Identificación automática del asunto de opiniones en texto en idioma español.

Informe de Proyecto de Grado

Yasim Zeballos

Tutores:
Aiala Rosá
Juan José Prada

Facultad de Ingeniería
Universidad de la República

2013

«La gran rueda de la historia raras veces se detiene, hay que luchar y vencer, ser yunque o ser martillo; por mucho tiempo hemos sido yunque, es hora de que nos transformemos en martillos para forjar nuestro destino»

Manuel Sadosky (1914 – 2005),
precursor de la computación argentina y uruguaya.

Resumen

Este documento presenta el diseño e implementación de un sistema basado en reglas para detectar automáticamente el tema del que trata una opinión expresada en un texto en español. El sistema se basa fuertemente en nociones establecidas en [Rosá2011]. Toma también información producida por el sistema del mismo trabajo, por lo que este sistema puede enmarcarse como extensión o componente de un proyecto más general.

Para el sustento teórico-lingüístico, se tienen en cuenta nociones morfológicas, sintácticas, así como ciertas características que poseen las opiniones, el discurso y el lenguaje en general.

La solución busca construirse utilizando reglas, no utilizándose enfoques estadísticos o de aprendizaje automático. En cuanto a los resultados logrados utilizando medidas típicas del área, en general se alcanzan valores superiores a otros trabajos.

Se generaron nuevos recursos lingüísticos, extendiendo el corpus anotado manualmente del trabajo en que nos basamos, mediante la anotación (también manual) de los asuntos de las 305 opiniones del corpus.

Palabras clave: minería de opiniones, sistemas de reglas, procesamiento de lenguaje natural, anotación de corpus, identificación de tema

Agradecimientos

A Aiala Rosá y Juan José Prada por las numerosas consideraciones a lo largo de todo el proyecto.

A Ismael Garrido, por la revisión preliminar y sus comentarios sobre el trabajo.

Índice

1 Introducción

Visto la gran cantidad de información que se encuentra en Internet, ya sea en foros de discusión, blogs, sitios especializados o portales de noticias, resulta de interés poder obtener información más específica de un texto, en principio solo estructurado en la forma que le dio el autor.

Existen numerosos trabajos que se enfocan en la identificación de las fuentes de las diversas opiniones que se transmiten en las noticias, así como del sentimiento u orientación semántica (positiva, negativa, neutra) que se expresa sobre un determinado tema, incluso algunos en el cambio de polaridad en función del tiempo [MLWS2007]. Identificar esta información permitiría predecir o averiguar las preferencias de las personas, obteniéndose importante información política, social, económica o de mercado [Liu2010][PL2008][SW2009][VC2012].

Un trabajo que brinda un buen contexto es [PL2008], donde explica el boom de las investigaciones en el análisis de subjetividad. Allí se sugiere que tanto la *minería de opiniones* como el *análisis de sentimientos* sean consideradas lo mismo. Estas áreas a su vez pertenecen al área más general Procesamiento de Lenguaje Natural (PLN).

En nuestro horizonte, se busca poder determinar el tema o asunto del que se está hablando en un determinado contexto. Se pueden sacar conclusiones analizando todo el contexto o se pueden tener niveles de granularidad más finos. Siguiendo la segunda opción en un texto se pueden delimitar las opiniones que se transmiten en él. En particular en este proyecto nos enfocamos en obtener el tema de cada una de las opiniones de una noticia, analizando su contenido. El resultado final es el tema de una opinión particular sin tener en cuenta el contexto de la opinión u otras opiniones de la misma noticia.

Un sistema que integre todo lo anterior nos permitiría buscar opiniones en portales de noticias (u otra fuente de información) por un determinado asunto, obteniendo su fuente y su orientación semántica, potencialmente extendiendo la capacidad de las herramientas de búsqueda de propósito general o sistemas de respuesta automática. Por ejemplo pudiendo responder a las preguntas: ¿Quiénes hablaron sobre el tema X?, ¿Cuáles son las opiniones sobre el tema X?, ¿Qué opina X sobre Y?.

Este proyecto tomará en cuenta definiciones de otros trabajos y fundamentalmente se "construye sobre" las nociones expuestas en [Rosá2011], aunque no tendremos en cuenta la orientación semántica del texto, que generalmente se utiliza para averiguar si alguien está a favor o en contra de algo (también llamado polaridad).

En PLN, existen dos grandes enfoques típicos para aproximarse a la solución, uno es mediante reglas -el enfoque que seguimos- y el otro es mediante métodos estadísticos (donde se enmarca el a*prendizaje automático*). El enfoque de reglas pone el énfasis en el conocimiento profundo del lenguaje, y poder expresar ese conocimiento con reglas específicas para automatizar el reconocimiento. Los métodos estadísticos no buscan representar el conocimiento de forma directa, sino elaborar relaciones o patrones que se puedan aprender mediante grandes juegos de datos. De alguna manera un enfoque es más teórico y otro más pragmático/empirista, aunque esto es discutible y no hay por qué verlos como técnicas opuestas, sino que son métodos complementarios, y ambos pueden ser usados para hacer un sistema híbrido como en [Rosá2011]

1.1 Motivaciones

Además de lo señalado anteriormente en la introducción, existe una carencia de herramientas para el español en el área de PLN, por lo que este proyecto contribuye al área de análisis de opiniones para nuestro idioma.

Como la búsqueda de nuevos horizontes está siempre latente [Mins2010], también es una pequeña contribución al área de la inteligencia artificial: el entendimiento del asunto de un determinado fragmento de texto avanza en la dirección de que una máquina pueda "entender" de qué se está hablando, si bien, este proyecto -tanto por las herramientas que usa como por los métodos usados- se encuentra en una etapa bastante rudimentaria de esta ambiciosa meta.

Una meta más terrenal es la de mejorar la interacción persona-computadora, al incrementar el potencial de búsqueda en texto libre.

1.2 Introducción a conceptos fundamentales

El presente proyecto tiene unos pocos conceptos centrales que forman parte de las estructuras básicas sobre las que se hace todo el trabajo, que se pueden ver en el recuadro de ejemplos elementales 1. En el recuadro se encuentran tres fragmentos de noticias distintas, donde se aprecian tres opiniones; a su vez cada opinión tiene una fuente y uno o más "temas centrales" o asuntos sobre los que versa la opinión de la fuente:

A) *El presidente de la Cámara de Diputados, Roque Arregui*, que presidió el acto de homenaje, dijo que **la obra de Ibarbourou** "no conoció límites de edad ni de intereses".

B) *El estudio*, que se basa en un análisis de los datos disponibles al cabo de ensayos comparativos de inhibidores de la neuraminidasa (enzima presente en el virus de la gripe) en los niños, subraya que **el Tamiflu** puede causar vómitos en algunos niños y que puede provocarles deshidratación y complicaciones.

C) Consultado **sobre la utilización de la palabra "secuestro" al ser detenido por inteligencia**, *Bruno* explicó que "existen determinados procedimientos que están establecidos por ley y que ahora afortunadamente han sido establecidos en la ley de establecimiento policial, pero que ya formaban parte de los costumbre y los usos. Todo un ritual procesal que en nuestro caso no se cumplieron".

Cuadro – Ejemplos elementales 1

Notando que las fuentes están en cursiva, mientras que los verbos que nos dan la pauta que se trata de una opinión subrayados y los asuntos en negrita, se puede ver que en el fragmento A la fuente es "el presidente de la Cámara de Diputados", y el asunto o tema del que habla, es "la obra de Ibarbourou". El texto se entiende que es una opinión por la aparición de una forma del verbo de comunicación *decir*.

Parecido es el fragmento B, donde la fuente es “El estudio” donde el verbo *subrayar* nos da el contenido de la opinión a través de la subordinada introducida por *que.*

El fragmento C tiene información adicional, el asunto o tema de la misma, se encuentra de manera explícita en el mensaje, precedido por la preposición *sobre.* Pero ¿qué ocurre cuando el asunto no está explícito en el texto?. Podríamos pensar en delimitarlo dentro del contenido de la opinión. Para delimitar el asunto, aplicaremos diversas reglas y heurísticas a este contenido, y a este contenido le llamaremos *mensaje* de la opinión, que desarrollaremos en la siguiente subsección.

1.3 Formalizando los elementos de la opinión

La noción de opinión de este proyecto sigue a [Rosá2011] que se inspira en ([BYTH2004], [WWC2005], [Liu2010]). La opinión y sus elementos estructurales más importantes se definen en el siguiente recuadro:

Opinión: segmento de texto que transmite las expresiones o posturas de alguna fuente sobre algún asunto.
Fuente: autor u origen de la opinión, que generalmente es una persona u organización.
Asunto: tema sobre el que se opina.
Mensaje: contenido de la opinión.
Predicado de opinión: elemento indicador de la presencia de una opinión.

Es claro que no se llega a una definición matemática, por lo que puede resultar insuficiente el nivel de precisión de las definiciones, hecho que refleja la dificultad intrínseca de la tarea que nos proponemos, particularmente en las definiciones de asunto y mensaje.

También salta a la vista cierta circularidad en la definición del asunto, pues si vamos al diccionario de la Real Academia Española, buscando la definición de tema nos encontraremos con:

tema(definición): Proposición o texto que se toma por **asunto** o materia de un discurso.

Se señala que la fuente es el *origen* de una opinión y se ve en los siguientes ejemplos:

A) Luego de que la policía recibiera *un llamado* que indicaba que **dos hombres intentaban ingresar por la fuerza** a una vivienda en Cambridge, Boston (noreste).

B) Según *este relevamiento*, **las emisiones, la calidad de aire y del agua** arrojan parámetros buenos y muy buenos.

C) *Chávez* alerta sobre **posibilidad de guerra**.

Cuadro – Ejemplos elementales 2

donde por ejemplo “un llamado” es la fuente de la opinión, más allá de que el llamado es realizado por una persona, o que los relevamientos también son hechos por personas.

El asunto puede estar mencionado de forma explícita como en el fragmento C del cuadro de ejemplos elementales 1, o en el ejemplo C del cuadro de ejemplos elementales 2, que incluso no posee mensaje. Sin embargo, el caso más común es que esté implícito por lo que debe ser deducido a partir del texto del mensaje, que es lo que nos proponemos resolver en este proyecto.

El mensaje de una opinión es el contenido o la información que la fuente expresa sobre determinado asunto, resaltado con fondo gris en el recuadro anterior. En el primer ejemplo se da un caso típico, en el que el mensaje de la opinión es expresado en una oración subordinada introducida por “que”. Reiteramos que en el tercer ejemplo no hay mensaje en la opinión, pues está compuesta solo por la fuente, el predicado de opinión y el asunto[1].

Finalmente un predicado de opinión permite introducir una opinión, por ejemplo:

[juzga, subraya, estima, explicó, según, condena, ...]

Como ya se señaló y se señala en [Rosá2011] algunos predicados son verbos de comunicación (decir, opinar, etc.), otros verbos que expresan subjetividad (apoyan, rechazan, agradan, etc.,), nombres deverbales (condena, opinión[2], etc.); también son predicados de opinión elementos específicos de atribución de autoría como “según”.

1.4 Objetivos del proyecto

Este proyecto se centra en una parte específica del análisis de opiniones. Tomando como insumo opiniones con sus elementos estructurales ya identificados se propone:

1. Crear un sistema basado en reglas para identificar el asunto de una opinión en textos periodísticos en español usando métodos simbólicos.
2. Contribuir a un sistema más general de identificación de opiniones ya existente del grupo de PLN del Instituto de Computación de la Facultad de Ingeniería.
3. Generar recursos lingüísticos en español para el área de PLN, mediante la anotación de asuntos en un *corpus*.

