *"http://www.wikidata.org/entity/Q298"*. La segunda entidad, etiquetada como *Mention* con el identificador *m1_dbps_twp_15*, representa la mención del texto "Chile" en una posición específica (caracteres 19 a 24) de un tuit en particular, identificada por el sistema *DBPedia Spotlight*.

La relación entre estas entidades se establece mediante la línea 4 del *Listing* 5.4, indicando que la entidad mencionada (*m1_dbps_twp_15*) está relacionada con la entidad de *Wikidata* (*wd_Q298*) a través de la conexión *:entity*. Además, se establece una conexión entre el tuit (*twp_15*) y la mención (*m1_dbps_twp_15*) mediante la línea 3 del mismo *listing*. Cabe destacar que los tuits ya se encuentran dentro de *TelarKG*, por lo que no debe ser definido.

```
1 wd_Q298 :PopulatedPlace :Place :Location :Country :Entity
    wikidata:"http://www.wikidata.org/entity/Q298"
2 m1_dbps_twp_15 :Mention text:"Chile" start:19 end:24 system:"
    DBPediaSpotlight"
3 twp_15->m1_dbps_twp_15 :mention
4 m1_dbps_twp_15->wd_Q298 :entity
```

Listing 5.4: Creación de entidades y relaciones en *QuadModel*: Ejemplo de vinculación de tuit a antidades de *Wikidata*

5.4. Consultas a *TelarKG*

QML destaca como el único lenguaje habilitado para la realización de consultas en bases de datos construidas con *MillenniumDB* que utilizan datos de entrada de tipo *QuadModel*. En esta sección se proporciona una detallada explicación de la sintaxis de *QML*, respaldada por un ejemplo práctico.

A continuación, se detallan los elementos de la sintaxis pertinentes para los objetivos de este trabajo.[1]:

- Declaración *MATCH*: Cada consulta debe iniciar con una declaración MATCH, seguida de un patrón de grafo que define nodos y relaciones, como se detalla con mayor precisión a continuación en la lista.
- Patrón de Nodo: Se define con paréntesis () y puede incluir identificadores, variables, etiquetas y propiedades. Los identificadores posibles son los que se muestran en la Tabla 5.1.

Tipo de Objeto	**Ejemplo**
Nodo nombrado	(Nombre_Apellido)
Nodo anónimo	(_a123)
Arco	(_e123)
Cadena Literal	"alguntexto"
Booleano Literal	true
Entero Literal	123
Flotante Literal	12,3

Tabla 5.1: Posibles identificadores de un nodo en *QML*

  Si no se utiliza algún identificador se pueden, o no, utilizar variables para representar al nodo de la siguiente manera: `(?x)`.

  También se pueden indicar una o más etiquetas, además de una variable que se quiera que posean los nodos, por ejemplo: `(?entidad :Entity :Person)`. En este caso se hace referencia a nodos que tengan ambas etiquetas *Entity* y *Person*.

  Además, se puede indicar el valor de uno o más atributos que posea el nodo que se quiere referenciar. La sintaxis para hacer referencia a todos los nodos que representan tuits que han sido retuiteados y marcados por alguien como favorito, por ejemplo, es como sigue: `(?tuit :Tweet retweeted:true, favorited:true)`.

- Patrón de arco: Conecta dos patrones de nodo y puede contener variables, tipos de borde y propiedades, aunque en la implementación actual de *QuadModel* no se encuentran disponibles propiedades en arcos o enlaces.

  Los arcos deben estar encerrados entre corchetes (`[ ]`) y tener una dirección `->` o `<-`, la cual indica de qué manera se relacionan los nodos a ambos extremos de estas.

  Un arco se representa por su nombre con la siguiente sintaxis: `(?x)-[_e12]->(?y)`. En este caso el arco tiene nombre "`_e12`".

  Las variables de arco se expresan como `?e`, siendo `e` un nombre a elección. Por otro lado las variables de tipo se describen como `:?e`, siendo `e` un nombre cualquiera. Un ejemplo en que se usen ambos es como sigue: `(?x)-[?e :?t]->(?y)`.

[1]La información detallada acerca de la sintaxis de *QML* se extrae de un archivo asociado a *MillenniumDB*. Para acceder a dicho archivo, consulte: `https://github.com/MillenniumDB/MillenniumDB/blob/main/doc/quad_model/query_language.md`

En vez de una variable de tipo también se puede fijar un tipo, por ejemplo como sigue: `(?x)-[?e :postedBy]->(?y)`. En este ejemplo se hace referencia a todos los pares de nodos que se relacionan entre sí con un enlace de tipo `:postedBy`.

- Patrón de Ruta: Conecta dos patrones de nodo mediante expresiones regulares de ruta, con direcciones y variables opcionales.

  Las rutas tienen cuatro componentes:

  1. Una expresión de ruta regular, similar a una expresión regular, pero que usa tipos en vez de caracteres como elemento atómico. La definición formal a continuación:

     Si t es un tipo válido, t es una expresión de camino regular. Si t es una expresión de camino regular:

     - $\hat{}t$ es una expresión de camino regular y representa que t ocurre en dirección invertida.
     - $t*$ es una expresión de camino regular y expresa que t ocurre 0 o más veces.
     - $t+$ es una expresión de camino regular y significa que t ocurre 1 o más veces.
     - *t?* es una expresión de camino regular e implica que t ocurre 0 o 1 vez.

     si *t1* y *t2* son expresiones de camino regular:

     - *t1 / t2* es una expresión de camino regular y representa que *t1* es seguido por *t2*.
     - *t1 | t2* es una expresión de camino regular y significa que ocurre *t1* o *t2*.
  2. Una variable de ruta opcional, como por ejemplo: `?variable`
  3. Una declaración semántica (ALL o ANY) para especificar si los resultados contendrán cualquier camino más corto (ANY) o todos los caminos más cortos (ALL) entre dos nodos. Si no se especifica ninguno, se asume ANY.
  4. Una dirección de la forma `=>` o `<=`

  Algunos ejemplos de caminos se encuentran en la siguiente lista:

  1. `(?x)=[:hasReply+]=>(?y)`: Une dos nodos donde el primero tiene una o más conexiones con otros nodos a través de un arco de tipo *:hasReply*.
  2. `(:Baquedano)=[ANY ?p :Linea1+/:Linea3?]=>(:Santa_Ana)`: Devuelve el camino más corto desde *:Baquedano* hasta *:Santa_Ana*, atravesando uno o más tramos de la línea 1 y cero o un tramo de la línea 3. La declaración semántica ANY especifica que se devuelva el camino más corto encontrado primero.

- Declaración *WHERE*: Opcional, filtra los resultados según una expresión.
- Declaración *ORDER BY*: Opcional, ordena los resultados según criterios específicos.
- Declaración *RETURN*: Esencial, especifica qué objetos o propiedades se devolverán en los resultados.
- Declaración *LIMIT*: Opcional, establece un límite en el número de resultados devueltos.

```
//Esta consulta pide el nombre y la edad de miembros de la convención
// que pertenecen al pacto político Pact1 y que tienen entre 60 y 70 años.
//Se ordenan los resultados por su edad de manera ascendente y por nombre de manera descendente.
MATCH (?x :ConventionMember)-[:pact]->(Pact1)
WHERE ?x.age => 60 AND ?x.age <= 70
ORDER BY ?x.name DESC, ?x.age ASC
RETURN ?x.age, ?x.name
LIMIT 1000
```

Listing 5.5: Consulta en *QML*

En el *Listing* 5.5, utilizando *TelarKG* como ejemplo, se presenta una consulta en formato *QML*, destinada a obtener información sobre miembros de la convención constitucional que pertenecen al pacto político *Pact1* y tienen edades comprendidas entre 60 y 70 años. Los resultados se ordenan de manera ascendente por edad y de manera descendente por nombre. A continuación, se ofrece una descripción detallada de los elementos que componen la consulta:

Con *MATCH* se busca un nodo `(?x)` de tipo *ConventionMember* que esté conectado mediante una relación *pact* al nodo *Pact1*.

Con la declaración *WHERE* se filtran los miembros cuya edad esté entre 60 y 70 años.

Debido a la declaración *ORDER BY* en esta consulta, se ordenan los resultados de manera ascendente por edad y de manera descendente por nombre.

La Declaración *RETURN* indica que se devuelven las propiedades de edad y nombre del miembro de la convención (?x).

Finalmente, la declaración *LIMIT* limita el número de resultados a 1.000.

La consulta descrita en el *Listing* 5.5 proporciona una visión detallada de cómo estructurar y personalizar las consultas para obtener información específica de un grafo de conocimiento. Es importante destacar que la información sobre la sintaxis de *QML* se ha obtenido a partir de un archivo de *MillenniumDB*.

5.5. Ampliación Práctica y Consultas en *TelarKG*

El término *dump* hace referencia al archivo que contiene tanto la estructura como los datos de una base de datos. Para generar el *dump* extendido de *TelarKG*, el primer paso consiste en definir todas las menciones y entidades de *Wikidata* que aún no están presentes en *TelarKG*, es decir, todas las menciones y todas las entidades que no corresponden a partidos políticos presentes en *TelarKG*, ni miembros de la convención constitucional. Además deben expresarse las conexiones entre las menciones y los tuits, todo ello en formato *QuadModel*. Es importante señalar que todos los tuits ya están declarados en *TelarKG*, por lo que no se debe repetir este

proceso.

Una vez que la información se ha organizado en el formato adecuado, se añade a continuación del archivo inicial del *dump* de *TelarKG*. Posteriormente, se lleva a cabo la indexación de los datos actualizados utilizando *MillenniumDB*, lo que resulta en una base de datos con 49.720.016 nodos, 61.297.839 conexiones, 49.752.376 etiquetas en total, 198 etiquetas diferentes y 380.127.784 propiedades en total.

En la Sección 6.3.2 se presentan los resultados de las consultas que se detallan en los *Listings* 5.6, 5.7, 5.8 5.9 y 5.10.

```
MATCH (?c :ConventionMember)->(?ta :TwitterAccount)<-(?t :Tweet
   )->(?m :Mention)->(?cc :ConventionMember)
WHERE ?c.name == ?cc.name
RETURN ?c, ?c.name, ?t, ?m.system, ?m.start, ?m.end,  ?m.text,
   ?cc
```

Listing 5.6: Consulta que retorna a los convencionales constituyentes que se mencionan a si mismos en un tuit

```
MATCH (?p :Party)<-(?j :ConventionMember)->(?t :TwitterAccount)
   <-(?x :Tweet)->(?y :Mention)->(?z :ConventionMember)->(?pp :
   Party)
WHERE ?p.name == ?pp.name AND ?j.name != ?z.name
RETURN ?p, ?j.name, ?y.text, ?y, ?z.name, ?pp
```

Listing 5.7: Consulta que retorna las menciones hechas por diferentes convencionales constituyentes que pertenecen al mismo partido político

```
MATCH (?j :ConventionMember)->(?t :TwitterAccount)<-(?x :Tweet)
   ->(?y :Mention)->(?z :ConventionMember)
WHERE ?j != ?z
RETURN ?j.name, ?x,  ?y, ?y.system, ?y.text, ?z.name
```

Listing 5.8: Consulta que retorna los tuits donde un convencional constituyente menciona a otro

```
MATCH (?p :Party)<-(?j :ConventionMember)->(?t :TwitterAccount)
   <-(?x :Tweet)->(?y :Mention)->(?pp :Party)
WHERE ?p.name != ?pp.name
RETURN ?y.text, ?y, ?y.system, ?pp
```

Listing 5.9: Consulta que retorna los tuits hechos por convencionales que mencionan a un partido político diferente al que pertenece

```
MATCH (?p :Party)<-(?j :ConventionMember)->(?t :TwitterAccount)
   <-(?x :Tweet)->(?y :Mention)->(?z :ConventionMember)->(?pp :
   Party)
WHERE ?p.name != ?pp.name
RETURN ?p, ?j.name, ?y.text, ?y, ?z.name, ?pp
```

Listing 5.10: Consulta que retorna la relación entre partidos políticos que fueron mencionados por algún convencional constituyente de otro partido político

Las consultas hechas a la versión extendida de *TelarKG* buscan resolver las siguientes preguntas:

1. ¿Cuáles son las personas que se mencionan a sí mismas con mayor frecuencia en los tuits?
2. ¿Cuál es la frecuencia de menciones entre convencionales de un mismo partido?
3. ¿Existen pares específicos de convencionales que se mencionan entre sí con mayor frecuencia?
4. ¿Cuáles son los partidos más mencionados por convencionales de otros partidos?
5. ¿Con qué frecuencia los convencionales de dos partidos políticos se mencionan mutuamente, revelando una relación destacada entre dichos partidos?