1.5 Estructura de la documentación

La documentación se estructura de la siguiente manera: en el capitulo 2 se profundiza la descripción general, haciéndose énfasis en la relación que hay con otros trabajos y se comenta los elementos que se tomaron en cuenta.

En el capítulo 3 se comienza un análisis de la problemática, centrándose en la definición conceptual de asunto, así como en su identificación concreta, mencionando casos especiales.

1 En lo que respecta al asunto y al mensaje, se presentan nociones de mayor profundidad en [Renk2004]

2 No sabemos si es estrictamente un nombre deverbal (que deriva de un verbo)

El capítulo 4 presenta la solución de manera general, así como sus detalles. Menciona las herramientas usadas a todo nivel. También se comentan las dificultades y escollos encontrados al implementar la solución.

El capítulo 5 presenta las medidas típicas de evaluación de resultados del área (precisión, recuperación), y se aplican a este trabajo. Se muestran cuadros comparativos con otros trabajos, haciendo comparaciones cuantitativas.

El capítulo 6 tiene las conclusiones y síntesis de todo el trabajo.

Es pertinente tener presente algunos conceptos del glosario que se encuentra al final del trabajo, pues se usan sin explicar conceptos que son familiares para personas vinculadas al área, pero no necesariamente lo son para el resto, salvo para quienes hayan profundizado en el estudio de la lengua española.

2 Visión global del proyecto

Complementando la introducción este capítulo da una visión global del proyecto como proceso de trabajo señalando los hitos más importantes. También se verán los trabajos relacionados de forma de introducir algunos elementos del estado del arte y ver cuáles fueron tenidos en cuenta (o no) y por qué.

2.1 Narración del proceso de realización del proyecto

El proyecto comenzó con la anotación manual de asuntos en los mensajes[3] del corpus del trabajo en que nos basamos. La opinión ya estaba anotada en un conjunto de noticias, mientras que el conjunto de opiniones ya tenía identificada la fuente y el mensaje, además del asunto en situaciones especiales como la que comentamos, por lo tanto se anotó en cada mensaje el asunto de la opinión teniendo en cuenta su contenido. Esta parte del proceso no fue sencilla y resultó trabajosa, puesto que fue necesario un estudio previo de todas las opiniones (cerca de 300) así como varias iteraciones en el corpus, donde nuevas anotaciones ayudaban a encontrar patrones y a redefinir asuntos ya anotados. Por ejemplo, en un principio se incluyeron muchos verbos en los asuntos anotados de donde resultó imposible abstraer un patrón común, por lo que se optó por procurar que los verbos no formen parte del asunto salvo que fuese necesario.

Una vez terminada la anotación completa del corpus, se hizo un extenso relevamiento de la literatura vinculada al tema, buscando las técnicas más usuales respecto a temáticas similares, así como una profundización en nociones básicas de gramática, como ser la sintaxis, en particular sobre sintagmas, categorías y funciones gramaticales.

En la literatura se encontraron algunos trabajos con el objetivo de detectar el asunto de una opinión, pero ninguno para el español. Uno de los artículos que resultó más útil tenía como idioma base el chino donde curiosamente es útil aplicar al español criterios análogos [Lu2010]. La diversidad de enfoques resultó tan grande, que se leyeron decenas de artículos, todos los cuales incorporaban nueva información y prácticamente una metodología nueva, por lo que no se llegó a una saturación por reiteración de información, sino que se cesó la incorporación de nueva información para no prolongarla en exceso.

En un principio se emplearon técnicas sencillas que tomaron como base algunos esquemas mencionados en la literatura, que en este trabajo se le dio el nombre de *agrupamiento*, posteriormente usado en la definición de la línea base. También se desarrolló la técnica de *lista negra,* que incorporan en menor medida nociones semánticas al proyecto.

3 Los mensajes se pueden ver en el anexo B. Como para identificar el asunto no se tiene en cuenta información de contexto el lector puede intentar determinar el asunto de los ejemplos planteados para tener una noción tanto de lo que se quiere hacer, como de su dificultad.

Aunque se contaba con información morfosintáctica obtenida del sistema desarrollado en [Rosá11], resultó insuficiente para obtener buenos resultados, por lo que se buscó obtener más información sintáctica. Para ello se relevaron herramientas que permitieran un análisis de texto libre en español. Se identificaron FreeLing[4][PS2012] y VISL[5] como posibles candidatos, resultando muy esclarecedor el trabajo de Nevena Tinkova, A State-of-the-Art Review on Automatic Parsing of Spanish [Tink2007] en lo que respecta a las capacidades de estas herramientas.

Para la identificación automática del asunto se utilizó la información contenida en los árboles de dependencias[6] obtenidos mediante FreeLing combinándose con la información morfosintáctica, además de algoritmos sencillos que operan con los *tokens*. Los árboles de dependencias reflejan, mediante árboles concretos, concepciones de la teoría lingüística de Tesnière, que postula que los elementos de una oración se pueden estructurar como dependientes entre sí bajo distintos tipos de relaciones [Tink2007].

Finalmente se aplicaron evaluaciones típicas del área, como ser la precisión, la recuperación y su media armónica, la medida F, comparando con resultados de otros y marcando las diferencias de los elementos sobre los cuales se basó cada trabajo.

Las conclusiones finales no solo son una síntesis del resultado del trabajo, sino que algunas se alcanzaron sobre la marcha, y tuvieron vigencia durante todo el proceso.

2.2 Trabajos relacionados

Es de destacar que, de la profusa literatura consultada, pocos elementos en concreto se pudieron extraer en lo que respecta a las metodologías, aunque sí aportaron en el bagaje conceptual. La gran amplitud de los temas vinculados a la minería de opinión se ve reflejada en la *survey* de [PL2008] que cuenta con más de 300 referencias. Es grande la diversidad de soluciones pues en general cada autor tiene su propio método y además la mayoría de los métodos no siguen el enfoque de reglas. Lo anterior conduce a que las alternativas vistas bien podrían servir de complemento para mejorar los resultados obtenidos. Por ejemplo, en [Rosá11] se presenta un enfoque por reglas y otro por aprendizaje, además en ese trabajo se hace la valoración de que el reconocimiento de un 24,5% de asuntos es demasiado bajo, siendo esa carencia, parte de lo que originó este trabajo.

Al comienzo de nuestro proyecto se obtuvieron conceptualizaciones importantes de [Llor2009], trabajo que se centra en resumir textos completos. Este trabajo aportó en tener conciencia de los distintos fenómenos lingüísticos que se dan en la práctica, pudiendo dimensionar algunas dificultades y entender las limitaciones de la restricción adoptada de tomar únicamente el contenido del mensaje para determinar el asunto de la opinión.

4 FreeLing, proyecto de la Universidad Politécnica de Catalunya (ver http://nlp.lsi.upc.edu/freeling)
5 Visual Interactive Syntax Learning, proyecto de una universidad de Dinamarca (ver http://beta.visl.sdu.dk/visl/es)
6 Ver la sección “Dependencias de FreeLing” del capítulo 4 para un mayor detalle de árboles de dependencias.

Algunos trabajos plantean posturas fuertes como [SC2008b] en el que se indica a través de un ejemplo que algunas veces es imposible determinar el asunto sin conocer el contexto, o que el asunto es fuertemente dependiente del contexto. En nuestro proyecto se parte de la premisa de que siempre hay un asunto.

El trabajo de [KKM2008], si bien explota características sintácticas y de dependencias, no se utilizó por tener un enfoque de aprendizaje automático, por el mismo motivo tampoco [KH2006] que utiliza el etiquetado de roles semánticos para identificar el asunto. Nuestro trabajo tiene en cuenta elementos semánticos en menor medida, como se detalla en secciones posteriores.

Como se destacó anteriormente, el trabajo de Bin Lu[Lu2010] fue muy útil a los efectos de este proyecto por inspirar los enfoques principales, en conjunto con la discusión de ideas con los tutores de este proyecto, en base a lo hecho en [RWM2010] [Rosá11].

En particular, Lu plantea una sucesión de reglas sencillas que tienen en cuenta la función sintáctica de las palabras para identificar el asunto. Si la primera falla, se sigue con la segunda, y así sucesivamente en unos pocos pasos, pues no hay demasiadas reglas. Esta idea cuyos detalles se pueden ver en el anexo D fue el núcleo inicial de nuestro proyecto. Además, después se agregaron otros pasos que incorporan la información de árboles de dependencia.

Lu aplica al asunto la expansión de candidato, que toma el asunto más grande posible, tomando todos los “modificadores” que se le apliquen al "asunto base". Esta noción se incorporó tempranamente en nuestro proyecto al anotar manualmente un asunto tendiendo a cierta extensión.

La información obtenida de una opinión puede ser usada en sistemas más generales. Por ejemplo la investigación mostrada en [SW2009] se centra en algo complejo como son las expresiones argumentativas para dilucidar posturas en debates ideológicos online. En este artículo se señala que se utiliza el asunto como insumo para una algoritmia más general, observándose que incorporar el asunto da “mejor” información que usar solo las palabras (sistemas basados en *unigramas* por ejemplo).

Se mencionan usos de la identificación de asuntos en [Gimp2006], entre los que se destaca:

- Recuperación de documentos: dada una consulta usualmente se quiere obtener documentos ordenados por orden de relevancia. Los asuntos pueden ser usados como un componente más para determinar la relevancia de un documento.
- Traducción automática: Utilizar el asunto como parte del contexto para orientar la traducción.

En los recursos usados, no se encontraron juegos de datos adecuados, por lo que se elaboraron nuevos para este trabajo.

Es de destacar que en el extenso relevamiento que hace [PL2008], solo se menciona que en “NTCIR-7 Multilingual opinion analysis task” (un evento de competencia con tareas vinculadas al lenguaje y recuperación de información) del año 2008, hay una tarea

vinculada a la obtención de asunto. El resto de los corpus y juegos de datos mencionados, no parecen tener nada vinculado al asunto.

No existen hasta donde sabemos otros trabajos que se centren en identificar componentes de la opinión para el español, más allá del trabajo en el que nos basamos.

Como información complementaria de trabajos relacionados, se puede ver el Anexo D, donde se hace un análisis de varios artículos en lo que respecta a la metodología que usaron, algunos ejemplos y los resultados que alcanzan.

3 Análisis del problema

Uno de los puntos cruciales del proyecto es lograr una buena definición de asunto, tal que haga sencilla y clara su anotación manual y automática. En esta sección se verá que es una tarea compleja.

Se entendió pertinente hacer énfasis tanto en el proceso que se siguió como en las dificultades que se encontraron para establecer criterios para anotar y reconocer asuntos. Se presentan ejemplos de distinta dificultad, y se reproducen sintéticamente las reflexiones que suscitaron.

Como se señaló en la introducción, los trabajos en los que nos basamos identifican cuatro elementos claves en una opinión: la fuente, el asunto, el predicado de opinión y el mensaje. Sobre la fuente, el predicado de opinión y el mensaje, se vio que ya eran buenas tanto las definiciones como los resultados alcanzados por los trabajos en que nos basamos [Rosá11], por lo que solo se profundizó en el estudio del asunto.