La pregunta 1 se responde mediante los datos que *TelarKG* proporciona al realizar la consulta del *Listing* 5.6. Esta consulta ofrece las menciones que los convencionales constituyentes hacen a sí mismos en los tuits de *TelarKG*.

Con respecto a la pregunta 2, los datos necesarios para responderla se obtienen al aplicar la consulta del *Listing* 5.7 a *TelarKG*, la cual entrega todas las menciones hechas por un convencional constituyente a otro del mismo partido.

Es necesario adquirir los datos resultantes de la ejecución de la consulta indicada en el *Listing* 5.8 para abordar la pregunta 3. Dicha consulta recupera la totalidad de los tuits generados por un convencional constituyente que menciona a otro convencional constituyente.

El *Listing* 5.9 presenta la consulta específica que debe realizarse en *TelarKG* con el fin de responder a la pregunta 4. Esta consulta proporciona todos los tuits en los cuales un convencional ha mencionado a un partido político distinto al que pertenece.

Para responder a las preguntas 5, es necesario obtener los datos de *TelarKG* que entrega la *query* del *Listing* 5.10. Esta consulta proporciona todas las menciones que un convencional constituyente hace a otro, donde el segundo debe pertenecer a otro partido político.

Capítulo 6

Experimentos y Evaluación

Para evaluar la viabilidad del enfoque propuesto en un entorno práctico, es esencial abordar varias preguntas relacionadas con las limitaciones temporales, de memoria y de ejecución del proceso de *entity linking*. Este desafío se ve acentuado por dos componentes fundamentales: la extensa cantidad de datos que se pretende etiquetar y la naturaleza intrínsecamente imperfecta de la tarea de *entity linking*, que es una tarea de Procesamiento del Lenguaje Natural (PLN). Hasta la fecha de esta publicación, alcanzar resultados perfectos en esta tarea se considera imposible con las tecnologías actuales.

Con el objetivo de validar y comprender el alcance potencial de este proyecto, se llevan a cabo diversos experimentos. Los cuáles surgen como respuesta a las siguientes preguntas clave:

1. ¿Es factible, en términos de tiempo y recursos, llevar a cabo el etiquetado de la totalidad de los tuits?
2. En caso afirmativo, ¿cuánto tiempo se estima que tomaría completar dicho proceso?
3. ¿Existen sistemas capaces de identificar entidades de manera efectiva tanto en inglés como en español, incluso en textos que no sigan necesariamente el formato buscado?
4. ¿Cuál es el nivel de precisión de las etiquetas generadas por estos sistemas?
5. ¿La búsqueda de entidades en textos carentes de contexto con formato de tuit demuestra un rendimiento satisfactorio en términos de eficacia y precisión?
6. Asumiendo que se pueden etiquetar los tuits, ¿los resultados de *entity linking* pueden habilitar análisis novedosos en el contexto de *TelarKG*?

Las respuestas a estas cinco preguntas se detallan en este capítulo. Las preguntas 1 y 2 encuentran su respuesta en la Sección 6.1, mientras que las preguntas 3 y 4 son abordadas en la Sección 6.2.1. La pregunta 5 se resuelve en la Sección 6.2.2. Finalmente, se busca la respuesta a la pregunta 5 en la Sección 6.3.

La ejecución de las tareas computacionales descritas en este capítulo se llevó a cabo en una máquina con las siguientes especificaciones:

- Información de la CPU
  - Arquitectura: x86_64
  - Modo de operación de la CPU: 32-bit, 64-bit
  - Orden de bytes: Little Endian
  - Número de CPUs: 30
  - Modelo de CPU: Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz
  - Frecuencia de la CPU: 2099,998 MHz
  - Tipo de virtualización: Xen (full virtualization)
- Tipo de Disco y Capacidad
  - Disco: xvda1
  - Tamaño: 3,6 TiB
- Capacidad de Memoria
  - Memoria total: 118 GiB
  - Memoria en uso: 35 GiB
  - Memoria libre: 1,2 GiB
  - Swap: 0B
- Sistema Operativo
  - Nombre: Devuan GNU/Linux 4 (chimaera)
  - Versión: 4 (chimaera)
  - ID: devuan
  - Tipo de sistema operativo: Devuan GNU/Linux
- Versión del Kernel
  - Linux s05-vm01 5.10.0-23-amd64 #1 SMP Debian 5.10.179-3 (2023-07-27) x86_64 GNU/Linux
- Conexión de Red
  - Red de la Universidad de Chile

6.1. Evaluación de Tiempos de Etiquetado

En esta sección, exploramos el rendimiento de los sistemas de *entity linking* (*TagMe*, *DBPedia Spotlight* y *OpenTapioca*) en diferentes configuraciones experimentales. Para respaldar la elección de agrupar los tuits en lotes, se presenta un experimento que estima los tiempos que cada sistema requeriría para procesar diferentes cantidades de peticiones y tuits.

En la Sección 6.1.1, analizamos el experimento diseñado para estimar el tiempo total de etiquetado de tuits utilizando los sistemas mencionados. Se evaluaron diversas configuraciones, variando la cantidad de peticiones y la cantidad de tuits por petición. Los resultados se presentan en la Tabla 6.1, proporcionando información sobre el tiempo estimado para procesar diferentes combinaciones de peticiones y tuits.

La Subsección 6.1.2 se centra en los tiempos reales de etiquetado para los tuits de las tablas *tw-convencionales* y *tw-streaming* de *TelarKG*. Se comparan los tiempos de *OpenTapioca* y *DBPedia Spotlight*, excluyendo el sistema *TagMe* debido a sus resultados deficientes en otros experimentos. La Tabla 6.2 presenta los resultados para la tabla *tw-convencionales*, y la Tabla 6.3 para la tabla *tw-streaming*. Estos resultados expresan el rendimiento real de los sistemas en la situación práctica del caso de uso escogido.

6.1.1. Rendimiento de Peticiones en Lotes

Para respaldar la decisión de agrupar los tuits en lotes, se llevó a cabo un experimento para estimar los tiempos que los tres sistemas tardarían en procesar todos los tuits. Este experimento consiste en medir el tiempo que cada sistema demora en etiquetar 1.000 peticiones de un tuit cada una, 10 peticiones de 100 tuits, 100 peticiones de 10 tuits y 1 petición de 1.000 tuits. Para calcular el tiempo total, se asume que el tiempo que tardarán las APIs en etiquetar se comporta de manera lineal. Por lo tanto, para estimar el tiempo total en segundos, se multiplica el tiempo obtenido en el experimento por 20.000, pues la muestra toma aproximadamente 1 de cada 20.000 tuits del conjunto total.

Al realizar este experimento, no se incluyeron tiempos de espera. Sin embargo, al procesar 20 millones de tuits, es necesario introducir un intervalo de espera de 2 segundos entre cada solicitud para evitar sobrecargar la API. Por lo tanto, el tiempo total de etiquetado también se estima teniendo en cuenta estos períodos de espera. En la Tabla 6.1, se detallan los tiempos estimados para cada sistema de *entity linking*.

Tabla 6.1: Estimación del tiempo total de etiquetado para diferentes configuraciones de peticiones y tuits

Tiempos:	**Experimento (s)**			**Total Estimado (días)**			**Total Estimado con Espera (días)**		
Cantidad de Peticiones	**TM**	**DB**	**OT**	**TM**	**DB**	**OT**	**TM**	**DB**	**OT**
1.000 peticiones de 1 tuit	888,6	770,3	927,6	205,9	178,3	214,7	668,9	641,3	677,7
100 peticiones de 10 tuits	167,5	84,4	116,6	38,8	19,5	37,6	85,1	65,8	83,9
10 peticiones de 100 tuits	18,6	10,8	17,2	4,3	2,5	4,0	8,9	7,1	8,6
1 petición de 1.000 tuits	2,2	1,1	2,0	0,5	0,2	0,5	1,0	0,7	1,0

La Tabla 6.1 presenta los resultados de un experimento diseñado para estimar el tiempo total de etiquetado de tuits utilizando tres herramientas de extracción de entidades (OT, DS y TM que representan a *Opentapioca*, *DBPedia Spotlight* y *TagMe*, respectivamente). Se han evaluado diferentes escenarios, variando la cantidad de peticiones y la cantidad de tuits por petición:

1. En el primer escenario, se realizaron 1.000 peticiones, cada una solicitando información de un solo tuit. Los tiempos de respuesta para cada herramienta oscilaron entre 770,3

y 927,6 segundos, y el tiempo total estimado para procesar todas las peticiones se extendería desde 178,3 hasta 214,7 días sin considerar tiempos de espera.

2. En el segundo escenario, se redujo el número de peticiones a 100, pero se aumentó la cantidad de tuits por petición a 10. Los tiempos de procesamiento disminuyeron en comparación con el primer escenario, variando entre 84,4 y 167,5 segundos por herramienta. El tiempo total estimado para esta configuración oscilaría entre 19,5 y 38,8 días.

3. En el tercer escenario, se realizaron 10 solicitudes, cada una compuesta por 100 tuits. Los tiempos de respuesta experimentaron una disminución adicional, oscilando entre 10,8 y 18,6 segundos. El tiempo total estimado para procesar estas peticiones varía de 2,5 a 4,3 días sin incluir tiempos de espera, mientras que, al considerar los tiempos de espera, se extiende desde 7,1 hasta 9,8 días.

4. En el último escenario, se ejecutó una única petición que solicitaba información sobre 1.000 tuits. Los tiempos de procesamiento más bajos se registraron en este caso, con valores entre 1,1 y 2,2 segundos. El tiempo total estimado para procesar todos los datos, utilizando lotes de este tamaño, oscilaría entre 0,2 y 0,5 días, y al incorporar los tiempos de espera, se estima un tiempo entre 0,7 a 1 día.

Estos resultados proporcionan una perspectiva sobre el rendimiento y los tiempos de respuesta de las herramientas de extracción de entidades en diferentes configuraciones experimentales.

Es importante resaltar que *DBPedia Spotlight* exhibe tiempos menores en cada configuración del experimento. A su vez, *OpenTapioca* y *TagMe* muestran resultados de tiempos similares, siendo *OpenTapioca* el que registra menor tiempo en las configuraciones 2, 3 y 4.

Los tiempos de ejecución están condicionados por las capacidades de la máquina utilizada y la velocidad de la conexión a Internet empleada para realizar las peticiones HTTP. Con una máquina diferente que admita una mayor paralelización mediante la utilización de más hilos, o con una conexión a Internet más robusta, se podrían lograr tiempos considerablemente reducidos para todos los experimentos centrados en la evaluación del tiempo. Cabe mencionar que esto sería con el riesgo de sobrecargar la API. La relevancia de este experimento radica en la evaluación de la viabilidad de etiquetar todos los tuits con los recursos materiales disponibles.

Los resultados presentados en la Tabla 6.1 indican que, con el objetivo de etiquetar la totalidad de los tuits en un intervalo de tiempo razonable, es necesario enviar lotes de 100 tuits o más. Estos hallazgos sugieren que el rendimiento, en cuanto a velocidad, de los sistemas de *entity linking* experimenta mejoras significativas al procesar lotes de mayor tamaño. La decisión de enviar lotes de al menos 100 tuits se respalda en la eficacia demostrada por los sistemas en términos de tiempo de tiempo de procesamiento. En consecuencia, enviar más peticiones HTTP con menos texto resultaría en un tiempo de etiquetado excesivo.

6.1.2. Tiempos Efectivos de Etiquetado

TagMe fue descartado debido a sus resultados deficientes en Precisión en el experimento realizado con *SemEval* (ver Sección 6.2.1) y por ser el sistema de *entity linking* con peores resultados generales en el experimento de etiquetado manual explicado en la Sección 6.2.2. Por esta razón, los resultados a continuación solo consideran los sistemas *DBPedia Spotlight* y *OpenTapioca*.

En una primera ronda, se realizó *entity linking* en los tuits publicados por los convencionales (aquellos que se encuentran en la tabla *tw-convencionales*), siendo este un conjunto de tuits de menor tamaño pero de mayor prioridad. En esta sección, se presentan primero los resultados obtenidos, con respecto al tiempo, al procesar la tabla *tw-convencionales* y luego los tiempos resultantes al etiquetar la tabla *tw-streaming*.

6.1.2.1. Tabla *tw-convencionales*

OpenTapioca emplea alrededor de 10.453,3 segundos, equivalente a unas 3 horas, para etiquetar los 600.000 tuits publicados por los convencionales en la tabla *tw-convencionales*, teniendo en cuenta los tiempos de espera establecidos. En cambio, el tiempo total de las solicitudes sin considerar el tiempo de espera es de 7.103,6 segundos, aproximadamente 2 horas. El tiempo promedio por petición es de alrededor de 4,6 segundos, con una desviación estándar de 0,73. Se realizaron un total de 1.541 peticiones HTTP.

Por otro lado, *DBPedia Spotlight* tarda 15.742,83 segundos (4,37 horas) en etiquetar todos los datos de la tabla *tw-convencionales*, considerando los tiempos de espera entre peticiones. Sin tener en cuenta los tiempos de espera, este sistema tarda 5.924,83 segundos (1,65 horas) en completar todas las peticiones. El tiempo promedio por petición es de 1,21 segundos, con una desviación estándar de 0,22. La cantidad total de peticiones realizadas es de 4.909.

Tabla 6.2: Comparación de tiempos de etiquetado entre *OpenTapioca* y *DBPedia Spotlight* para tabla de tuits *tw-convencionales*

	Tiempo total (h)	**Tiempo sin esperas (h)**	**Promedio por petición (s)**	**Desviación estándar**
DBPedia Spotlight	4,4	1,7	1,2	0,2
OpenTapioca	2,9	1,9	4,6	0,7

6.1.2.2. Tabla *tw-streaming*

El tiempo de procesamiento necesario para que *OpenTapioca* etiquetara todos los tuits en la tabla *tw-streaming*, considerando los tiempos de espera, fue de aproximadamente 92 horas. Excluyendo los tiempos de espera, esta tarea se completó en 66,6 horas. Durante este proceso, se realizaron un total de 45.663 peticiones, con un tiempo promedio de 5,3 segundos por petición y una desviación estándar de 1,5.

En comparación, *DBPedia Spotlight* requirió aproximadamente 305,8 horas para etiquetar todos los tuits, incluyendo los tiempos de espera. Excluyendo los tiempos de espera, este

proceso tomó 100,5 horas. Se llevaron a cabo un total de 369.720 peticiones, con un tiempo promedio por petición de 0,98 segundos y una desviación estándar de 0,18.

Tabla 6.3: Comparación de tiempos de etiquetado entre *OpenTapioca* y *DBPedia Spotlight* para tabla de tuits *tw-streaming*

	Tiempo total (h)	**Tiempo sin esperas (h)**	**Tiempo promedio por petición (s)**	**Desviación estándar**
DBPedia Spotlight	305,86	100,46	0,98	0,18
OpenTapioca	91,98	66,61	5,25	1,52

6.2. Evaluación de la Calidad de Enlaces

Para determinar la eficacia de los sistemas seleccionados en la identificación de entidades, es esencial someterlos a evaluación. En el ámbito del Procesamiento del Lenguaje Natural y los modelos de aprendizaje automático (*machine learning*), se recurre habitualmente a tres métricas fundamentales para medir el rendimiento de dichos modelos: Precisión, Recuperación y F1 [3]. Cada métrica ofrece diferentes perspectivas para evaluar el desempeño de los modelos. En la evaluación de modelos de *entity linking* con estas métricas, resulta fundamental disponer de las entidades esperadas, es decir, los resultados ideales. Posteriormente, el conjunto de respuestas esperadas y obtenidas por cada sistema se segmenta en tres subconjuntos excluyentes.

- Verdaderos Positivos (TP): Representa el número de muestras correctamente predichas como positivas. En el contexto de *entity linking*, se refiere a todas las entidades identificadas correctamente, es decir, aquellas que estaban presentes en los resultados esperados.
- Falsos Positivos (FP): Indica el número de muestras equivocadamente predichas como positivas. En nuestra evaluación, esto se refiere a todas las entidades identificadas por el sistema que no estaban incluidas en las respuestas esperadas.
- Falsos Negativos (FN): Representa el número de muestras incorrectamente predichas como negativas. En nuestro contexto, son las entidades esperadas que el sistema de *entity linking* no logró identificar.

$$\text{Precisión} = \frac{TP}{TP + FP} \tag{6.1}$$

$$\text{Recuperación} = \frac{TP}{TP + FN} \tag{6.2}$$

$$\text{Puntaje F1} = \frac{2 \cdot \text{Precisión} \cdot \text{Recuperación}}{\text{Precisión} + \text{Recuperación}} \tag{6.3}$$

Utilizando las clasificaciones mencionadas, se calculan la Precisión, Recuperación y Puntaje F1 según las fórmulas recién proporcionadas, en particular, las formulas (6.1), (6.2) y (6.3), respectivamente. En primer lugar, la precisión representa la proporción de entidades relevantes encontradas con respecto a todas las entidades identificadas; es decir, evalúa la calidad de las entidades descubiertas por el sistema. Por otro lado, la recuperación indica la proporción de entidades correctamente identificadas por el sistema en comparación con todas las entidades que se esperaba encontrar, ofreciendo una medida de la capacidad del modelo para descubrir las entidades esperadas. Finalmente, el puntaje F1 se calcula como el promedio armónico de la recuperación y la precisión. Todas estas métricas pueden variar entre 0 y 1, ambos inclusive. Un valor más alto en cada métrica indica una mayor calidad en los resultados proporcionados por el modelo en esa dimensión de evaluación.

Tras la evaluación de los diversos sistemas de *entity linking*, se toma la decisión de cuál o cuáles se incorporarán en el desarrollo del proyecto. Las opciones consideradas incluyen la utilización exclusiva de un sistema, así como la combinación mediante la unión o intersección de las entidades identificadas por dos o más sistemas. Estas alternativas plantean desafíos debido a las distintas bases de datos de entidades empleadas por los sistemas de *entity linking*. Sin embargo, como se mencionó previamente en la Sección 2.4.1, todas las entidades pueden ser encontradas en *Wikidata*, lo que permite unificar los resultados obtenidos por los diferentes sistemas.

Para abordar la pregunta que motivó esta sección, la dividiremos en dos partes. En primer lugar, examinaremos el rendimiento de los sistemas en diferentes idiomas utilizando textos formales y bien escritos, que proporciona más contexto que un tuit y, por ende, puede generar entidades más relevantes. Posteriormente, evaluaremos si estos resultados son consistentes al evaluar sobre tuits reales. Especificaciones sobre cada una de estos enfoques se encuentran en las Secciones 6.2.1 y 6.2.2 de este capítulo.

6.2.1. *SemEval*

Con la finalidad de evaluar y comparar los resultados obtenidos usando los diferentes sistemas de *entity linking* se utilizó *SemEval*[1]: Serie de evaluaciones de sistemas computacionales que resuelven dos tipos de problemas en el campo del procesamiento de lenguaje natural. Estos son Desambiguación del Sentido de las Palabras* y *Entity Linking* (*WSD* y *EL* sus siglas en inglés).

WSD o Desambiguación del sentido de las palabras es una tarea del campo de procesamiento de lenguaje natural, que busca determinar el sentido o significado correcto de una palabra dado un contexto. Por ejemplo, el significado de la palabra "interés/es" en las siguientes frases tienen significados diferentes.

1. Pedí un préstamo con una tasa del 15 por ciento de *interés*
2. Ha expresado tener *interés* por las actividades solitarias.

[1] `https://alt.qcri.org/semeval2015/task13/index.php?id=data-and-tools`

La palabra interés en la primera frase hace referencia al lucro producido por el banco a causa del capital. En cambio en la segunda frase hace referencia a las inclinaciones de ánimo y personales de la persona de quien se habla.

Al utilizar esta herramienta para evaluar los tres sistemas candidatos de *entity linking*, se realizó una alteración a los datos de salida que provee *SemEval*. La versión de *SemEval* utilizada se denomina *SemEval-2015 Task 13*, y tiene como datos de entrada algunos textos de diversos temas e idiomas y evalúa los resultados al comparar los entregados por el sistema a prueba con los datos de salida esperados. Las etiquetas esperadas no contemplan solamente las entidades de *Wikidata*, sino que también incluye entidades de *BabelNet* y *WordNet* (otras fuentes de entidades). Dado que todos los sistemas de *entity linking* empleados utilizan *Wikidata* como fuente principal de entidades, o permiten la transformación de las entidades identificadas a formato *Wikidata*, se optó por filtrar los resultados esperados, eliminando aquellos que carecían de una etiqueta asociada en *Wikidata*. Este enfoque tiene como objetivo evaluar el desempeño real de los sistemas de *entity linking* seleccionados previamente, excluyendo falsos negativos, es decir, las etiquetas que no tienen correspondencia en *Wikidata*; estas etiquetas eliminadas están más enfocadas en la tarea de WSD que la tarea de *entity linking*.

El gráfico presentado en la Figura 6.1 representa el rendimiento de varios sistemas de *entity linking* evaluados en el contexto de la competición *SemEval*. La evaluación se realiza utilizando tres métricas clave: Precisión, Recuperación y F1. Es importante destacar que la evaluación en *SemEval* se hizo utilizando los resultados esperados en el mismo idioma con el que se configuraron los sistemas de *entity linking*.

- Eje X (Horizontal): Cada barra en el gráfico representa un sistema de *entity linking*, o una combinación de ellos, en una configuración específica de idioma. Los sistemas incluidos son *TagMe*, *DBPedia Spotlight* (en inglés y español), y *OpenTapioca* (en inglés y español). En este contexto, "en" se refiere a inglés, "it" a italiano y "es" a español. Además, las notaciones *TM*, *DS* y *OT* representa los sistemas *TagMe*, *DBPedia Spotlight* y *Opentapioca*, respectivamente. Vale recordar que *TagMe* no tiene una configuración en español; se seleccionó el idioma más parecido a este.
- Eje Y (Vertical): La escala en el eje y varía de 0 a 1 e indica el rendimiento de los sistemas en términos de Precisión, Recuperación y F1.
- Barras de Colores: Cada grupo de tres barras de colores diferentes representa la Precisión (en azul), Recuperación (en rojo) y F1 (en café) respectivamente para cada sistema e idioma evaluado.

Para facilitar la interpretación de los datos expuestos en el gráfico de la Figura 6.1 se comenta cada configuración de sistema e idioma por separado:

- *TagMe* (en inglés): Exhibe una buena Precisión (0,7), pero Recuperación (0,4) y F1 (0,5) más bajas.
- *TagMe* (en italiano): Similar a la configuración en inglés.