Recordar que existen casos particulares (un 25% del corpus aproximadamente) donde ya se reconoció el asunto mediante el sistema de [Rosá11] pues en general lo introduce la preposición *sobre* o similares *(en lo que respecta a*, etc). Por ejemplo observar el asunto reconocido en las siguientes opiniones:

Opinion 12.8[7]

<opinion>
<fuente>La ministra de Salud Pública</fuente>
realizó algunas
<predicado>precisiones</predicado>
sobre **<asunto>la información que recorrió este jueves la prensa local e internacional</asunto>**
</opinion>

Opinión 4.8

<opinion>
sobre **<asunto>los grupos de presión que pudieron haber precipitado su procesamiento</asunto>**,
<predicado>dijo</predicado>
<mensaje>que "hubo de todo. Grupos de la mafia, grupos de presión político, económico, empresarial, gremial y policial . Fue un juego con muchas puntas"</mensaje>
</opinion>

Es decir, que en general el sistema intentará encontrar el asunto para el resto de los casos donde el asunto no se introduce “trivialmente” con una palabra particular.

Cuando se trata un problema complejo, una estrategia posible es buscar un encuadre o patrón para los casos más sencillos, y luego ir alterando la solución para que incluya los

7 El número de opinión refleja cierta nomenclatura del corpus de trabajo, en particular 12.8 significa “opinión 8 de la noticia 12”, se utiliza esta numeración por no crear una nueva numeración, aprovechando que cada opinión ya tiene un “identificador único”.

casos más complejos. En esta sección se seguirá esta estrategia, que se verá deja algunos cabos sin atar, pues como se ejemplifica en [Renk2004] para otra área de la lingüística como el caso de la teoría de actos del habla[8], a veces incluir todos los casos resulta demasiado difícil y la teoría sencillamente no los contempla, o quedan como contradicciones a la espera de ser superadas.

3.1 La definición del asunto, un problema en sí mismo

Como este proyecto se basa en texto ya procesado donde las opiniones, fuentes y mensajes ya están delimitados, resulta de mayor importancia la definición de en qué consiste un asunto y el estudio de los resultados de una anotación humana de esa definición puesta en práctica. Por ejemplo Elena Lloret en *Topic Detection and Segmentation in Automatic Text Summarization [Llor2009]* define asunto como "lo que trata una unidad de discurso[9]" además señala las distintas ópticas desde las que se puede abordar la estructura de un texto:

> *(traducción) Si un texto es considerado como un todo usualmente trata sobre un solo asunto, aunque en un análisis más profundo, varios subtemas pueden ser identificados, dando información adicional sobre el tema principal. Por otra parte, teniendo en cuenta la estructura de la oración, podemos encontrar que cada oración tiene un asunto.*

Compartiendo especialmente lo último, se trabajará con la hipótesis de que todo mensaje tiene el asunto contenido en él.

Otra forma de pensar la definición de asunto, como la forma de determinarlo es en función de para qué puede ser usado: para búsqueda de noticias, para investigación, para uso en tiempo real en debates políticos, o de manera universal sin tener en cuenta un uso específico.

3.2 Construyendo la noción de asunto como un proceso dialéctico

Se invita al lector a un esfuerzo en tratar de seguir la explicación como el proceso dialéctico que fue en la elaboración del proyecto, así como en la propia explicación de los ejemplos. Al no contarse con una base teórica sólida de la lengua española, deben utilizarse aproximaciones e intuiciones, frente a elementos o "formas lingüísticas" que seguramente ya estén definidas, pero escapan al alcance del proyecto y a la capacidad de elaboración. También será pertinente prestar atención a los ejemplos como forma de entender a qué se refieren las distintas preguntas y planteos que se hacen en el proceso.

Comenzamos con casos sencillos de analizar: los ejemplos van a estar en un recuadro indicando la numeración interna del corpus de opiniones, con un número indicando la

8 Una teoría de la pragmática formulada originalmente por J. Austin, dentro de la filosofía del lenguaje.
9 "The aboutness of a unit of discourse"

noticia (se trabajó con unas 38 noticias) y otro el número de opinión dentro de la noticia. El corpus de opiniones inicialmente solo tenía anotados los asuntos que se mencionaban de manera especial como vimos en secciones anteriores, por lo que se etiquetó manualmente el asunto de todas las opiniones que no lo tenían.

El primer caso que vemos es un ejemplo con un mensaje corto:

Opinion 1.3

La ministra destacó la prevención y
<opinion>
<predicado>reiteró</predicado>
<mensaje>que **<asunto>la pandemia</asunto>** es "leve a moderada".</mensaje>
</opinion>

donde es claro que el asunto es “la pandemia” por ser el sujeto principal del que se habla. Extender el asunto sobre lo que *es* la pandemia o qué características tiene, resulta excesivo para este ejemplo en concreto, pues el asunto termina siendo casi lo mismo que el mensaje, además en general tampoco se extendió el asunto cuando ya hay un verbo (y el verbo *ser* incluido) de por medio.

El siguiente es un ejemplo con un mensaje más largo:

Opinion 3.5

<opinion>
<fuente>El director del Maciel</fuente>
<predicado>explicó</predicado>
<mensaje>que hubo **<asunto>un acuerdo bilateral</asunto>** en hacer esa devolución en horas de trabajo, "porque como nosotros comenzamos a abrir más servicios y más salas, empezamos a requerir más horas de limpieza. Preferimos que nos devuelvan con más horas de trabajo"</mensaje>
</opinion>

Notar que en esta opinión el asunto corresponde a la pregunta de ¿qué ocurrió/qué hubo? y el asunto es un grupo nominal.

Otro ejemplo sencillo, que tiene un grupo nominal al comienzo:

Opinión 1.8 (simplificada)

Muñoz precisó que las cifras manejadas por Basso se refieren a "los puestos centinelas del MSP" en distintas instituciones médicas y
<opinion>
<predicado>recalcó</predicado>
<mensaje>que "no quiere decir que no haya" casos de gripe común, porque, recordó,
"**<asunto>la gripe</asunto>** no es una enfermedad de denuncia obligatoria, ni la común ni la H1N1"</mensaje>
</opinion>

Pasando a generalizar un poco más, si buscásemos la pregunta que nos obligaría a responder con el asunto, esta sería: ¿de qué habla la fuente? y más específicamente si quisiéramos tener orientaciones de la forma que tenemos que responder, teniendo en cuenta las preguntas de ¿Qué, Quién, Cómo, Cuándo, Dónde?, nos estaríamos enfocando en el “quién/qué”, y prácticamente dejando de lado el “cuándo” y el “dónde”. Es

decir que ponemos énfasis en los grupos nominales, y restamos atención a los tiempos o lugares. Por ejemplo en:

Opinión 37.1

<opinion>
<mensaje>**<asunto>La industria manufacturera</asunto> atenuó <asunto>la caída de su producción</asunto>** en volumen físico durante el sexto mes del año</mensaje>,
según
<fuente>datos del INE</fuente>
</opinion>

es claro que “sexto mes del año”(vinculado a “¿Cuándo?”) no es el asunto del mensaje, ni que en:

Opinión 29.2 modificada

<opinion>Según <predicado>indicó</predicado>
un comunicado emitido por
<fuente>UTE</fuente>
<mensaje>**<asunto>una falla en el suministro eléctrico</asunto>** dejó sin abastecimiento a la zona Oeste de Montevideo y a zonas adyacentes de los departamentos de San José y Canelones</mensaje>
</opinion>

los lugares formen parte del asunto (vinculados a “¿Dónde?”) .

Opinion 2.3 (simplificada)

<opinion>Según
<fuente>Helen Stancey, psicóloga del citado centro</fuente>,
<mensaje>"**<asunto>el cerebro</asunto>** procesa de modo distinto la información visual del espacio próximo y lejano que le llega por distintas vías "</mensaje>
</opinion>

El caso anterior se enfoca en responder “Quién” es “el sujeto” del asunto. Notar que en caso de que se busque responder las preguntas de ¿Qué hace? o ¿Cómo lo hace? debería de marcarse un asunto más completo usualmente incluyendo un verbo, en el sentido de decir que el asunto de la opinión anterior es “el cerebro procesa de modo distinto la información visual” y no simplemente “el cerebro” que responde la pregunta de ¿Qué/quién?. Este enfoque de priorizar el ¿Qué/quién? frente a ¿Cuándo/Dónde? para el caso de que el dominio de las opiniones[10] se encuentre fuertemente basados en fechas o lugares, tiene como principal desventaja que no es bueno para los casos en que que el asunto está vinculado a lugares o al tiempo. Por ejemplo en:

Opinión 33.2

<opinion>
<fuente>El presidente de ANCAP, Raúl Sendic</fuente>

10 Problemas por dominios específico también se presentan en [VAG2013]

,<predicado>aseguró</predicado>
<mensaje>que a partir de 2010 **<asunto>Uruguay</asunto>** ahorrará US$70 millones por concepto de importación de petróleo
</mensaje>
</opinion>

el asunto se entendió que es Uruguay, que es un lugar. Si bien justo es un caso discutible, donde puede considerarse “importación de petróleo” como asunto.

La opinión:

Opinión 13.6

<opinion>
<mensaje>"<asunto>La Eurozona</asunto> está en recesión, con fuertes señales de mejora que todavía deben transformarse en recuperación"</mensaje>,
<predicado>explicó</predicado>
<fuente>el FMI</fuente>
</opinion>

es un caso similar y menos ambiguo en lo que respecta al asunto.

En:

Opinion 6.2

<opinion>
<fuente>
Obama, el primer presidente negro del país , sumó leña a la controversia al
</fuente>
<predicado>declarar</predicado>
<mensaje>que **<asunto>la policía</asunto>** había actuado "estúpidamente" por arrestar a su amigo, lo que elevó la polémica a nivel nacional</mensaje>
</opinion>

se marcó como asunto únicamente el sujeto sin incorporar el verbo. Nuevamente en caso de querer extender el asunto, debería haberse marcado como “la policía había actuado ‘estúpidamente’”. Y en general veremos que la acción no es contemplada dentro del asunto, lo cual ayuda a la determinación del asunto, porque los mensajes en el corpus suelen tener al comienzo un sujeto -o más bien un grupo nominal- seguidos por un verbo.

Opinión 12.14

Esta semana,
<opinion>
<fuente>el portavoz de la comisión parlamentaria para los temas de los prisioneros, Kazem Jalali,</fuente>
<predicado>anunció</predicado>
<mensaje>que aun **<asunto>250 personas permanecen detenidas</asunto>**, entre ellas 50 personalidades políticas</mensaje>
</opinion>

En el recuadro anterior se ve un un caso donde el asunto parece necesariamente tener

que incluir un verbo, sobretodo si se piensa en un posible caso de uso, que salvo una búsqueda muy específica sobre “250 personas”, su estado (permanecer detenidas) es esencial al asunto. Siendo la pregunta asociada no solo vinculada a ¿Quién/quienes?, sino a ¿en qué estado están?.

Para esta opinión:

Opinion 11.1

<opinion>
<fuente>Carlos Julio Pereyra</fuente>
<predicado>confirmó</predicado>
a Montevideo Portal
<mensaje>que "no hubo **<asunto>un pronunciamiento oficial</asunto>"** para dejar en **<asunto>la libertad de acción</asunto>** a los votantes del PN ante el plebiscito por la anulación de la ley de caducidad</mensaje>
</opinion>

es discutible si lo esencial al asunto es resaltar la negación de que no hubo un pronunciamiento oficial, o basta destacar que se está hablando de “un pronunciamiento oficial”. Entendimos que “el plebiscito por la anulación de la ley de caducidad” es información circunstancial, por lo que no es lo suficientemente relevante como para ser tomada como asunto del mensaje.