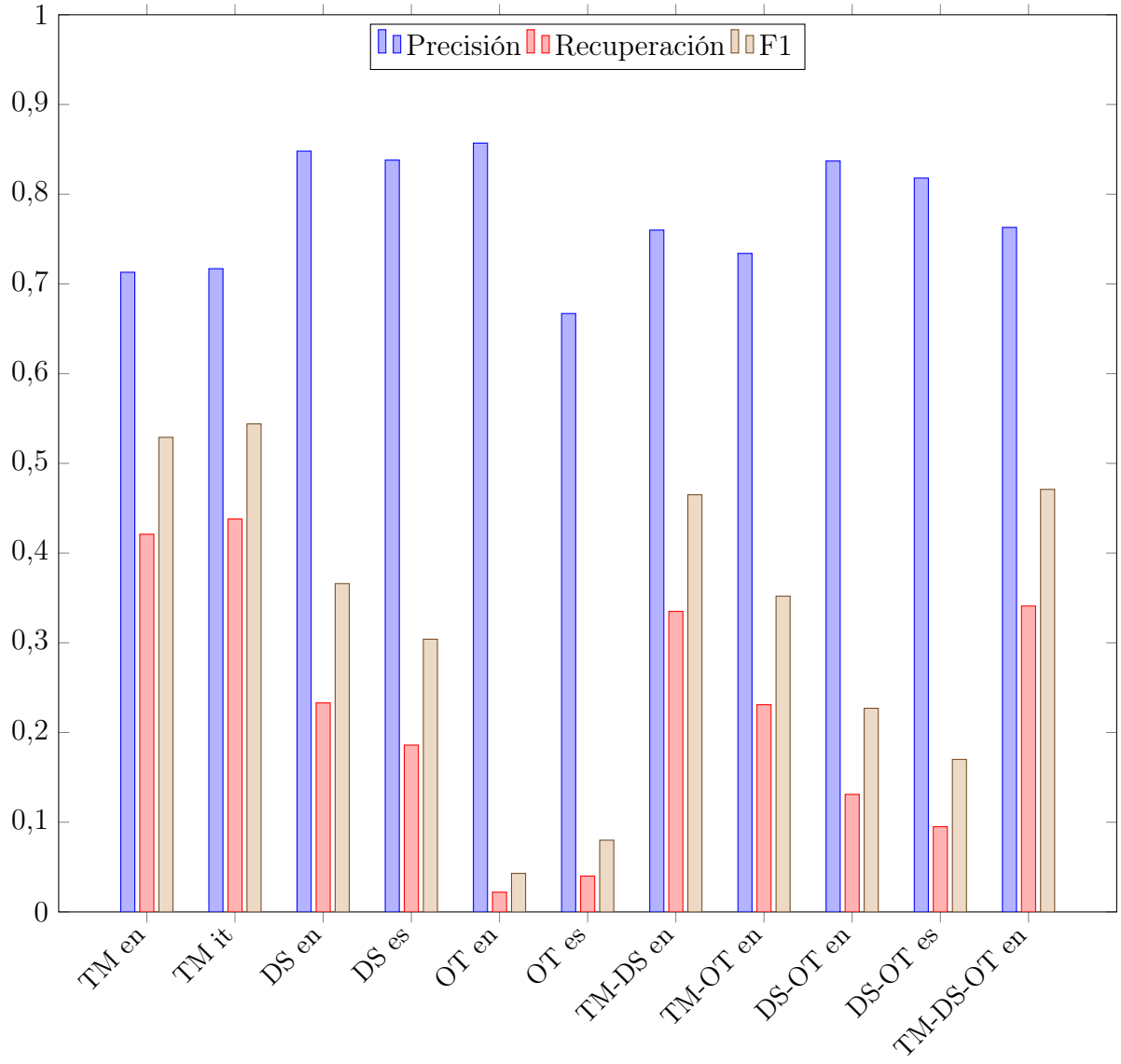

Figura 6.1: Desempeño de los sistemas en evaluación con *SemEval*: Precisión, Recuperación y F1.

- *DBPedia Spotlight* (en inglés): Alta Precisión (0,83), pero mucho menor Recuperación (0,25) y F1 (0,35).
- *DBPedia Spotlight* (en español): Similar a la configuración en inglés, pero con un rendimiento general más bajo.
- *OpenTapioca* (en inglés): Alta Precisión (0,85), pero muy baja Recuperación (0,02) y F1 (0,04).
- *OpenTapioca* (en español): Precisión (0,7) significativamente más alta que la Recuperación (0,04) y F1 (0,08).
- *TagMe-DBPedia Spotlight* (en inglés): Exhibe una buena Precisión (0,76), pero una Recuperación (0,34) y F1 (0,47) más bajas.

- *TagMe-OpenTapioca* (en inglés): Buena Precisión (0,73), pero bajas F1 (0,35) y Recuperación (0,23).
- *DBPedia Spotlight-OpenTapioca* (en inglés): Precisión (0,84) considerablemente más alta que Recuperación (0,13) y F1 (0,23).
- *DBPedia Spotlight-OpenTapioca* (en español): Muestra una alta Precisión (0,82), pero Recuperación (0,09) y F1 (0,17) muy bajas.
- *TagMe-DBPedia Spotlight-OpenTapioca* (en inglés): Exhibe una Precisión (0,76), Recuperación (0,34) y F1 (0,47) comparable a otras configuraciones.

En resumen, este gráfico ofrece una visión comparativa del rendimiento de diferentes sistemas de *entity linking* y sus combinaciones en diversos idiomas, evaluados en términos de Precisión, Recuperación y F1 en el marco de *SemEval*. La métrica F1 se emplea para clasificar los sistemas, dado que representa el promedio armónico de las otras dos métricas. En consecuencia, el sistema con los mejores resultados es *TagMe* en ambos idiomas evaluados, seguido por la combinación *TagMe-DBPedia Spotlight-Opentapioca* y, finalmente, la unión *TagMe-DBPediaSpotlight*.

Es importante destacar que, en cuanto a Precisión, *TagMe* (en ambas configuraciones de idiomas) y *OpenTapioca* (en español) lograron rendimientos muy similares, siendo *TagMe* el que alcanzó la puntuación más alta entre ambos. Por otro lado, *OpenTapioca* (en inglés) exhibió la mejor precisión entre los tres sistemas, lo que indica que este sistema presenta entidades encontradas con mayor exactitud, es decir, cuenta con más entidades correctas en comparación con las incorrectas que los demás sistemas.

En cuanto a la recuperación, el sistema líder es *TagMe*, seguido por la combinación *TagMe-OpenTapioca-DBPedia Spotlight*, y en última posición se encuentra *OpenTapioca*. Esto implica que *TagMe* es el sistema que identifica la mayor cantidad de entidades esperadas, mientras que *OpenTapioca* se posiciona en último lugar en este aspecto.

6.2.2. Etiquetas Manuales

Con el propósito de validar el trabajo propuesto, buscamos responder a la pregunta: ¿la búsqueda de entidades en textos carentes de contexto con formato de tuit demuestra un rendimiento satisfactorio en términos de eficacia y precisión?. No se encontró ninguna herramienta de evaluación de etiquetados específica para textos en español de tipo tuit, es decir, con una longitud de hasta 280 caracteres. Por este motivo, se etiquetaron manualmente 100 tuits seleccionados al azar desde el conjunto de 20 millones disponibles en *TelarKG* al que tenemos acceso. Cada tuit seleccionado contiene al menos una entidad mencionada.

Un formato comúnmente utilizado para representar los resultados esperados en Procesamiento del Lenguaje Natural (PLN) se conoce como *Interchange Format* (*NIF*), el cual se basa en *Resource Description Framework* (*RDF*) y *Web Ontology Language* (*OWL*). Todas las entidades identificadas en el conjunto de 100 tuits etiquetados manualmente fueron registradas en este formato. Para realizarlo se empleó una herramienta llamada *NIFify_ v2*,

desarrollada por Henry Rosales (Doctor en Computación), que proporciona una interfaz gráfica para etiquetar uno o más documentos de manera organizada, eficiente y sin posibilidad de errores en el formato.

Una vez que las entidades fueron identificadas manualmente y registradas en un formato estándar, se procedió con la evaluación de las etiquetas proporcionadas por cada sistema. Para llevar a cabo esta evaluación de manera efectiva, es fundamental describir las etiquetas en el mismo formato. Con este fin, se utilizó la librería "*nifwrapper*", diseñada para transformar datos *NIF* en clases de *Python*, facilitando así el procesamiento eficiente de la información. Además dicha librería proporciona comandos de evaluación que nos permite tener métricas estándar (Precisión, Recuperación y F1) para los resultados obtenidos por los tres sistemas de *entity linking* (*TagMe*, *DBPedia Spotlight* y *OpenTapioca*).

Existen diferentes enfoques para calcular las métricas de precisión, recuperación y F1, conocidos como Micro Promedio y Macro Promedio. La distinción entre ambos radica en la manera de abordar los datos. El Micro Promedio calcula los valores de las métricas de manera individual para cada documento y luego los combina en un único valor global. Por otro lado, el Macro Promedio utiliza todos los datos para calcular directamente las métricas finales. La elección entre uno u otro depende de la naturaleza de los datos y de la importancia asignada a cada documento [3].

El Micro Promedio otorga más peso a los documentos con más instancias, mientras que el Macro Promedio asigna igual peso a todos, independientemente de su tamaño. En nuestro contexto, cuando nos referimos a documentos, estamos hablando de tuits, donde los documentos con más instancias serían aquellos tuits que contienen más entidades. Sin embargo, la cantidad de entidades por tuit es variable, ya que los tuits pueden ser tan cortos como una palabra o tan largos como aproximadamente 31 palabras en español.

En nuestra evaluación, la diferencia en la cantidad de entidades entre los tuits más cortos y más largos es significativa, con una variación máxima de 30. Dado que esta diferencia es lo suficientemente grande, se considera más apropiado utilizar el Macro Promedio. Esto garantiza que se le dé igual importancia a las entidades encontradas en tuits con menos menciones y en aquellos con más menciones, contribuyendo así a una evaluación más equitativa.

Para emplear el Macro Promedio, la única consideración a tener en cuenta es integrar todos los tuits en un único documento para su evaluación, en lugar de tratar cada tuit como un documento independiente.

En la Tabla 6.4 se presentan los resultados de rendimiento al etiquetar 100 tuits, utilizando la configuración en español de los tres sistemas de *entity linking* de manera individual propuestos en este trabajo. Es relevante señalar que en el proceso de etiquetado, se generó una petición por cada tuit, generando así un total de 100 peticiones por cada sistema de *entity linking*.

Como se puede observar en la Tabla 6.4 el sistema con el mejor resultado es *Opentapioca*, con una Precisión, Recuperación y puntaje F1 aproximados de 0,4.

Con el propósito de destacar de manera comparativa el rendimiento de los sistemas de *entity linking*, tanto entre ellos como entre sus diversas configuraciones de idioma, se presentan

Tabla 6.4: Resultados de Precisión, Recuperación y Puntuación F1 de las entidades identificadas por los sistemas de *entity linking* al realizar el etiquetado de tuits en español

	Precisión	**Recuperación**	**Puntaje F1**
TagMe	0,0041	0,0128	0,0063
DBPedia Spotlight	0,3151	0,1474	0,2009
OpenTapioca	0,4379	0,4295	0,4337

los resultados de las métricas (Precisión, Recuperación y F1) en el gráfico de la Figura 6.2, el cual se describe a continuación:

- Eje X (Horizontal): Cada punto en este eje en el gráfico representa uno o más sistemas de *entity linking* unidos en una configuración específica de idioma. Los sistemas incluidos son *TagMe* (en inglés señalado como *TM en*), *DBPedia Spotlight* (indicado como *DS en* para inglés y *DS es* para español)), y *OpenTapioca* (abreviado como *OT en* para inglés y *OT es* para español). Además, se consideraron todas las posibles combinaciones de uno, dos o tres de estos sistemas para diferentes idiomas, por ejemplo, *TagMe* con *DBPedia Spotlight* para inglés *(TM-DS en)*, *DBPedia Spotlight* con *OpenTapioca* para español *(DS-OT es)*, etc.
- Eje Y (Vertical): La escala en el eje y varía de 0 a 1. Indica el rendimiento de los sistemas en términos de Precisión, Recuperación y F1.
- Barras de Colores: Cada grupo de tres barras de colores diferentes representa la Precisión (en azul), Recuperación (en rojo) y F1 (en café) respectivamente para cada sistema e idioma evaluado.

Con respecto a los resultados expuestos en el gráfico de la Figura 6.2 se puede observar:

1. *TagMe* (en inglés): Exhibe muy bajos valores de Precisión Recuperación y F1. Todos con valores cercanos a 0,01.
2. *DBPedia Spotlight* (en inglés): Muestra una Precisión (0,2), Recuperación (0,1) y F1 (0,1) bajas, pero significativamente mayores que *TagMe*.
3. *DBPedia Spotlight* (en español): Similar a la configuración en inglés, con resultados un poco más altos.
4. *OpenTapioca* (en inglés y español): Presenta resultados consistentes de Precisión, Recuperación y F1, manteniendo niveles comparables entre ellos. Destaca como el sistema con los mejores resultados, ya que sus valores de Precisión, Recuperación y F1 para ambas configuraciones se sitúan entre 0,4 y 0,5.
5. *TagMe* y *DBPediaSpotlight* (en inglés): En general, se observaron resultados bajos, todos por debajo de 0,2. Sin embargo, todas las métricas superaron los resultados individuales de *TagMe*, destacando especialmente la Recuperación, que alcanzó un valor de 0,1.

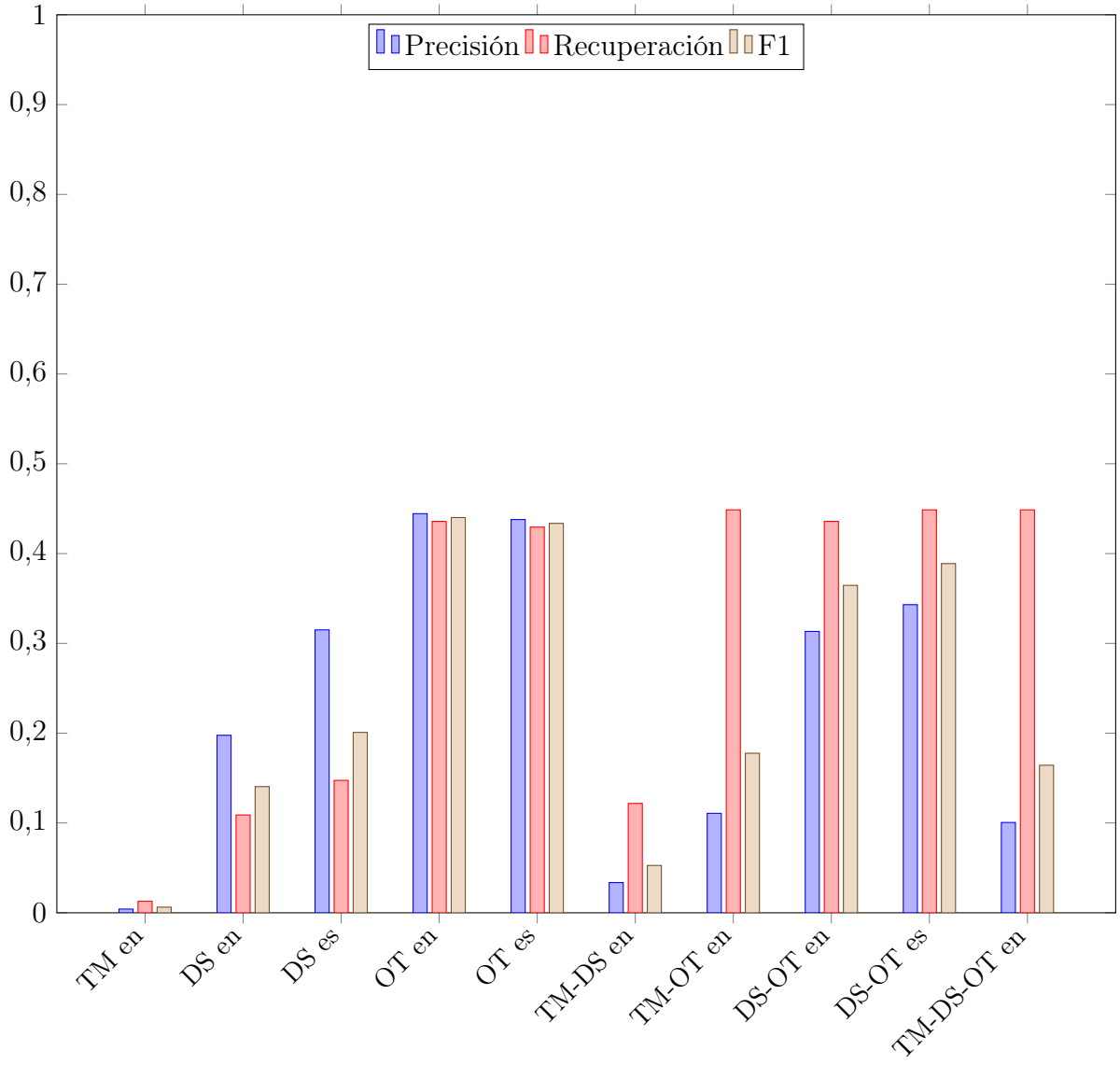

Figura 6.2: Desempeño de los sistemas en evaluación de tuits: Precisión, Recuperación y F1.

6. *TagMe* y *Opentapioca* (en inglés): La combinación de estos sistemas resultó en mejoras en la Precisión y F1 en comparación con los resultados de cada sistema por separado. A su vez, se registró una Recuperación más alta para ambos sistemas combinados, alcanzando un valor de 0,4.

7. *DBPedia Spotlight* y *Opentapioca* (en inglés y español): Los resultados de las tres métricas fueron ligeramente superiores en la configuración en español en comparación con la configuración en inglés. Tanto la Precisión como la F1 obtuvieron valores entre los resultados de los sistemas individualmente, pero la Recuperación fue notablemente superior, con un valor de 0,4.

8. *TagMe*, *DBPedia Spotlight* y *Opentapioca* (en inglés): Aunque se logró la Recuperación más alta del experimento (0,4) junto a otras configuraciones, la Precisión y F1 no superaron el umbral de 0,2.

En resumen, este gráfico proporciona una comparación visual del rendimiento de diversos sistemas de *entity linking*, así como de todas sus combinaciones, en la tarea de etiquetar tuits. Se evalúa su desempeño en términos de métricas clave en dos idiomas: inglés y español. En cuanto a los resultados de los sistemas evaluados individualmente, *OpenTapioca* tiene los valores más altos, para todas las métricas, en cualquier configuración de idioma; en segundo lugar *DBPedia Spotlight* y en tercero *TagMe*.

En cuanto a los resultados de las combinaciones de sistemas, la combinación de *DBPedia Spotlight* y *OpenTapioca* obtiene el mejor puntaje general (F1), siendo ligeramente mayor en español por aproximadamente 0,02 en comparación con la configuración en inglés. Le sigue en puntaje cercano la combinación de *TagMe* y *OpenTapioca*, seguida por *TagMe*, *DBPedia Spotlight* y *OpenTapioca* en conjunto, y finalmente, la puntuación más baja de este grupo corresponde a la combinación de *TagMe* junto con *DBPediaSpotlight*.

En cuanto a la Recuperación en particular, se alcanzó el mejor puntaje de todo el experimento (0,449) con la combinación de *DBPedia Spotlight* y *OpenTapioca* en español, así como con *TagMe* y *OpenTapioca* en inglés y la combinación de los tres sistemas juntos.

Este experimento presenta varias limitaciones. En primer lugar, la cantidad de tuits utilizados es reducida, lo que puede afectar la representatividad de los resultados. En segundo lugar, las etiquetas fueron asignadas por una persona, lo que introduce la posibilidad de no reconocer alguna entidad al leer el texto. Además, la tarea de *entity linking* depende del contexto, y al utilizar solo tuits, se prescinde de gran parte de este contexto, lo que podría considerarse un entorno desfavorable para evaluar de manera óptima los sistemas de *entity linking*.

A pesar de las limitaciones inherentes a este experimento, los resultados preliminares presentados proporcionan una primera comparación entre los tres sistemas en el entorno real del caso de uso específico. Es importante considerar que estos resultados, aunque son preliminares, ofrecen valiosas conclusiones iniciales que pueden orientar la elección del o los sistemas de *entity linking* más adecuados.

6.3. Resultados sobre Tuits Etiquetados

En el marco del caso de uso detallado en el Capítulo 5, se llevó a cabo el etiquetado de aproximadamente 20 millones de tuits con contenidos asociados al proceso constituyente en Chile durante el año 2021. Este proceso de etiquetado se realizó mediante *entity linking*, identificando entidades de *Wikidata* presentes en el texto de los tuits almacenados en *TelarKG*.

Con el objetivo de evaluar los resultados obtenidos y abordar la pregunta 6 planteada al inicio de este capítulo (Asumiendo que se pueden etiquetar los tuits, ¿los resultados de *entity linking* pueden habilitar análisis novedosos en el contexto de *TelarKG*?), se llevó a cabo un análisis de los datos en bruto y posteriormente se realizaron consultas sobre la versión extendida de *TelarKG*.

Los resultados obtenidos tras aplicar la técnica de *entity linking* a los tuits almacenados

en *TelarKG* se exponen detalladamente en la Sección 6.3.1. Estos resultados proporcionan información crucial, incluyendo el recuento de nodos y conexiones que se han incorporado a la base de datos *TelarKG*. Dichos nodos y enlaces se identifican con las etiquetas *Mention* y *Entity*, y las conexiones pueden clasificarse como *mention* y *entity*. Es importante resaltar que los tuits que contienen al menos una mención establecen una conexión de tipo *mention* con un nodo etiquetado como *Mention*. Estos nodos *Mention* a su vez están enlazados mediante una conexión de tipo *entity* a una entidad, que puede ser un nodo recién creado (con la etiqueta *Entity*) o uno ya existente en *TelarKG*, en caso de que la entidad mencionada represente a un convencional constituyente o a uno de los partidos políticos que cuenta con una entidad en *Wikidata*. Para obtener detalles más específicos sobre estos resultados, se sugiere revisar la Sección 6.3.1.

Además, se llevaron a cabo análisis de los resultados obtenidos al realizar consultas sobre la versión extendida de *TelarKG*. Esta aproximación posibilitó la incorporación tanto de las entidades originales de *TelarKG* como de las menciones agregadas en la versión ampliada. Con el propósito de alcanzar este objetivo, se plantearon cinco preguntas específicas, cuyas respuestas se buscaron a través de consultas en *QML* (*QuadModel Query Language*), un lenguaje diseñado para consultar datos en formato *QuadModel*. Se ofrece más información al respecto en la Sección 6.3.2.

6.3.1. Análisis y Métricas de Entidades Encontradas en *TelarKG*

Al llevar a cabo el etiquetado de los tuits, se procedió inicialmente con aquellos pertenecientes a la tabla *tw-convencionales* debido a su prioridad. Los resultados de este procesamiento, que incluyen el promedio de menciones encontradas por tuit, la cantidad de tuits con al menos una mención, el porcentaje de tuits con al menos una mención y la cantidad de entidades diferentes identificadas por los dos sistemas de *entity linking*, se detallan en la Tabla 6.5.

Tabla 6.5: Métricas de entidades en la tabla *tw-convencionales*

Tabla tw-convencionales	**DBPedia Spotlight**	**OpenTapioca**
Número de tuits	616.486	616.486
Promedio de menciones por tuit	0,26	0,12
Tuits con al menos una mención	96.518	36.653
Porcentaje de tuits con menciones	15,7 %	5,9 %
Entidades diferentes encontradas	11.002	2.979

Los resultados presentados en la Tabla 6.5 revelan que *DBPedia Spotlight* generó un mayor número de menciones, abarcando un mayor número de tuits en comparación con su contraparte. Además, este sistema de *entity linking* identificó aproximadamente cuatro veces más entidades únicas en los tuits de esta tabla que *OpenTapioca*.

La Tabla 6.6 presenta los mismos resultados que se muestran en la Tabla 6.5, pero en relación con la tabla de tuits *tw-streaming*.

Según los resultados presentados en la Tabla 6.6, es evidente que la cantidad de tuits eti-

Tabla 6.6: Métricas de entidades en la tabla *tw-streaming*

Tabla tw-streaming	DBPedia Spotlight	OpenTapioca
Número de tuits	19.545.968	18.444.637
Promedio de entidad por tuit	1,36	0,22
Tuits con al menos una mención	15.874.133	2.398.851
Porcentaje de tuits con menciones	81,2 %	13 %
Entidades diferentes encontradas	62.713	19.207

quetados por ambos sistemas de *entity linking* no coincide. Esto se atribuye a fallos técnicos durante el procesamiento de los tuits, específicamente a la pérdida de algunos paquetes enviados entre la máquina y la API. A pesar de esto, *DBpedia Spotlight* etiquetó todos los tuits de la tabla *tw-streaming*, mientras que los tuits perdidos representan un 5,6 % para *OpenTapioca*, lo que sugiere que este error no implica una pérdida significativa de información.

Al igual que en el procesamiento de los tuits en la tabla *tw-convencionales*, *DBpedia Spotlight* identificó más menciones y abarcó una mayor cantidad de tuits. En esta ocasión, *DBpedia Spotlight* encontró alrededor de 3,3 veces más entidades únicas que su contraparte.