3.2.1 Casos problemáticos

En esa subsección se explican algunos casos complejos y se aclara cuando no entren dentro del alcance del proyecto.

3.2.1.1 Asunto complejo

Opinion 12.8

Pero
<opinion>
<mensaje>"no fue autorizado a recitar los versos del Corán que se dicen en estas ocasiones e inmediatamente fue rodeado por la policía antidisturbios que lo llevó hasta su coche"</mensaje>,
<predicado>indicó</predicado>
<fuente>un testigo</fuente>
</opinion>

¿Cuál es el asunto del mensaje? ¿La no autorización a recitar versos, lo ejecutado por la policía o ninguna de las dos cosas?. Mirando la noticia en su contexto:

Noticia Nº 12 del corpus, contexto de la opinión.

IRÁN . **<asunto 1>Choque entre policías y manifestantes en Teherán</asunto 1>**. El enfrentamiento se produjo en el cementerio de la ciudad donde decenas de personas rendían homenaje a las víctimas de las protestas que estallaron tras la reelección de Ahjmadinejad.

(...)

Los testigos indicaron que **<asunto 2>Musavi</asunto 2>** logró salir de su coche y emprender el camino hacia la tumba de Neda Agha Soltan, la joven que murió baleada el 20 de junio pasado, convirtiéndose en símbolo de las protestas contra el resultado de las elecciones. <opinión>Pero "**<asunto 3>no fue autorizado <objeto directo>a recitar los versos del Corán que se dicen en estas ocasiones</objeto directo></asunto 3>** e inmediatamente fue rodeado por la policía antidisturbios que lo llevó hasta su coche", indicó un testigo. </opinión>
Mehdi Karubi pudo quedarse en el lugar y recogerse sobre las tumbas.

Pueden aventurarse tres asuntos si se tiene en cuenta el contexto de la opinión:

- El que está en el resumen de la opinión.
- El sujeto implícito del mensaje.
- El que forma parte del mensaje.

Viendo las posibilidades, si se determina el asunto como la primera frase que aparece en el resumen al comienzo del texto, se obtiene *Choque entre policías y manifestantes en Teherán*, aunque también puede pensarse que *no fue autorizado a recitar los versos del Corán...* es el asunto, siendo el complejo caso que una *negación* y una *conjugación del verbo ser* estarían perteneciendo al asunto. La última posibilidad es tomar el sujeto implícito de la autorización que se menciona en el texto y determinar que se está hablando de *Musavi*. Este camino nos llevaría por el lado de entender la progresión temática de un discurso así como la resolución de menciones implícitas.

Otro asunto complejo se presenta en:

Opinión 29.3

<opinion>
<mensaje>**<asunto>Un desperfecto en una estación de transformación</asunto>** ocasionó **<asunto>un importante corte de energía</asunto>**
</mensaje>
</opinion>

Cuyo contenido tiene la forma *causa-efecto*, donde es discutible si no corresponde marcar solo alguna de las dos cosas como asunto, o incluso todo como un único asunto.

3.2.1.2 Asuntos "implícitos" por correferencias

Opinion 11.2
(contexto de la noticia: la fuente es Carlos Julio Pereyra)

<opinion>
<mensaje>" Yo mismo cuando voté en contra
<opinion>
<predicado>reclamé</predicado>
<asunto>libertad de acción</asunto>
</opinion>"
</mensaje>
,<predicado>recordó</predicado>
</opinion>

Opinion 11.2 simplificada, asuntos potenciales resaltados en negrita

<opinion>
<mensaje>"**Yo mismo** cuando voté en contra **reclamé libertad de acción**"</mensaje>
,<predicado>recordó</predicado>
</opinion>

En el caso anterior (opinión simplificada) donde existen dos potenciales asuntos, hay uno especialmente complejo, por contener una correferencia. Tanto “reclamé libertad de acción” como “yo mismo” pueden ser los asuntos. Si se consideraran las correferencias se podría querer que el asunto fuese Carlos Julio Pereyra. No se incluye la resolución de correferencias para determinar el asunto en este proyecto pues es un problema complejo dentro del área y escapa al alcance de este proyecto, ver [ARZ2010] para resolución de correferencias.

3.2.1.3 Indirecciones

Opinión 11.4 simplificada

<opinion>
<fuente>El presidente del Honorable Directorio del Partido Nacional</fuente>
<predicado>dijo</predicado>
<mensaje>que **<asunto>coincide con <asunto>la opinión de varios correligionarios </asunto> </asunto>**, entre ellos Luis Alberto Lacalle
</mensaje>
</opinion>

Es polémico si *coincidir* es esencial al asunto en el sentido de que el asunto no es “la opinión de varios correligionarios” sino lo central es la coincidencia. Este problema ocurre porque o bien se piensa que el asunto es “la opinión de varios correligionarios” o bien algo que se aplica sobre la misma, que es la coincidencia, esto es, hay dos niveles de análisis, uno sobre el “asunto concreto” y otro que estaría “por encima” de el “asunto concreto”, pudiendo considerarse también como asunto.

3.2.1.4 Inclusión de subordinadas

Opinión 27.2

<opinion>
<mensaje>**<asunto>Esta práctica de derivar a los hogares del INAU a aquellos jóvenes que aún siendo mayores de edad deben cumplir con su condena anterior</asunto>** es frecuente
</mensaje>
, así lo
<predicado>señalaron</predicado>
<fuente>varios funcionarios del sindicato del INAU</fuente>
</opinion>

A veces no corresponde incluir en el asunto complementos con la preposición “de”. En el caso anterior la alternativa es elegir un asunto que sintetice el mensaje, o entender que casi todo el mensaje es el asunto.

Debido a la complejidad intrínseca del lenguaje, frente a este casos (y muchos otros) no se pudo establecer un criterio claro de como un humano puede determinar el asunto. En principio resultaría demasiado pobre dejar solo como asunto “Esta práctica” por ser un concepto muy general, por lo que se optó porque el asunto sea casi todo el mensaje, siendo una situación poco feliz. Como complemento, piénsese el problema de cuando se le pregunta a alguien sobre el asunto de algún mensaje que le hayan transmitido, hay personas que tienden a ser breves y concisas, y hay otras que entienden que el asunto es algo más extenso. A su vez pensando en casos de uso, nótese que es complejo buscar el asunto anterior, cuando la propia “especificación” de *esta práctica* está en el mensaje.

Opinión 34.3

```
<opinion>
<mensaje><asunto>espera hacerlo</asunto> hasta el fin de su carrera</mensaje>
según lo
<predicado>anunció
</predicado>
</opinion>
```

Otro caso de cierta complejidad es cuando el asunto es la acción principal de algún sujeto, que en este caso, además es implícito.

Como último comentario respecto al análisis, hay que tener presente que los casos mostrados anteriormente son tomados tras una revisión general y no necesariamente incluyen todos los tipos de problema que se hayan presentado en el corpus.

4 Solución propuesta

En esta sección se hará una síntesis de los criterios para identificar asuntos, se presentarán las características de los elementos de los que se parte para identificar el asunto describiendo el archivo de entrada, como también se presentará la arquitectura de la solución centrándonos en el núcleo de la solución. En el núcleo de la solución se destaca la generación de reglas para agrupamiento y el uso de árboles de dependencias para cuestiones lingüísticas de cierta complejidad. Se hace una sucinta descripción de las herramientas usadas, del corpus de trabajo y de la línea base.

Finalmente se hará una enumeración de las principales dificultades encontradas y se complementa con información sobre opiniones anidadas.

4.1 Síntesis de criterios para identificar asuntos

Los criterios surgidos del proceso de anotación fueron los siguientes:

1. El asunto en general es el primer grupo nominal del mensaje, por ejemplo la entidad de la que se habla o el sujeto que ejecuta la "acción principal" sin tener en cuenta la acción que están realizando, salvo que el grupo nominal no aporte "demasiada información".
2. El asunto no debería ser el lugar en que se desarrolla la acción, ni una referencia temporal.
3. Se extiende el asunto con oraciones subordinadas o algunos complementos, por entenderse que estas aportan a la comprensión del asunto. Además, desde un punto de vista sintáctico, forman parte del grupo nominal.

Para reafirmar el punto 1, ver que en:

Opinión 4.5

<opinion>
Según <fuente>información que difundieron a través de su abogado,</fuente>
<mensaje>"**<asunto> la demanda </asunto>** supera el millón de dólares, al amparo de lo establecido por el Articulo 4 de la Ley 15.859"</mensaje>
</opinion>

se incorporó al asunto la entidad principal (la demanda) y no "lo que hace" (superar el millón de dólares).

Lo mismo ocurre en el siguiente ejemplo:

Opinión 9.6 simplificada

<opinion> De acuerdo a lo
<fuente>lo indicado a la agencia
por el abogado de Infanzón, Darío Saldaño</fuente>,
<mensaje>**<asunto>el programa conducido por Ernestina Pais</asunto>** realizó una investigación sobre la comercialización de combustible y otros negocios de su representado</mensaje>
</opinion>

donde además de no incorporarse la acción principal al asunto, se ve el caso de la voz

pasiva, donde el asunto no es solo Ernestina Pais, sino su programa, ya que la voz pasiva usualmente se utiliza para resaltar la importancia del *objeto directo*.

Un ejemplo de 2 es:

Opinión 11.12

<opinion> Además (<fuente>Carlos Julio Pereyra</fuente>)
<predicado>recordó</predicado>
<mensaje>que en ese momento contrajo "**<asunto>un compromiso público</asunto>**" pues "como legislador tenía una postura tomada" que fue "la misma como ciudadano"</mensaje>
</opinion>

pues el asunto no es “ese momento” (referencia temporal).

Como ejemplo de 3:

Opinión 5.3
(El ex ministro de Economía, Danilo Astori , retornó de Estados Unidos y ...)

<opinion>
<predicado>informó</predicado>
<mensaje>que mantuvo **<asunto>una reunión con el subsecretario de dicho organismo financiero</asunto>**</mensaje>
</opinion>

ya que no solo “una reunión” es el asunto, sino también con quién fue mantenida, que es información complementaria. Notar que el grupo preposicional “con el subsecretario...” no forma parte del grupo nominal. Pero sí se incorporó al asunto, pensando en posibles casos de uso.

4.2 Corpus de trabajo

En lo cuantitativo se tiene un corpus etiquetado con unas **305 opiniones** de 38 noticias diferentes de las cuales **73 (24%)** ya tienen un asunto marcado por tener alguna palabra clave que permite al sistema en que nos basamos identificar al asunto mientras que unas 13 opiniones **(4%)** no tienen mensaje. En el corpus se etiquetaron manualmente los asuntos de acuerdo a los criterios descriptos al comienzo de la sección. Es de destacar que este es un corpus pequeño.

La anotación de asuntos fue un proceso que tomaba como retroalimentación los propios ejemplos dentro del corpus, donde los criterios finalmente adoptados fueron el resultado de refinamientos sucesivos que buscaban dar cierta armonía a lo que se considerase un asunto, además de repercutir directamente en la dificultad de su identificación. Fueron necesarias varias iteraciones y revisiones de casos particulares como para llegar a criterios que resultasen satisfactorios como a su vez sencillos en su enunciación y aplicación.