En la Tabla 6.7, se detalla la cantidad de nodos adicionales incorporados a *TelarKG*. Estos nodos se añaden al incluir las menciones a entidades, nuevas o previamente existentes, que son identificadas a través de aplicar *entity linking* en todos los tuits de *TelarKG*. La versión expandida de *TelarKG* incluye nuevos nodos clasificados como *Mention* y *Entity*. Los nodos de tipo *Mention* están siempre conectados a un tuit y, además, tienen una segunda conexión que puede enlazarse a un nodo clasificado como *Entity*, *ConventionMember* o *Party*.

Tabla 6.7: Estadísticas globales de entidades creadas en *TelarKG*

Nodos Entity	71.596
Nodos Mention	29.311.087
menciones a Entity	29.081.567
Menciones a ConventionMember	117.680
Menciones a Party	111.840
ConventionMember mencionados	83
Party mencionados	15

Dado que cada nodo con la etiqueta *Mention* siempre cuenta con dos enlaces, uno clasificado como *mention* y otro como *entity*, la cantidad total de estos enlaces coincide con el número de nodos *Mention* creados, que asciende a 29.311.087. La creación de 71.596 nodos etiquetados como *Entity* se llevó a cabo, y se establecieron 29.081.567 enlaces entre estos nodos y aquellos que representan menciones.

Adicionalmente, se generaron 117.680 conexiones entre menciones y 83 nodos diferentes de convencionales constituyentes, de un conjunto total de 155. También se incorporaron 111.840 enlaces entre menciones y 15 nodos distintos clasificados como *Party*, de un conjunto total de 25.

En la Tabla 6.8, se presentan de manera descendente los miembros de la convención

constitucional que fueron más mencionados, considerando los 117.680 enlaces generados por *DBPedia Spotlight* y *OpenTapioca* en conjunto.

Tabla 6.8: Diez miembros de la convención constitucional más mencionados en los tuits de *TelarKG*

Nombre de Convencional Constituyente	**Número de menciones**
Jorge Baradit Morales	20.728
Marcela Cubillos	17.835
Fernando Atria Lemaitre	16.387
Benito Baranda Ferrán	9.128
Elisa Loncon Antileo	7.614
Mauricio Daza Carrasco	6.994
Patricia Politzer Kerekes	6.733
Malucha Pinto Solari	5.117
Cristina Dorador Ortiz	4.928
Marcos Barraza Gómez	4.094

Los datos expuestos en la Tabla 6.8 indican que, según la información recopilada, el convencional constituyente más mencionado es Jorge Baradit Morales, con un total de 20.728 menciones. Le sigue Marcela Cubillos, con 17.835 menciones, ocupando el segundo lugar. Fernando Atria Lemaitre se sitúa en la tercera posición, con un reconocimiento de su entidad de *Wikidata* en 16.387 tuits de *TelarKG*.

En la Tabla 6.9, se presenta un *ranking* de los diez partidos políticos cuyas entidades fueron identificadas con mayor frecuencia. Estas menciones a partidos políticos, que suman un total de 111.840, fueron registradas por los sistemas de *entity linking* al procesar los tuits de *TelarKG*.

Tabla 6.9: Diez partidos políticos más mencionados en los tuits de *TelarKG*

Partido Político	**Número de Menciones**
Partido Comunista de Chile	39.040
Partido Radical	32.325
Partido Socialista Chileno	14.620
Unión Demócrata Independiente	7.440
Renovación Nacional	5.447
Evolución Política	4.938
Democracia Cristiana de Chile	2.235
Partido Por la Democracia	1.890
Partido Republicano	1.266
Convergencia Social	1.128

Como se evidencia en la Tabla 6.9, el partido político más mencionado, según los datos derivados del proceso de *entity linking* aplicado a los tuits en *TelarKG*, fue el Partido Comunista de Chile, con un total de 39.000 menciones. En la segunda posición se sitúa el Partido

Radical, con 32.325 menciones. En la parte inferior del *ranking* se ubica Convergencia Social, con un total de 1.128 menciones.

La información proporcionada por los sistemas de *entity linking* seleccionados incluye, además de la entidad y la posición en el texto donde se menciona, el tipo de entidad en la mayoría de los casos. Como se muestra en la Tabla 6.7, se crearon 71.596 nodos de tipo *Entity*, cada uno etiquetado con la palabra clave *Entity* y todos los tipos a los que pertenecen. Es importante destacar que *OpenTapioca* solo proporciona un tipo por entidad (*Organization*, *Location* o *Person*), mientras que *DBPedia Spotlight* incluye otros más específicos. Los diez tipos de entidades más frecuentes encontradas en *TelarKG* se detallan en la Tabla 6.10.

Tabla 6.10: Los diez tipos de entidades más comunes hallados en los tuits de *TelarKG*

Label	**Número de Entidades**
Person	16.628
Agent	16.496
Location	13.874
Place	8.398
Work	8.315
Organization	7.541
PopulatedPlace	6.773
Region	6.385
AdministrativeRegion	6.385
Artist	3.634

En la Tabla 6.10, se observa que el tipo de entidad más frecuente en los tuits de *TelarKG* es *Person*, con un total de 16.628 menciones, seguido de cerca por *Agent*, con 16.496 entidades. Los demás tipos de entidades, en orden descendente según la cantidad de estas, son *Location*, *Place*, *Work*, *Organization*, *PopulatedPlace*, *Region*, *AdministrativeRegion* y *Artist*. Se destaca que la mayoría de estos tipos de entidades están estrechamente relacionados con los temas pertinentes a *TelarKG*, lo que sugiere que los resultados obtenidos reflejan el contenido de la información analizada. Aunque es importante tener en cuenta que los resultados no son perfectos, la identificación de estos tipos de entidades puede ser valiosa para comprender mejor el contenido y los temas discutidos en la plataforma de *TelarKG*.

6.3.2. Consultando a *TelarKG*

Con el objetivo de examinar las menciones de entidades en los tuits de *TelarKG*, los cuales fueron sometidos al proceso de *entity linking*, se llevaron a cabo consultas en formato *QML* (consulte el Capítulo 5.4) en la versión actualizada de *TelarKG*. Esta versión incluye todas las menciones a entidades de *Wikidata* identificadas en los tuits que ya estaban previamente registrados en *TelarKG*.

Tanto los datos de entrada de *TelarKG* como su versión extendida están en formato *QuadModel* (consulte la Sección 5.1). Para la creación e inicialización de la base de datos, se emplea *MillenniumDB*, un motor de base de datos.

Las consultas en *QML* realizadas sobre la versión extendida de *TelarKG* se detallan en la Sección 5.4 y se diseñaron con el propósito de poner a prueba o refutar el valor real de las menciones obtenidas mediante *entity linking*. Dado que *TelarKG* contiene información sobre el proceso constitucional en Chile del año 2021, todas las preguntas planteadas, en el contexto de evaluar la utilidad de las menciones, están relacionadas con partidos políticos y miembros de la convención constitucional de esa época.

Cada pregunta a responder está asociada a una consulta en *QML*, detallada en la Sección 5.4. Las preguntas seleccionadas se describen a continuación:

1. ¿Cuáles son los convencionales constituyentes que se mencionan a sí mismos con mayor frecuencia en los tuits? Consulta en *QML* en el *Listing* 5.6
2. ¿Cuál es la frecuencia de menciones entre convencionales de un mismo partido? Consulta en *QML* en el *Listing* 5.7
3. ¿Existen pares específicos de convencionales que se mencionan entre sí con mayor frecuencia? Consulta en *QML* en el *Listing* 5.8
4. ¿Cuáles son los partidos más mencionados por convencionales de otros partidos? Consulta en *QML* en el *Listing* 5.9
5. ¿Con qué frecuencia los convencionales de dos partidos políticos se mencionan mutuamente, revelando una relación destacada entre dichos partidos? Consulta en *QML* en el *Listing* 5.10

Las respuestas y detalles de las preguntas 1, 2, 3, 4 y 5 se presentan en las Secciones 6.3.2.1, 6.3.2.2, 6.3.2.3, 6.3.2.4 y 6.3.2.5, respectivamente. Además, es relevante destacar que las consultas en *QML* no ofrecen una funcionalidad directa para contar, por lo que los resultados obtenidos requieren de un procesamiento adicional para obtener las cifras.

6.3.2.1. ¿Cuáles son los convencionales constituyentes que se mencionan a sí mismos con mayor frecuencia en los tuits?

Aunque apenas el 0,4 % de las menciones identificadas en los tuits corresponden a miembros de la convención constitucional, estos datos son esenciales para analizar la efectividad real de las menciones detectadas. Para abordar la pregunta que guía esta sección, se llevó a cabo la consulta en *QML* detallada en el *Listing* 5.6. Esta consulta proporciona todos los tuits publicados por un convencional que tiene una mención dirigida a sí mismo, logrando esto al igualar los atributos *name* del convencional que publica el tuit y el mencionado en él.

Existiendo alrededor de 600.000 tuits publicados por convencionales constituyentes y 126 cuentas de *Twitter* asociadas a dichos convencionales en *TelarKG*, se identificaron un total de 719 menciones a sí mismos realizadas por 52 miembros de la convención constitucional. Los 10 convencionales constituyentes con más menciones a sí mismos, en orden descendente, se presentan en la Tabla 6.11.

Tabla 6.11: Los 10 convencionales constituyentes con más menciones a sí mismos en *TelarKG*

Convencional Constituyente	**Automenciones**
Marcela Cubillos	103
Cristina Dorador	89
Mauricio Daza	80
Marcos Barraza	50
Ignacio Achurra	45
Jorge Baradit	43
Fernando Atria	36
Malucha Pinto	34
Amaya Alvez	24
Elisa Loncón	23

La primera persona en el *ranking* es Marcela Cubillos, con 103 menciones a sí misma, seguida por Cristina Dorador, con 89 automenciones. En la décima posición se encuentra Elisa Loncón, con 23 menciones a sí misma identificadas.

6.3.2.2. ¿Cuál es la frecuencia de menciones entre convencionales de un mismo partido?

Dado que cada convencional constituyente está vinculado a un partido político dentro de *TelarKG*, es posible identificar el partido político del emisor de un tuit, siempre y cuando este sea un miembro de la convención constitucional. También es posible identificar el partido político de cualquier convencional constituyente mencionado en algún tuit. Para abordar la pregunta de esta sección, se llevó a cabo la consulta en *QML* detallada en el *Listing* 5.7. Esta consulta recupera todos los tuits publicados por un miembro de la convención constitucional que menciona a otro de su mismo partido político en el texto. Este resultado se logra mediante la igualación de los atributos *name* de los partidos políticos del emisor y del mencionado, al mismo tiempo que se garantiza que no se trate de la misma persona, mediante la imposición de que los atributos *name* sean diferentes para ambos convencionales constituyentes.

En última instancia, el total de tuits publicados por un miembro de la convención constitucional que menciona a otro miembro de la convención perteneciente al mismo partido político asciende a 205 menciones. La distribución de estas menciones, desglosada por partido político, se presenta detalladamente en la Tabla 6.12.

Como se exhibe en la Tabla 6.12, los candidatos independientes son los que más menciones realizan entre sí, con un total de 161 menciones. En segundo lugar, con significativamente menos menciones (15), se encuentran los convencionales constituyentes pertenecientes a los Pueblos Originarios. En la última posición se hallan Renovación Nacional y Evolución Política, con tan solo una mención registrada.

Tabla 6.12: Menciones a convencionales constituyentes hechas por miembros de su partido político

Partido Político	Menciones entre Convencionales del Mismo Partido
Independientes	161
Pueblos Originarios	15
Unión Demócrata Independiente	13
Partido Socialista	5
Revolución Democrática	5
Convergencia Social	2
Partido Comunista	2
Renovación Nacional	1
Evolución Política	1

6.3.2.3. ¿Existen pares específicos de convencionales que se mencionan entre sí con mayor frecuencia?

A través de la consulta en *QML* especificada en el *Listing* 5.7, es posible recuperar todos los tuits publicados por un miembro de la convención constitucional que contiene una mención a otro convencional constituyente. Esto se logra al imponer la diferencia en el atributo *name* de ambos convencionales constituyentes. En total, se identificaron 470 menciones realizadas por un convencional constituyente a otro. Los nueve pares de personas que más se mencionan entre sí se presentan en orden descendente, según la cantidad de menciones, en la Tabla 6.13.

Tabla 6.13: Relación de menciones entre convencionales constituyentes más frecuentes

Convencionales Constituyentes		Cantidad de Menciones Cruzadas
Cristina Dorador	Elisa Loncón	24
Ignacio Achurra	Cristina Dorador	14
Elisa Loncón	Teresa Marinovic	12
Ignacio Achurra	Malucha Pinto	7
Cristina Dorador	Malucha Pinto	6
Alondra Carrillo	Elisa Loncón	6
Fernando Atria	Bernardo Fontaine	6
Jorge Baradit	Marcela Cubillos	6
Amaya Alvez	Tomás Laibe	6

Los datos proporcionados en la Tabla 6.13 revelan que Cristina Dorador y Elisa Loncón son las convencionales constituyentes que más se han mencionado mutuamente, sumando un total de 24 menciones. En segundo lugar, nuevamente Cristina Dorador, pero esta vez con Ignacio Achurra, registran un total de 14 menciones.

Dada la escala limitada de los resultados de este experimento, se llevó a cabo una verificación manual para garantizar su precisión. Se comprobó si cada texto identificado como mención de un miembro de la convención constitucional correspondiera a una mención real de este. Durante la revisión, se encontraron solo 4 casos, de un total de 470, en los que las entidades mencionadas no estaban efectivamente presentes en el texto del tuit al que se hacía

referencia. Además, en todos los demás casos, se constató que el nombre y apellido del convencional eran mencionados en el tuit, asegurando así la conexión precisa con las entidades encontradas. Este experimento logró identificar con un 99,1 % de precisión un total de 470 menciones a convencionales constituyentes entre los tuits publicados por otros miembros de la convención constitucional.

6.3.2.4. ¿Cuáles son los partidos más mencionados por convencionales de otros partidos?

TelarKG dispone de información sobre los convencionales constituyentes, incluyendo la afiliación política a la que pertenecen, lo que permite abordar la pregunta planteada en esta sección. En la versión ampliada de *TelarKG*, se incorporaron los identificadores de *Wikidata* de los partidos políticos que tienen una entidad en *Wikidata*. Esta ampliación facilita la conexión entre los tuits publicados por los convencionales constituyentes y los partidos políticos mencionados en su contenido.

Para abordar esta pregunta, se llevó a cabo la consulta en *QML*, detallada en el *Listing* 5.9, en la versión extendida de *TelarKG*. La consulta devuelve los tuits publicados por un convencional constituyente que menciona un partido político diferente al que pertenece. Esto se logra al exigir que sean diferentes los atributos *name* del partido político mencionado y del partido político del convencional constituyente que publica el tuit.

En total, se identificaron 1.067 menciones a partidos políticos realizadas por convencionales constituyentes que pertenecen a una agrupación política diferente. En la Tabla 6.14, se presentan los 10 partidos políticos más mencionados por convencionales constituyentes que no pertenecen a ellos, ordenados por la cantidad de menciones de manera descendente.

Tabla 6.14: Menciones a partidos políticos por convencionales de agrupaciones diferentes.

Partidos Políticos	Cantidad de Menciones Realizadas por Convencionales de Otros Partidos
Partido Radical	360
Partido Comunista	326
Partido Socialista	140
Evolución Política	72
Partido Por la Democracia	64
Renovación Nacional	35
Convergencia Social	24
Democracia Cristiana	20
Partido Federación Regionalista Verde Social	8
Partido Republicano	7

Como se evidencia en la Tabla 6.14, el partido político más mencionado por convencionales constituyentes pertenecientes a agrupaciones diferentes fue el Partido Radical, con un total de 360 menciones. A continuación, el Partido Comunista ocupa la segunda posición con 320 menciones. En la última posición de los 10 partidos políticos más mencionados se encuentra el Partido Republicano, con un total de 7 menciones.

6.3.2.5. ¿Con qué frecuencia los convencionales de dos partidos políticos se mencionan mutuamente, revelando una relación destacada entre dichos partidos?

El objetivo de esta sección es investigar las relaciones entre partidos políticos a través de las menciones presentes en los tuits de *TelarKG*. Para ello, se utiliza la consulta del *Listing* 5.10 en la versión extendida de *TelarKG*. Esta consulta recupera todos los tuits publicados por un convencional constituyente que menciona a otro convencional constituyente perteneciente a una agrupación política diferente. La condición esencial para obtener estos resultados es la discrepancia entre los atributos *name* de los partidos políticos del emisor del tuit y del convencional constituyente mencionado. Es relevante destacar que no es necesario verificar que el convencional constituyente que emite el tuit sea distinto al mencionado, ya que, al pertenecer a diferentes agrupaciones políticas, se descarta la posibilidad de que sean la misma persona.

En total, se identificaron 1.077 menciones realizadas entre convencionales constituyentes pertenecientes a partidos políticos diferentes. La Tabla 6.15 presenta las 10 primeras posiciones de pares de agrupaciones políticas con la mayor cantidad de menciones cruzadas. Cabe destacar que en la última posición existen dos pares de partidos políticos con la misma cantidad de menciones. La tabla está ordenada por la cantidad de menciones en orden descendente.

Tabla 6.15: Pares de partidos políticos con mayor cantidad de menciones cruzadas entre sus convencionales constituyentes

Pares de Partidos Políticos		**Cantidad de Menciones Cruzadas**
Independientes	Unión Demócrata Independiente	291
Partido Socialista	Unión Demócrata Independiente	132
Independientes	Revolución Democrática	131
Independientes	Pueblos Originarios	88
Convergencia Social	Independientes	68
Independientes	Renovación Nacional	52
Partido Comunista	Independientes	29
Partido Socialista	Unión Demócrata Independiente	20
Independientes	Partido Progresista	18
Independientes	Partido Liberal	15
Pueblos Originarios	Renovación Nacional	15

Los resultados presentados en la Tabla 6.15 revelan que la pareja de agrupaciones políticas con la mayor cantidad de menciones cruzadas son los Independientes y La Unión Demócrata Independiente, con un total de 291 menciones. En segundo lugar, se encuentra nuevamente La Unión Demócrata Independiente, pero esta vez con el Partido Socialista, con un total de 132 menciones. En la última posición, con 15 menciones cruzadas, se destacan los pares Independientes con el Partido Liberal y Pueblos Originarios con Renovación Nacional.

6.3.2.6. Comentarios Finales

Dados los resultados expuestos en esta sección, es posible afirmar que la aplicación de *entity linking* a los tuits de *TelarKG* permite realizar análisis novedosos. Por ejemplo, se identificaron 1.077 tuits (detallados en la Sección 6.3.2.5) que evidencian conversaciones entre miembros constitucionales de partidos políticos e ideologías diferentes, como es el caso del Partido Socialista con Unión Demócrata Independiente. Asimismo, se encontraron 1.067 tuits realizados por miembros de la convención constitucional que mencionan un partido al que no pertenecen (más información en la Sección 6.3.2.4). Estos ejemplos proporcionan la reducción significativa de la cantidad de tuits que deben examinarse manualmente, para realizar un estudio sobre la conversación.

Aún siendo una muestra pequeña de los resultados obtenidos con *entity linking*, se observó que el 99,1 % de las menciones identificadas en la Sección 6.3.2.3 estaban correctamente asignadas, que es un aumento significativo con respecto al etiquetado general; en este caso, solo se consideran menciones a convencionales constituyentes indicados en *TelarKG*, que reduce significativamente los falsos positivos. En las instancias correctamente identificadas, aunque se mencionara al convencional por nombre y apellido, no era necesario que estuviera su nombre completo para ser encontrado. Además, al tener estas menciones dentro de *TelarKG*, es posible aplicar condiciones sobre quién publica el tuit al realizar búsquedas, lo que representa una ventaja respecto al buscador implementado en *Twitter*.

6.4. Discusiones Generales

Los resultados obtenidos en los experimentos de *SemEval* y de etiquetas manuales revelan discrepancias significativas. En *SemEval*, *TagMe* destaca como el mejor sistema en términos de F1 (rendimiento general) y *OpenTapioca* en la última posición. En contraste, con el etiquetado manual, *OpenTapioca* emerge como el sistema líder en todas las métricas y *TagMe* obtiene el peor resultado.

Aunque estos resultados pueden parecer contradictorios, se pueden explicar considerando la influencia del tamaño del texto en la tarea de *entity linking*. El contexto de estos, representado por todas las palabras en el texto, desempeña un papel crucial en la capacidad de los sistemas para seleccionar entidades. Los resultados podrían indicar que *OpenTapioca* logra un buen desempeño incluso con menos contexto, a diferencia de *TagMe*, que muestra mejores resultados cuando se dispone de un contexto extenso.

Al evaluar el rendimiento en cuanto a las métricas de recuperación, precisión y F1, se toma una decisión sobre qué sistema o sistemas utilizar para el etiquetado de todos los tuits. Esto es fundamental para el proyecto, ya que estos sistemas serán responsables de enlazar los tuits con entidades de *Wikidata*, facilitando consultas avanzadas. Dado que *DBPedia Spotlight* demuestra resultados consistentes en ambos experimentos de manera individual, y se posiciona favorablemente cuando se combina con otros sistemas, ya sea con *OpenTapioca*, *TagMe*, o ambos simultáneamente, se concluye que es la elección preferida como sistema de *entity linking*.

Considerando el experimento de *SemEval*, se observa que *OpenTapioca* exhibe buena precisión pero menor recuperación, indicando que no encuentra muchas de las entidades esperadas, pero las que identifica suelen ser correctas. En el experimento manual, *OpenTapioca* muestra valores cercanos en recuperación y precisión. En este contexto, es preferible tener menos entidades pero con alta precisión, evitando así proporcionar entidades incorrectas a los tuits etiquetados. Aunque se puedan perder algunas entidades, esto puede mejorarse en futuros trabajos, asegurando resultados de mejor calidad. Por lo tanto, se prefiere *OpenTapioca* sobre *TagMe*, ya que muestra los mejores resultados en el experimento de etiquetado de tuits y un desempeño suficientemente bueno en *SemEval* para ser seleccionado. Dado el bajo rendimiento tanto de *TagMe* de forma individual como en conjunto con otros sistemas en el experimento de tuits, se concluye que este sistema requiere más contexto para lograr respuestas satisfactorias. Entonces, aunque haya sido el sistema con mejores resultados en *SemEval*, no se elige para etiquetar los 20 millones de tuits.

En relación con el experimento de etiquetado manual descrito en la Sección 6.2.2 de este capítulo y para optimizar la precisión de los resultados en un experimento de esta índole, se proponen algunas sugerencias. En primer lugar, se sugiere aumentar significativamente la cantidad de tuits etiquetados, fomentando la participación de un mayor número de personas en la anotación de datos. Esta estrategia tiene el potencial de amplificar la identificación de entidades esperadas y obtener resultados más precisos sobre el desempeño de los sistemas de *entity linking* puestos a prueba.

Al analizar la calidad de los enlaces efectuados, según se detalla en la Sección 6.13, se evidencia que al restringir los resultados a los convencionales constituyentes, la Precisión aumenta significativamente al 99,1 %. Este incremento es consecuencia de la reducción de falsos positivos que apuntan hacia entidades no relevantes para *TelarKG*. Sin embargo, solo se logran identificar menciones de 83 convencionales de un total de 155; asumiendo que cada convencional es mencionado al menos una vez en los datos, esto indica una Recuperación a nivel de convencionales de aproximadamente un 54 %. Los resultados presentados en la Tabla 6.8 sugieren que se logra una identificación más precisa de convencionales con nombres menos comunes, y así, menos ambiguos. Finalmente, se propone realizar una evaluación manual de las menciones que apuntan a un nodo de tipo *Party* en *TelarKG*, con el fin de analizar más a fondo el rendimiento de los sistemas de *entity linking* al etiquetar entidades relevantes al caso de estudio.

Capítulo 7

Conclusión

Millones de individuos comparten publicaciones a diario en plataformas de redes sociales, como *Twitter*, abordando una amplia variedad de temas; desde lugares y personas hasta empresas y mucho más. Por otro lado, existen repositorios robustos, como *Wikidata*, que albergan millones de entidades representando conceptos simples o complejos que van desde números hasta teorías filosóficas. Este proyecto surge con la premisa de unir estos dos universos, identificando las entidades de *Wikidata* que se mencionan en conversaciones en redes sociales. El propósito es clasificar de manera automática el contenido de las publicaciones en redes sociales, facilitando así su análisis posterior, volviéndolo más eficiente y menos costoso. Actualmente, la revisión manual de cada publicación es necesaria para comprender sobre qué trata o a qué hace referencia. Este proyecto busca aportar en la superación de esa limitación.