4.3 Archivo de Entrada

Le llamaremos *entrada* del sistema que estamos construyendo a un XML que tiene anotadas las opiniones como sus elementos estructurales. Además esta entrada es la salida del sistema descripto en [Rosá11], que toma información de FreeLing. Como se señaló anteriormente el asunto ya lo tienen incorporado las opiniones en que en el mensaje tiene un indicador particular, por ejemplo la preposición *sobre*. Cada token de la entrada cuenta además con información morfosintáctica, que naturalmente se expresa en el XML mediante etiquetas con atributos. Este conjunto de etiquetas y sus atributos se basan en las etiquetas propuestas por el grupo EAGLES[11] para la anotación de lexicones y corpus de todas las lenguas europeas. Además se tienen marcados los grupos nominales (gn).

Para introducir el etiquetado, veamos el siguiente ejemplo:

"No es que haga desaparecer la común, sólo que aparece más la A en los puestos centinela", dijo a Montevideo Portal María Julia Muñoz.

tiene el siguiente análisis:

```
<opinion>
<mensaje>
<F Lema="""" Atributos="[abre]">"</F>
<R Lema="no" Atributos="[N]">No</R>
<V Lema="ser" Atributos="[S, I, P, 3, S, 0]">es</V>
<C Lema="que" Atributos="[S]">que</C>
<V Lema="hacer" Atributos="[M, S, P, 3, S, 0]">haga</V>
<V Lema="desaparecer" Atributos="[M, N, 0, 0, 0, 0]">desaparecer</V>
<gn Atributos="[A, común, simple, C, S, 0, 0]">
  <D Lema="el" Atributos="[A, 0, F, S, 0]">la</D>
  <A Lema="común" Atributos="[Q, 0, C, S, 0]">común</A>
</gn>
<F Lema="," Atributos="[c]">,</F>
<R Lema="sólo" Atributos="[G]">sólo</R>
<P Lema="que" Atributos="[R, 0, C, N, 0, 0, 0]">que</P>
<V Lema="aparecer" Atributos="[M, I, P, 3, S, 0]">aparece</V>
<R Lema="más" Atributos="[G]">más</R>
<gn Atributos="[P, a, simple, 0, 0, V, 0]">
  <D Lema="el" Atributos="[A, 0, F, S, 0]">la</D>
  <N Lema="a" Atributos="[P, 0, 0, V, 0, 0]">A</N>
</gn>
<S Lema="en" Atributos="[P, S, 0, 0]">en</S>
<gn Atributos="[C, puesto, simple, M, P, 0, 0]">
  <D Lema="el" Atributos="[A, 0, M, P, 0]">los</D>
  <N Lema="puesto" Atributos="[C, M, P, 0, 0, 0]">puestos</N>
</gn>
<N Lema="centinela" Atributos="[C, C, S, 0, 0, 0]">
   <gn Atributos="[C, centinela, simple, C, S, 0, 0]">centinela</gn>
</N>
<F Lema="""" Atributos="[cierra]">"</F>
</mensaje>
```

11 http://www.lsi.upc.edu/~nlp/tools/parole-sp.html

```
<F Lema="," Atributos="[c]">,</F>
<V Lema="decir" Atributos="[M, I, S, 3, S, 0]">
<predicado>dijo</predicado>
</V>
<S Lema="a" Atributos="[P, S, 0, 0]">a</S>
<N Lema="montevideo portal" Atributos="[P, 0, 0, O, 0, 0]">
  <gn Atributos="[P, montevideo portal, simple, 0, 0, O, 0]">Montevideo Portal</gn>
</N>
<N Lema="maría julia muñoz" Atributos="[P, 0, 0, S, P, 0]">
<gn Atributos="[P, maría julia muñoz, simple, 0, 0, S, P]">
<fuente>María Julia Muñoz</fuente>
</gn></N>
</opinion>
```

Cuadro de análisis con etiquetado.

Notar que aparte de las etiquetas EAGLES se ven agrupamientos que contienen más de un token al demarcarse un grupo nominal. En particular, como se señala en [ARZ2010] el formato utilizado para los atributos de un grupo nominal (gn) es:

```
<gn>: [categoría, lema, ..., <atributos del núcleo>]
```

Se destaca que las categorías son:

P	Nombre Propio
N	Nombre Común
Pron	Pronombre
A	Adjetivo

Para entender un poco el cuadro de análisis con etiquetado se muestran las distintas categorías:

Categoría	Código
Adjetivo	A
Adverbio	R
Determinante	D
Nombre	N
Verbo	V
Pronombres	P
Conjunciones	C
Interjecciones	I
Preposición	S
Signo de puntiación	F
Cantidad*	Z

*: Es una extensión de EAGLES. No posee atributos.

4.4 Arquitectura de la solución

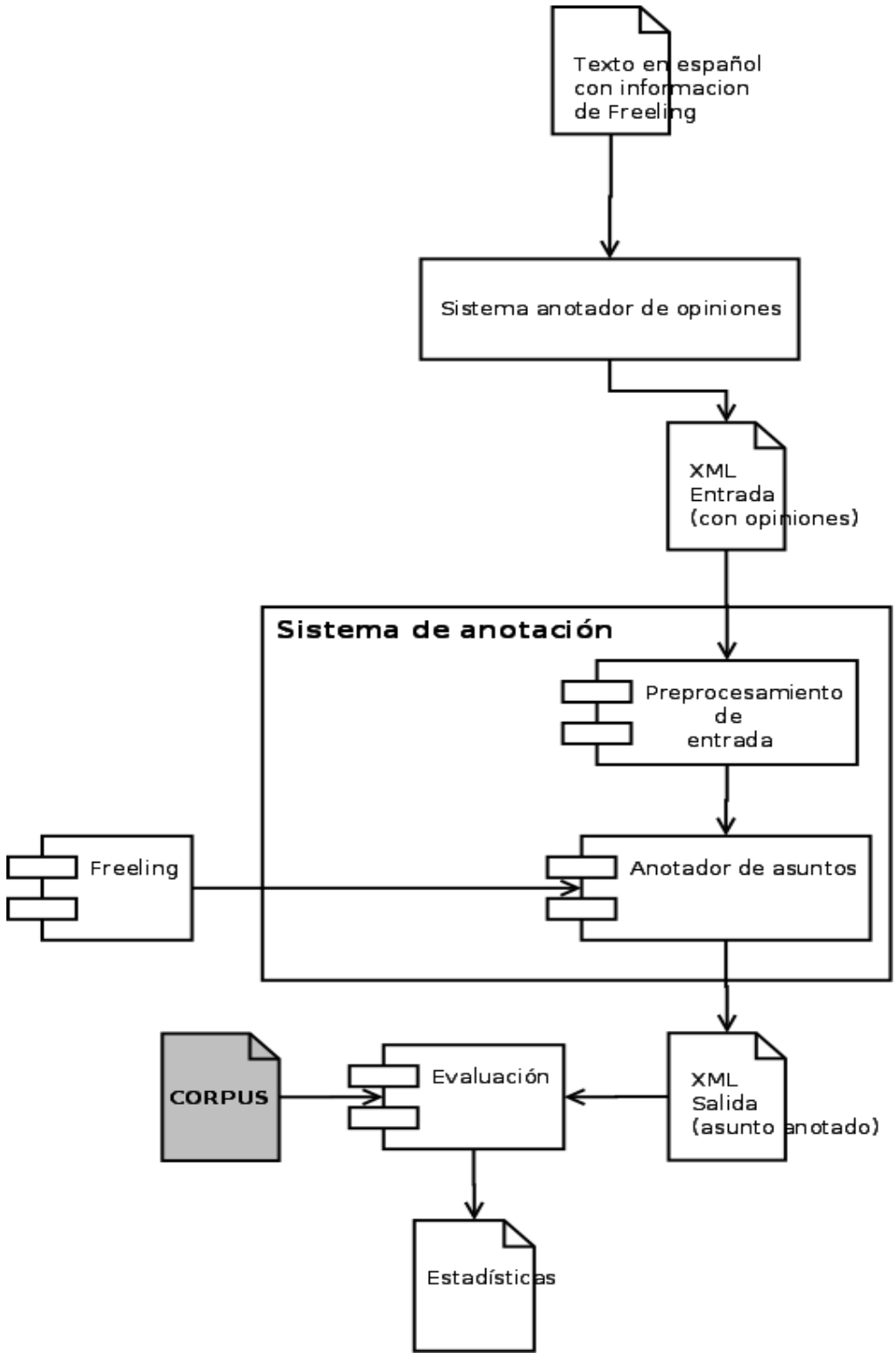


Figura – Arquitectura de la solución

En la figura se puede ver la arquitectura de la solución:

- El sistema anotador de opiniones no forma parte de nuestro sistema, pero es necesario que exista para poder obtener un XML adecuado para nuestro sistema.

- El XML de entrada puede tener o no las etiquetas morfosintácticas del grupo EAGLES, pero necesariamente debe tener los elementos estructurales de la opinión.
- La herramienta FreeLing la usamos para incorporar los elementos morfosintácticos a aquellos mensajes que no los tienen, también se obtienen los árboles de dependencia de un texto.
- El módulo de preprocesamiento de entrada se usa para corregir elementos menores del XML de entrada.
- El módulo de anotación de asuntos, núcleo de nuestro sistema, se detalla en la siguiente subsección.
- El módulo de evaluación, compara el resultado de nuestro sistema con el *gold standard* que es el corpus, produciendo resultados estadísticos de precisión, recuperación y medida F. También incorpora funciones para mostrar en pantalla de diversas maneras los árboles etiquetados, para facilitar la visualización de resultados y por ende su comparación con el corpus, como la elaboración de nuevas reglas.
- El XML de salida tiene las etiquetas de asunto agregadas.

4.5 Diseño de la solución

El módulo anotador de asuntos comprende 5 etapas distintas que se combinan:

1. El uso de agrupamiento
2. El uso de algoritmia simple que busca corregir detalles para mejorar el reconocimiento
3. Información posicional
4. Lista negra
5. El uso de la información del árbol de dependencias de FreeLing

Es importante destacar que todas estas etapas se aplican a cada mensaje de cada opinión de manera independiente, a su vez, el orden en que se aplica cada etapa se puede ver en la sección Flujo de trabajo.

En lo que resta de la subsección se verá una profundización de cada etapa:

4.5.1 Agrupamiento

Como se señala en [JM2008], muchas tareas de procesamiento no necesitan un árbol sintáctico completo, sino que basta con un parseo superficial/parcial (o *shallow parse*), a este parseo de agrupar segmentos que no se superponen se le llama *chunking*.

Estrictamente no estamos haciendo chunking, pues este es para reconocer en general grupos nominales, o grupos verbales, y si bien se utiliza para reconocer una “entidad

lingüística” como es el asunto, no cae dentro de los usos usuales del chunking.

Por lo tanto, llamaremos agrupamiento cuando se procesa la entrada de la siguiente forma: Se marca un conjunto de tokens como asunto, que siguen determinado patrón (o regla), teniendo en cuenta que si el token no pertenece a un grupo nominal, se toma la categoría de los tokens, cuando el token pertenece a un grupo nominal, se toma solo esa información y se descartan otras características morfosintácticas.