En relación al proceso constituyente en Chile del año 2021, el Instituto Milenio Fundamento de los Datos lanzó el proyecto *TelarKG*, una base de conocimientos que almacena información pertinente a este evento histórico. Dentro de los datos almacenados se encuentran alrededor de 20 millones de tuits relacionados al tema y publicados tanto por personas comunes, como por convencionales constituyentes de la época. Utilizando este caso de uso se propone una solución para el enriquecimiento de los tuits de *TelarKG* mediante el enlazamiento a entidades de *Wikidata*.

Se evaluaron tres sistemas de *entity linking*, eligiendo *DBPedia Spotlight* y *OpenTapioca* tras analizar su desempeño. En el procesamiento, se etiquetaron todos los tuits con estas herramientas. La unión de los resultados de ambos sistemas de *entity linkin* suman un total de 29.311.087 menciones; estas hacen referencia a 71.596 entidades nuevas y 98 entidades pertenecientes a los datos originales de *TelarKG* (15 partidos políticos y 83 convencionales constituyentes). Los resultados se transformaron al formato de datos en bruto *TelarKG*, conocido como *QuadModel*. Se utilizó el motor de bases de datos *MillenniumDB* para crear una base de datos que contuviera la nueva información asociada a la ya existente en *TelarKG*. Finalmente, se realizaron consultas avanzadas sobre las nuevas entidades de la base de datos para vislumbrar cualquier relación entre ellas con respecto a la temática común de los tuits.

Al implementar la versión extendida de *TelarKG*, la cual permite consultas especializadas utilizando el lenguaje *QML*, se logró el objetivo general de este proyecto: "Implementar una estructura de datos que enlace tuits archivados y recientes de *Twitter* con *Wikidata*, permi-

tiendo ejecutar consultas más avanzadas sobre los tuits que las que permite la plataforma de *Twitter* actualmente. Esperamos que el trabajo a realizar facilite búsquedas en una muestra grande de contenido de *Twitter* para quién desee estudiar tuits sobre algún tema particular".

Con respecto al primer objetivo específico, que consiste en "Encontrar relaciones entre entidades mencionadas en los tuits y sus autores, con aquellas entidades existentes en *Wikidata*", los resultados obtenidos en la Sección 6.3.1 respaldan la consecución de este objetivo. Se logró identificar entidades correspondientes a los miembros de la convención constitucional del caso de uso en un total de 117.680 ocasiones (sumando los resultados de ambos sistemas de *entity linking* seleccionados). Además, se detectaron 11.840 menciones a partidos políticos relevantes para el proceso y período en Chile.

Se logró alcanzar el objetivo específico 2: "Generar una estructura de datos que almacene los tuits y sus entidades relacionadas, permitiendo ejecutar consultas avanzadas sobre sus datos". Esto al escribir los resultados obtenidos al aplicar *entity linking* a los tuits de *TelarKG* en formato *QuadModel*. Esto se hizo para extender la base de datos con nueva información, creando así una estructura de datos que facilita consultas especializadas, tal como se detalla en la Sección 6.3. A través de los experimentos mencionados en la Sección 6.3, se estableció una relación entre las entidades mencionadas en los tuits y algunos de sus autores, cuando estos eran miembros de la convención constitucional. Se identificaron 205 menciones realizadas por miembros de la convención constitucional a otros colegas de la misma agrupación política. Asimismo, se registraron 470 menciones entre miembros de la convención constitucional. También se documentaron 1.067 menciones efectuadas por convencionales constituyentes a partidos políticos diferentes al suyo, y 1.077 menciones entre convencionales constituyentes de diferentes partidos.

Con respecto al objetivo específico 3 "Realizar diversas consultas sobre una estructura de datos creada a partir de una muestra de todos los datos que se poseen y corroborar manualmente los resultados obtenidos", al momento de llevar a cabo el trabajo se encontró uno de los desafíos fundamentales en este proyecto: Determinar qué tuits, dentro del vasto conjunto de 20 millones, eran verdaderamente representativos. La complejidad radica en que cualquier muestra de datos verificable manualmente resultaba demasiado reducida para arrojar conclusiones que se pudieran generalizar.

Dada esta limitación, se realizó un ajuste en el objetivo para que fuera alcanzable con los limitados recursos que se poseen y, al mismo tiempo, proporcionara información valiosa sobre la calidad de las menciones encontradas. La modificación incluye que en vez de ejecutar varias consultas a una versión reducida de la estructura de datos, se realiza una consulta relevante al caso de estudio a la base de datos completa. Además, se etiquetan manualmente una muestra de 100 tuits, para comprender la calidad de los enlaces que se pueden generar con los sistemas de *entity linking* seleccionados. Este ajuste se llevó a cabo mediante dos experimentos clave: uno que implicaba la revisión manual de etiquetas, siendo alcanzable al tratarse de 100 tuits, y otro que aborda la pregunta crucial: "¿Existen pares específicos de convencionales que se mencionan entre sí con mayor frecuencia?".

Estos experimentos fueron estratégicos, ya que el primero proporcionaba una muestra manejable para una evaluación manual detallada, y el segundo ofrecía perspectivas valiosas sobre las menciones específicas entre convencionales. Esta modificación no solo hizo que el

objetivo fuera más accesible, sino que también generó una comprensión más profunda y significativa de la calidad de las menciones dentro del contexto del caso de uso seleccionado.

Las limitaciones significativas identificadas durante la realización de este trabajo incluyen; la vasta cantidad de datos que requerían etiquetado y análisis posterior. También se encuentran las restricciones propias de las APIs de los sistemas de *entity linking*, incluyendo la restricción en longitud del texto que puede ser etiquetado y la existencia de un límite en la cantidad de consultas que se pueden enviar de manera simultánea. Además, el formato del texto a etiquetar es una limitación importante para los sistemas de *entity linking*, ya que los tuits, textos cortos y propensos a abreviaciones, contienen un contexto limitado, lo que dificulta la correcta identificación de entidades, colocando a las herramientas de *entity linking* en un escenario desafiante.

La evaluación de la calidad de las menciones encontradas en los tuits propone dos limitaciones. La primera es que, para evaluar un sistema de *entity linking*, se deben encontrar las entidades mencionadas de manera manual. Además, debido a limitaciones temporales, fue imposible etiquetar un gran conjunto de datos, logrando esta tarea solo con 100 tuits. Asimismo, este tipo de trabajo obtiene mejores resultados si se realiza en grupo, ya que es más probable encontrar una mayor cantidad de entidades al contar con más de una persona etiquetando los mismos datos, además de poder validar que hay un acuerdo entre las personas sobre las entidades presentes.

Para mejorar y ampliar este trabajo, se sugiere evaluar la posibilidad de etiquetar los datos utilizando el sistema de *entity linking REL* [23]. También se sugiere analizar los resultados de aplicar *entity linking* sobre textos de tipo tuit con una muestra, etiquetada manualmente, más extensa. Además, podría explorarse la identificación de las entidades más mencionadas y verificar si existe alguna entidad que no corresponda con el texto al que se le asocia y eliminar todas las menciones a esta mención, lo que podría eliminar menciones a palabras comunes del español (“les”, “como”, “mas”, entre otras). Asimismo, la creación de un diccionario de abreviaturas comunes y su sustitución por las palabras completas podría proporcionar datos de entrada más sólidos a los sistemas de *entity linking*.

Resultaría sumamente beneficioso para investigaciones futuras relacionadas con tuits contar con un sistema de *entity linking* diseñado específicamente para identificar entidades en situaciones de escaso contexto, pero que comparten un tema común. Este es el caso de los tuits analizados, pues estos no abordan temas aleatorios, ya que fueron recopilados con la intención de ser relevantes en la discusión sobre la creación de una nueva constitución en Chile.

Se plantea la posibilidad de definir una entidad que abarque el contexto y que sirva como referencia para las entidades identificadas por el sistema, exigiendo cierta cercanía o similitud con dicha entidad. La creación e implementación de este sistema se vislumbra como un valioso trabajo futuro, con beneficios no solo para el proyecto actual, sino también para estudios en el ámbito de las redes sociales en general.

Bibliografía

[1] Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. Multimodal entity linking for tweets. In *Lecture Notes in Computer Science*, pages 463–478. Springer International Publishing, 2020.

[2] Marcelo Arenas and Pablo Barceló. Chile's new interdisciplinary institute for foundational research on data. *Commun. ACM*, 63(11):78–83, 2020.

[3] Hercules Dalianis. *Evaluation Metrics and Evaluation*, pages 45–53. Springer International Publishing, 2018.

[4] Antonin Delpeuch. Opentapioca: Lightweight entity linking for wikidata. *CoRR*, abs/1904.09131, 2019.

[5] Antonin Delpeuch. OpenTapioca: Lightweight Entity Linking for Wikidata. In Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic, editors, *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[6] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49, March 2015.

[7] Yue Feng, Fattane Zarrinkalam, Ebrahim Bagheri, Hossein Fani, and Feras Al-Obeidat. Entity linking of tweets based on dominant entity candidates. *Social Network Analysis and Mining*, 8(1), June 2018.

[8] Paolo Ferragina and Ugo Scaiella. Fast and Accurate Annotation of Short Texts with Wikipedia Pages. *IEEE Softw.*, 29(1):70–75, 2012.

[9] Joseph Gibbons, Robert Malouf, Brian Spitzberg, Lourdes Martinez, Bruce Appleyard, Caroline Thompson, Atsushi Nara, and Ming-Hsiang Tsou. *Twitter-based measures of neighborhood sentiment as predictors of residential population health*, Jul 2019.

[10] Bo Han, Paul Cook, and Timothy Baldwin. Text-based twitter user geolocation prediction. *J. Artif. Intell. Res.*, 49:451–500, 2014.

[11] Manoj Prabhakar Kannan Ravi, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Jens Lehmann. Cholan: A modular approach for neural entity linking on wikipedia and wikidata. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021.

[12] Ioannis Karatzas. and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, Berlin, 2nd edition, 2000.

[13] Marcus Klang and Pierre Nugues. Hedwig: A named entity linker. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4501–4508. European Language Resources Association, 2020.

[14] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. End-to-end neural entity linking. *arXiv preprint arXiv:1808.07699*, 2018.

[15] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015.

[16] Ying Lin. *10 Twitter Statistics Every Marketer Should Know in 2020 [Infographic]*, Oct 2020.

[17] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia Spotlight: shedding light on the web of documents. In Chiara Ghidini, Axel-Cyrille Ngonga Ngomo, Stefanie N. Lindstaedt, and Tassilo Pellegrini, editors, *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*, ACM International Conference Proceeding Series, pages 1–8. ACM, 2011.

[18] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bring order to the web. Technical report, Technical report, stanford University, 1998.

[19] Philip Protter. *Stochastic Integration and Differential Equations*. Springer, 1990.

[20] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*. Number 293 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin [u.a.], 3. ed edition, 1999.

[21] Ibrahim Sabuncu, Mehmet Ali Balci, and Ömer Akgüller. Prediction of usa november 2020 election results using multifactor twitter data analysis method. *CoRR*, abs/2010.15938, 2020.

[22] Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570, 2022.

[23] Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. Rel: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20. ACM, 2020.

[24] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

[25] Domagoj Vrgoč, Carlos Rojas, Renzo Angles, Marcelo Arenas, Diego Arroyuelo, Carlos Buil-Aranda, Aidan Hogan, Gonzalo Navarro, Cristian Riveros, and Juan Romero. MillenniumDB: An Open-Source Graph Database System. *Data Intelligence*, pages 1–39, 06 2023.

[26] Jia Xue, Junxiang Chen, Ran Hu, Chen Chen, Chengda Zheng, Yue Su, and Tingshao Zhu. Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. *Journal of Medical Internet Research*, 22(11):e20550, November 2020.

[27] Miguel Ángel Zúñiga González. Análisis del diálogo sobre políticos latinoamericanos en twitter, 2020.