Por lo tanto, para la opinión cuyo texto de mensaje es:

```
que " tiene por <gn>objetivo evidente</gn> <gn>el congelamiento de la realidad electoral</gn>"
```

Se tiene el siguiente etiquetado:

```
[(u'que', u'C'), (u'"', u'F'), (u'tiene', u'V'), (u'por', u'S'),
(u'objetivo', 'gn'), (u'evidente', 'gn'), (u'el', 'gn'),
(u'congelamiento', 'gn'), (u'de', 'gn'), (u'la', 'gn'), (u'realidad',
'gn'), (u'electoral', 'gn'), (u'"', u'F')]
```

Ya que su correspondiente XML de entrada es:

Opinión en el XML de entrada

```
<opinion>
<V Lema="agregar" Atributos="[M, I, S, 3, S, 0]">
<predicado>agregó
</predicado>
</V>
<mensaje>
<C Lema="que" Atributos="[S]">que</C>
<mensaje>
<F Lema="&amp;quot;" Atributos="[abre]">"</F>
<V Lema="tener" Atributos="[M, I, P, 3, S, 0]">tiene</V>
<S Lema="por" Atributos="[P, S, 0, 0]">por</S>
<gn Atributos="[C, objetivo, simple, M, S, 0, 0]">
<N Lema="objetivo" Atributos="[C, M, S, 0, 0, 0]">objetivo</N>
<A Lema="evidente" Atributos="[Q, 0, C, S, 0]">evidente</A>
</gn>
<gn Atributos="[C, congelamiento, simple, M, S, 0, 0]">
<D Lema="el" Atributos="[A, 0, M, S, 0]">el</D>
<N Lema="congelamiento" Atributos="[C, M, S, 0, 0, 0]">congelamiento</N>
<S Lema="de" Atributos="[P, S, 0, 0]">de</S>
<D Lema="el" Atributos="[A, 0, F, S, 0]">la</D>
<N Lema="realidad" Atributos="[C, F, S, 0, 0, 0]">realidad</N>
<A Lema="electoral" Atributos="[Q, 0, C, S, 0]">electoral</A>
</gn>
<F Lema="&amp;quot;" Atributos="[cierra]">"</F>
</mensaje>
</mensaje>
</opinion>
```

Para explorar esta parte de la solución se utiliza la herramienta chunker de *nltk*[12] que

12 Natural Languaje Toolkit (Python): http://nltk.org/

puede recibir reglas con una notación parecida a la de expresiones regulares.

Cabe aclarar que a lo largo del proyecto se probaron varios conjuntos de reglas, y algunas reglas fueron descartadas para tratar de encontrar el patrón a reconocer en otra etapa. Se presenta un cuadro que resume el proceso:

Reglas del sistema:
Regla 1: {<gn>+<S><gn>+}
Descripción: Marcar como asunto los conjuntos de tokens que de izquierda a derecha tengan:un gn, una preposición y un gn.
Regla 2: {<gn>+}
Descripción: Marcar como asunto todos los grupos nominales.

Reglas experimentales:
Regla 4. {<gn>+<cc>+}
Descripción: Marcar como asunto todos los grupos nominales seguidos de un complemento circunstancial.
Regla 5. {<gn>+<grup-sp>+}
Descripción: Marcar como asunto todos los grupos nominales seguido de un grupo preposicional (notar que es una generalización de la Regla 1)
Regla 6. {<gn>+<subord-rel>+}
Descripción: Marcar como asunto todos los grupos nominales seguidos de una oración subordinada relativa.

Reglas descartadas:
Regla 7. {<gn>+<F><gn>+}
Descripción: Marcar como asunto los conjuntos de tokens que de izquierda a derecha tengan:un gn, un signo de puntuación, un gn
Regla 8. {<D>+<S>+<gn>+}
Descripción: Marcar como asunto los conjuntos de tokens que de izquierda a derecha tengan:un determinante, una preposición, un gn
Regla 9. {<gn>+<F><gn><F>}
Descripción: Marcar como asunto los conjuntos de tokens que de izquierda a derecha tengan: un gn seguido de un gn entre dos signos de puntuación.

El sistema tiene pocas reglas en esta etapa porque se vio que era más complicado el tratamiento del texto en etapas posteriores como peores los resultados en las evaluaciones.

Las reglas experimentales son útiles cuando se las combina con una etapa de añadir información de la función sintáctica que se obtiene con FreeLing mediante el parseo de dependencias a los tokens, pero no mejoraron sustancialmente los resultados por lo que no se incorporaron en el sistema final, se verán ejemplos en una subsección posterior, cuando se hayan introducido los árboles de dependencias.

Las reglas descartadas sirven para ilustrar las complicaciones que se pueden generar ante un etiquetado temprano de asuntos que no conduzca a nada bueno, por ejemplo, la regla 1 ({<gn>+<F><gn>+}), donde F es cualquier signo de puntuación, eventualmente reconoce asuntos con un punto en el medio. En el ejemplo:

"Si era por mí me hubiera quedado **<gn>10 años</gn> <gn>en Nacional</gn> (. , F) <gn>Las razones</gn>** de por qué no vine antes ya no me interesan y no quiero ni pensar en eso. Lo que importa es que estoy nuevamente en el club de mis amores"

se tomaría todo lo resaltado en negrita como asunto, cuando lo que interesa son los

primeros grupos nominales (el asunto de este caso se entendió que era “me hubiera quedado 10 años en Nacional”).

Podría restringirse a que el signo de puntuación para reconocer como asunto, sea una coma, y aunque se subsane esto, se notó que en general los asuntos no están de los “dos lados” de la coma, como en:

A) (…) "Nosotros estudiamos, a través del **<gn>Laboratorio de Higiene Pública</gn>(, , F) <gn>la vigilancia epidemiológica</gn>** (…)
B) Según **<gn>este relevamiento</gn> (, , F) <gn>las emisiones</gn>, la calidad de aire y del agua</gn>** arrojan parámetros buenos y muy buenos.

si bien hay algunas excepciones, por ejemplo:

C) (…) y permitió que <gn>el actual ministro de Ganadería(, , F) Ernesto Agazzi</gn>(...)

en la cual se da el fenómeno que hay un grupo nominal, una coma, y luego un grupo nominal que especifica el grupo nominal anterior. En el caso C, ambos grupos nominales forman parte del asunto, mientras que en los casos A y B, forma parte del asunto el grupo nominal que está a la izquierda de la coma.

La regla 3 busca reconocer grupos nominales cuyas siglas se especifican entre paréntesis, por ejemplo “Producto Bruto Interno (PBI)”. Esta regla fue descartada pues se reconoce con menos ruido mediante lo que llamamos algoritmia de corrección.

El hecho que se use el “+” para los grupos nominales, que significa 1 o más grupos nominales, es por un defecto de construcción, ya que cada palabra dentro de un grupo nominal tiene asociada la etiqueta de grupo nominal. Tiene la desventaja de que se reconoce como un asunto, dos grupos nominales distintos que estén uno a continuación del otro, de todas formas en la práctica se vio que esto no perjudica los resultados, e incluso reconoce indirectamente algunos asuntos, que en algunas ocasiones es el asunto, por ejemplo:

(…) tiene por [objetivo evidente] [el congelamiento de la realidad electoral]

Que tiene 2 grupos nominales (siendo el segundo el asunto), es tomado en esta etapa como:

tiene por <asunto>objetivo evidente el congelamiento de la realidad electoral</asunto>

por lo que se logra un reconocimiento parcial.

4.5.2 Información posicional

En el proceso de anotar el corpus manualmente se notó que el asunto generalmente estaba al comienzo del mensaje, y que en particular era el primer grupo nominal que aparecía en el mensaje. De la etapa de agrupación se desprende que se marcan más de un asunto por mensaje (tantos como grupos nominales haya por ejemplo). Nuestro sistema devuelve el primer asunto identificado leyendo de izquierda a derecha, descartando el resto. Si bien es un criterio sencillo, resultó ser de los más útiles y fue

utilizado para construir la línea base y la solución final, con un reconocimiento bastante alto.

4.5.3 Lista negra

Se utiliza una lista de palabras que no deberían de estar en el asunto, que le llamamos *lista negra*:

Ejemplo de lista negra

```
[<pronombres>, <referencias de tiempo>, <palabras de frecuencia>,
<palabras vinculadas a fechas(meses, años, etc.)>]
```

en base a un análisis del corpus se vio que los pronombres personales, y las expresiones vinculadas a lugar/tiempo tienden a generar grupos nominales que no contienen al asunto.

Opinión 1.9

<opinion>
<mensaje>"**Nosotros** estudiamos, **a través del Laboratorio de Higiene Pública**, <asunto>**la vigilancia epidemiológica**</asunto>. (...)
</mensaje>,
<predicado>detalló</predicado>
</opinion>

En negrita y subrayado se ven los grupos nominales que se obtienen del XML de entrada. En el cual se ve que el pronombre “Nosotros”, no corresponde al asunto anotado.

En principio se puede pensar en eliminar todo asunto reconocido en una etapa anterior que contenga una palabra prohibida, luego se hizo el refinamiento de eliminar del asunto todas las palabras que hay hasta la última aparición de la palabra de la lista (inclusive).

Otro refinamiento podría ser tener distintas listas negras dependiendo del dominio que se esté analizando, pero no se incorporó esto último en nuestro sistema por lo pequeño del corpus..

4.5.4 Algoritmia de corrección

Se entiende algoritmia corrección la que ajusta el asunto ya identificado en base a reconocimiento muy simple de tokens o de estructuras.

Por ejemplo, cuando se busca reconocer un asunto de la forma *gn (SIGLAS)* en general no se capturan las siglas y los paréntesis como pertenecientes al asunto, por lo tanto es necesario diseñar un algoritmo que extienda el asunto marcado incluyendo las

siglas y los paréntesis dentro del asunto, en el entendido que siempre que aparecen siglas, son en referencia al gn anterior:

Opinión 36.16

El militar recordó
<mensaje>que <asunto>la jefa de la Liga Nacional por la Democracia **(LND)**</asunto> es hija del general Aung San, el héroe de la independencia birmana que fue asesinado en 1947
</mensaje>

En el ejemplo anterior se subraya lo reconocido por la etapa de agrupamiento como asunto, y luego, en negrita lo que se agrega con el algoritmo de corrección, que efectivamente permite reconocer el asunto completamente.

4.5.5 Árboles de dependencia

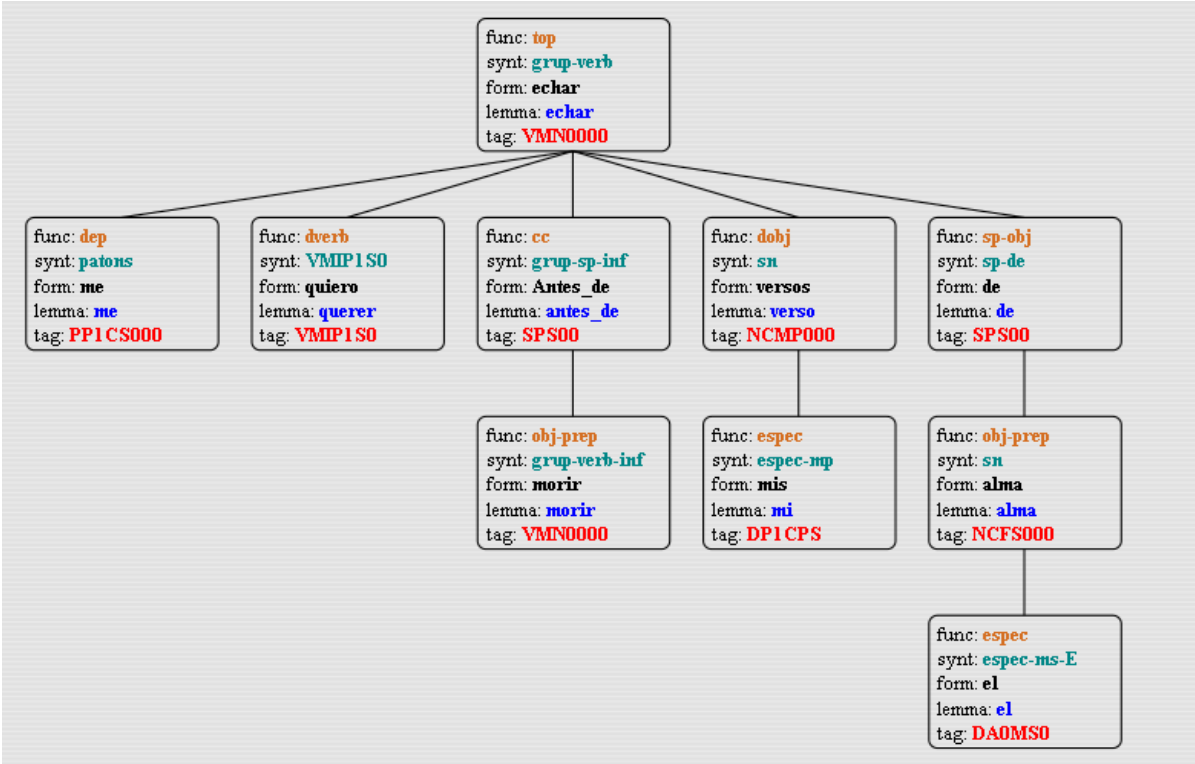
Las técnicas que permitieron hilar más fino y capturar fenómenos lingüísticos más complejos y más específicos fueron las aplicadas sobre el árbol sintáctico de dependencias producido por FreeLing [ACM2005][Padr2011][PS2012]. Un árbol sintáctico es un árbol que captura ciertas relaciones de dependencias entre las palabras de acuerdo al paradigma de las Gramáticas de Dependencia de Tesnière como indica [Tink2007] que lo resume de cierta forma:

> *(traducción) "citando a Vauvenargues "La ley suprema es subordinación y dependencia", Tesnière (…) indica que se establecen conexiones estructurales indicando relaciones de dependencia. Cada conexión relaciona un término superior (regente) con un término inferior (dependiente)."*

Por ejemplo la frase del conocido poema de José Martí:

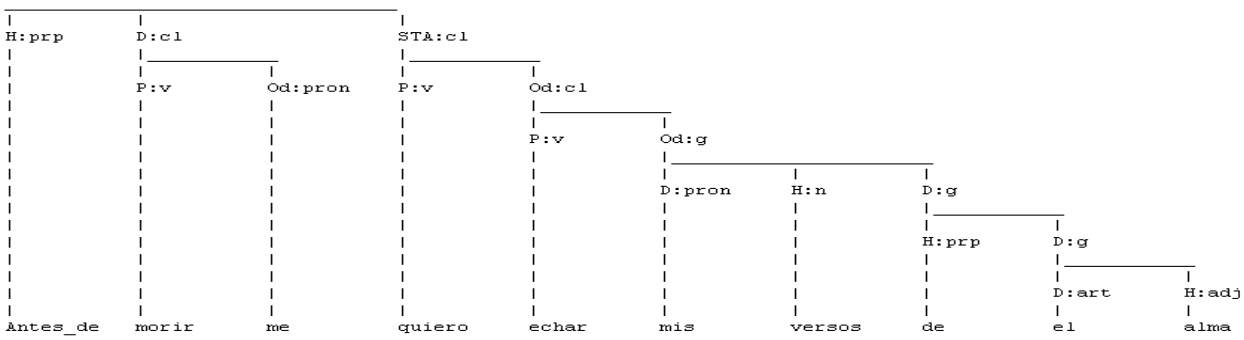
Antes de morirme quiero / echar mis versos del alma

Tiene el siguiente árbol de dependencia según FreeLing (no necesariamente correcto):



donde se ve que el verbo “echar” forma parte de un grupo verbal con función *top* (núcleo principal), y el token versos que forma parte de un *sn* (grupo nominal), cumple la *func* (función sintáctica) de *dobj* (mis versos es *objeto directo* del verbo echar).

Notar que el árbol presenta un error al establecer el pronombre *me* como dependiente de *echar*. Mientras en la siguiente figura que muestra el análisis de VISL se lo señala correctamente como dependiente de *morir:*



Por lo tanto de FreeLing se utilizan tanto las funciones sintácticas como las categorías para ayudar a reconocer asuntos, tomándose como insumo los asuntos marcados en etapas previas.

Tras algunas observaciones del corpus se crean las siguientes reglas:

1. Si el asunto marcado previamente es un objeto directo dependiente de algo que no es el verbo principal y hay un objeto directo dependiente del verbo principal, se marca como asunto un grupo nominal asociado (o dependiente) a este último objeto directo.
2. Si el asunto marcado está precedido por la palabra “en” entonces redefinir el

asunto como el verbo principal.

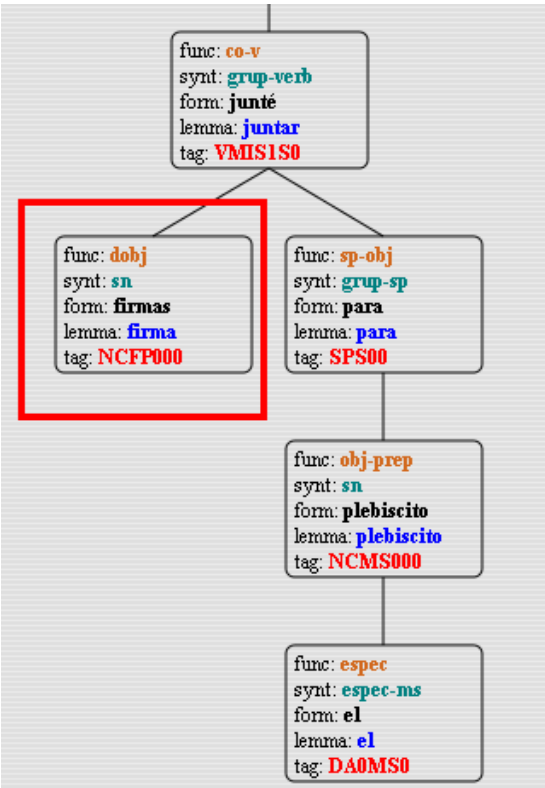
El primer caso se hace bajo la idea de entender que si hay un verbo principal, con un objeto directo que depende del mismo, probablemente este objeto directo sea “más importante” que cualquier otro, entonces corresponde marcarlo como asunto.

El segundo caso se debe a que si hay una acción (denotada por el verbo) y un lugar (precedido por en) probablemente se trate de un posible grupo nominal sin importancia por lo que damos prioridad a la acción antes que al lugar.

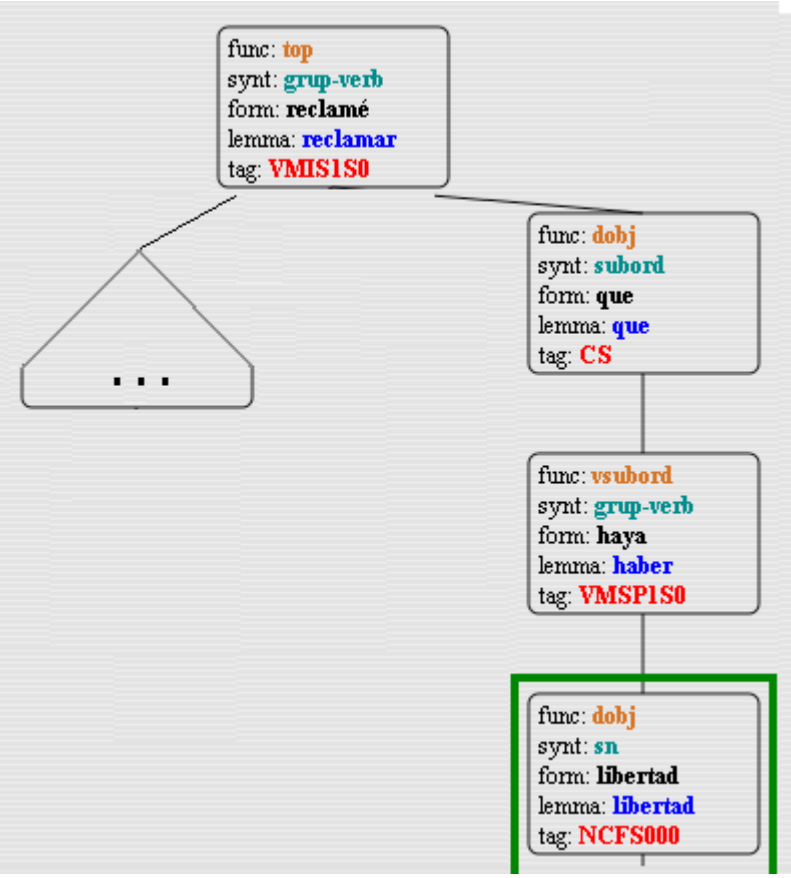
El primer caso se ilustra en el siguiente mensaje:

Yo mismo cuando voté en contra y junté firmas para el plebiscito, reclamé que haya libertad de acción.

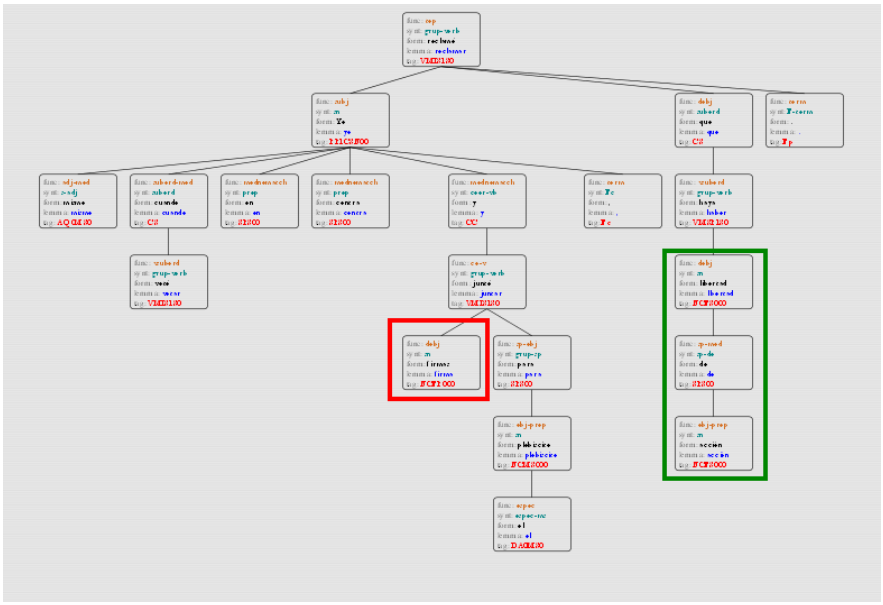
que viendo el árbol de dependencias en partes:



vemos recuadrado el asunto reconocido (por ser un grupo nominal y por su posición), pero que según nuestro último criterio, es un objeto directo de menor jerarquía.



En la figura anterior se ve recuadrado el comienzo de un grupo nominal que está en un objeto directo dependiente del verbo principal. Finalmente solo por completitud se muestra el árbol completo en una resolución poco visible[13] por razones de espacio:



Es claro que esta etapa es la más creativa de todas, donde el campo de posibilidades es realmente grande, se vieron algunas otras posibles situaciones donde utilizar las

13 El lector conectado a la red de redes puede apreciar el árbol de manera legible si utiliza la demo online de FreeLing: http://nlp.lsi.upc.edu/freeling/demo/demo.php

dependencias, como descartar del asunto ciertos complementos circunstanciales, incorporar algunas subordinadas y grupos preposicionales. Como ya se mencionó, estos elementos no se incorporaron a la solución final por no obtener buenos resultados. El desarrollo de estas soluciones estuvo limitado por el factor tiempo, ya que abstraer situaciones como la implementación de reconocimiento lleva un tiempo no menor. Por ejemplo, un buen criterio hubiese sido diseñar 10 patrones que contemplen la mayor cantidad de casos posibles, necesariamente en un corpus más grande del que disponemos.

4.6 Interacción entre dependencias y agrupamiento

Un camino que no se incorporó a la solución final, fue agregar etiquetas de complemento circunstancial, subordinadas relativas, o de grupo preposicional, (aparte de las de grupo nominal) como información adicional a usarse en la técnica de agrupamiento.

Por ejemplo, para el caso (con asunto en negrita) de:

> **Los niños que de forma preventiva recibieron Tamiflú** tuvieron efectos secundarios como náuseas o pesadillas.

si se etiqueta con la información de dependencias se obtiene:

> <gn>Los niños </gn> <subord-rel>que <cc>de forma preventiva</cc> recibieron Tamiflú</subord-rel> tuvieron efectos secundarios como náuseas o pesadillas.

Este es uno de los pocos casos que una regla de la forma <gn>+<sub-rel>+ permitiría reconocer el asunto. Notar que la regla <gn>+<cc>+ no reconocería este caso.

En:

> El director del Maciel explicó que hubo <gn>**un acuerdo bilateral**</gn> <cc>en hacer esa devolución en horas de trabajo</cc>, " porque como nosotros comenzamos a abrir más servicios y más salas, empezamos a requerir más horas de limpieza. Preferimos que nos devuelvan con más horas de trabajo "</mensaje>

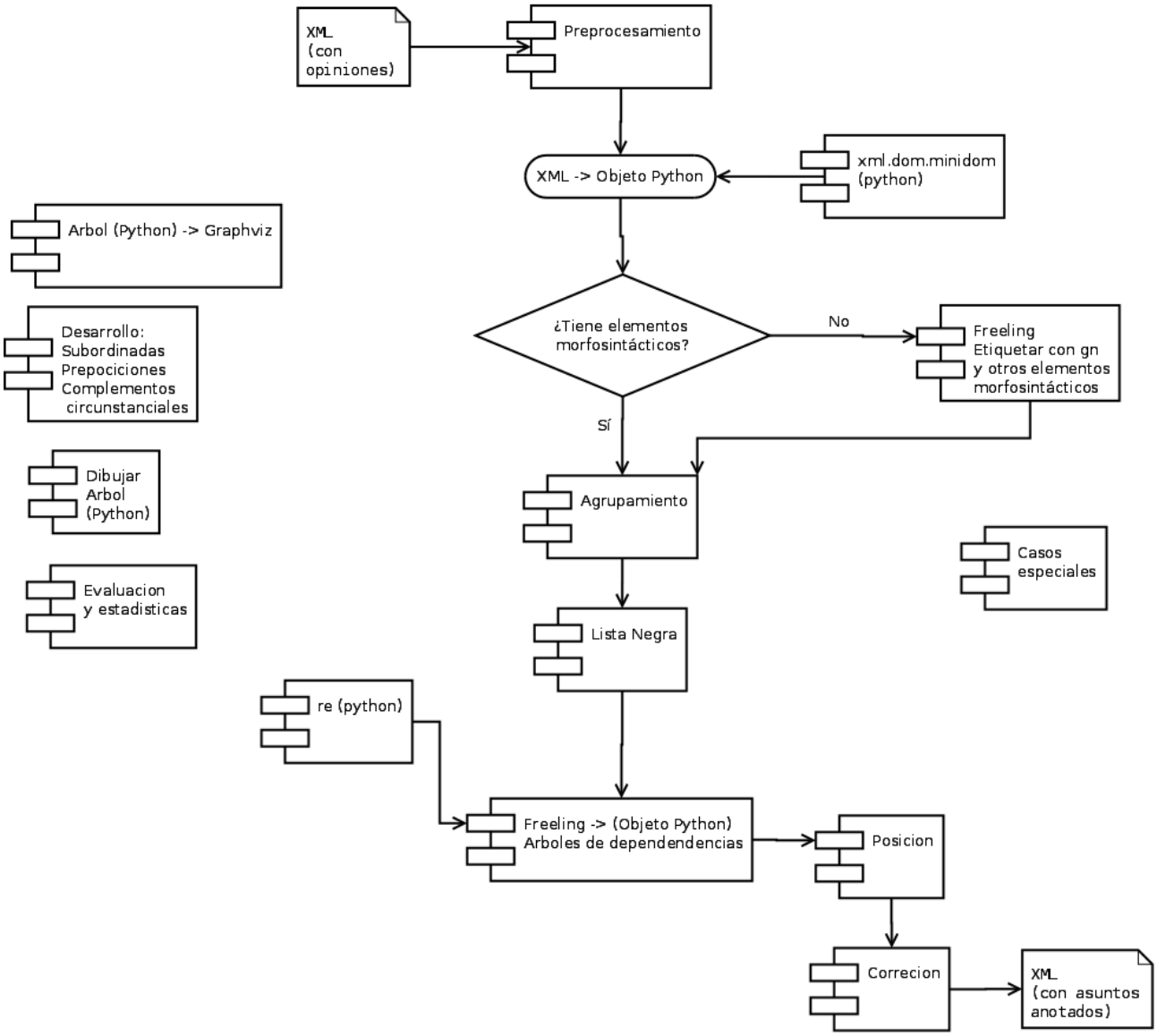
con la regla (<gn>+<cc>+) se reconocen textos más amplios, perjudicando el reconocimiento completo del asunto, rebajándolo a reconocimiento parcial.

4.7 Línea base

Se realizó la línea base obedeciendo una regla muy simple, tomar el primer grupo nominal del mensaje, luego de aplicar la etapa de agrupación, que logró una precisión y recuperación bastante buena desde el comienzo.

4.8 Flujo de trabajo

Ya comprendidos los principales procesos, se puede ver el flujo de trabajo típico:



- El elemento de entrada es un XML con una o varias opiniones.
- El módulo de preprocesamiento corrige elementos menores del XML
- Luego, usando xml.dom.minidom se pasa el texto del XML a un objeto Python.
- La única decisión que se muestra en el diagrama es si es necesario incorporar información básica faltante. El requerimiento mínimo es que la noticia tenga delimitadas la parte de opinión, fuente y mensaje. Si el XML no tiene las etiquetas de cada palabra, ni los grupos nominales, se utiliza FreeLing para incorporarlas. Esto no afecta el desempeño general pues el sistema en que nos basamos utiliza mecanismos parecidos.
- Luego comienza una sucesión de etapas, donde se aplican las técnicas detalladas, y en el orden que se muestran: Agrupamiento, lista negra, las técnicas que utilizan la información de los árboles de dependencias de FreeLing y finalmente se toma el primer asunto (Posición) y luego se hacen las correcciones (módulo Corrección).

- Hay que mencionar que hubo casos de opiniones que no funcionaron por diversas razones, para eso hubo que añadir cierta lógica para tratar de corregirlos, o eventualmente descartarlos. Por ejemplo, al haber múltiples transformaciones de árboles expresados de distintas maneras, tokens como “de” “él”, requerían cierto procesamiento para transformarlas en el token “del” con todo el cuidado que implica para el etiquetado y los árboles. También las locuciones “a_través_de” exigieron cierto cuidado.
- Se desarrolló un módulo para poder visualizar los árboles con la herramienta Graphviz[14], una visualización clara acelera el proceso de comprensión del problema y resultados, al facilitar el análisis caso a caso. También se pueden visualizar árboles con las propias herramientas de nltk (que se usa en varios módulos).
- En el procesamiento final de la identificación de asuntos no se utilizó el módulo desarrollado que tenía en cuenta algunos cuestiones vinculadas a preposiciones, complementos circunstanciales y oraciones subordinadas. Este módulo debería actuar como primera etapa, antes del agrupamiento, para poder añadir reglas de la forma <gn>+<cc>+ en el módulo de agrupamiento, para marcar como asunto los grupos nominales seguidos de un complemento circunstancial.
- También hay un módulo de evaluaciones y estadísticas, que interactúa con el corpus, y con los módulos de visualización.

4.9 Herramientas

En esta subsección hay una pequeña argumentación de todas las herramientas usadas, desde el lenguaje de programación, pasado por herramientas típicas de Unix, así como piezas clave como FreeLing y bibliotecas de Python, donde se destaca nltk.

4.9.1 Python

Se utilizó Python por ser un lenguaje moderno y expresivo, que permite enfocarse en el problema a resolver, además de contar con bibliotecas para el procesamiento de lenguaje natural bastante buenas.

4.9.2 Herramientas Unix

Se utilizaron las clásicas herramientas de texto para preprocesamiento de texto (*sed*, *egrep*) así como algunas vinculadas a la codificación (*file*, *iconv*). Fueron útiles en la eliminación de espacios en blanco para optimizar la velocidad de lectura de archivo, así como en la eliminación de alguna información sobrante. El uso de caracteres del español (eñe, tildes), que no son caracteres comunes de la anglosajona tabla ASCII, motivan un manejo con cierto cuidado de las codificaciones de los archivos.

4.9.3 FreeLing

FreeLing es una biblioteca de código abierto para el procesamiento automático de varias lenguas, entre ellas en español, que proporciona varios servicios de análisis

14 http://www.graphviz.org/

lingüístico: etiquetado gramatical, parsing de dependencias, entre otros. Si bien FreeLing es personalizable y ampliable, en nuestro proyecto lo usamos tal cual es, usando los análisis básicos que permiten cierta configuración a través de parámetros introducidos por la línea de comandos.

Elegimos FreeLing por estar disponible, ser fácilmente accesible e integrable a las etapas de procesamiento y en particular porque permite obtener árboles de dependencia para textos en español. Sobre los árboles de dependencia ya se profundizó en la subsección 4.5.5. Un análisis de la teoría detrás de FreeLing como de sus capacidades y limitaciones se puede profundizar en [Tink2007]. En nuestro sistema usamos la versión 3 de FreeLing.

4.9.4 Bibliotecas usadas

Paquete nltk:

Natural Language Toolkit [BKL2009]. Es un paquete muy completo y potente. Entre otras muchas cosas sirve para hacer tokenización de palabras, trabajar con chunkers y con otras herramientas vinculadas al PLN. También se utiliza un módulo para manipular árboles, que permite operar con ellos y dibujarlos en pantalla.

Paquete xml.dom.minidom:

Es una implementación mínima del Document Object Model. Se utilizó en el procesamiento y recorrida de los datos del XML de entrada, para obtener un objeto Python con la estructura de las opiniones, los atributos y las etiquetas.

Paquete pickle:

Permite la serialización de objetos Python. Utilizado en la optimización de no procesar cosas que son comunes a varias ejecuciones del programa, sobretodo en la generación de árboles y su recorrida.

Paquete re:

El uso de expresiones regulares permite hacer procesamiento de strings en pasos intermedios, útiles para el balanceo de paréntesis.

4.10 Dificultades encontradas

Se enumeran dificultades encontradas de distintos marcos: ya sea de implementación con las herramientas que se usan, de índole práctica particular a este proyecto o que involucren nociones de ingeniería de software, como también dificultades de los fundamentos teóricos.

4.10.1 Arrastre de error de otras herramientas

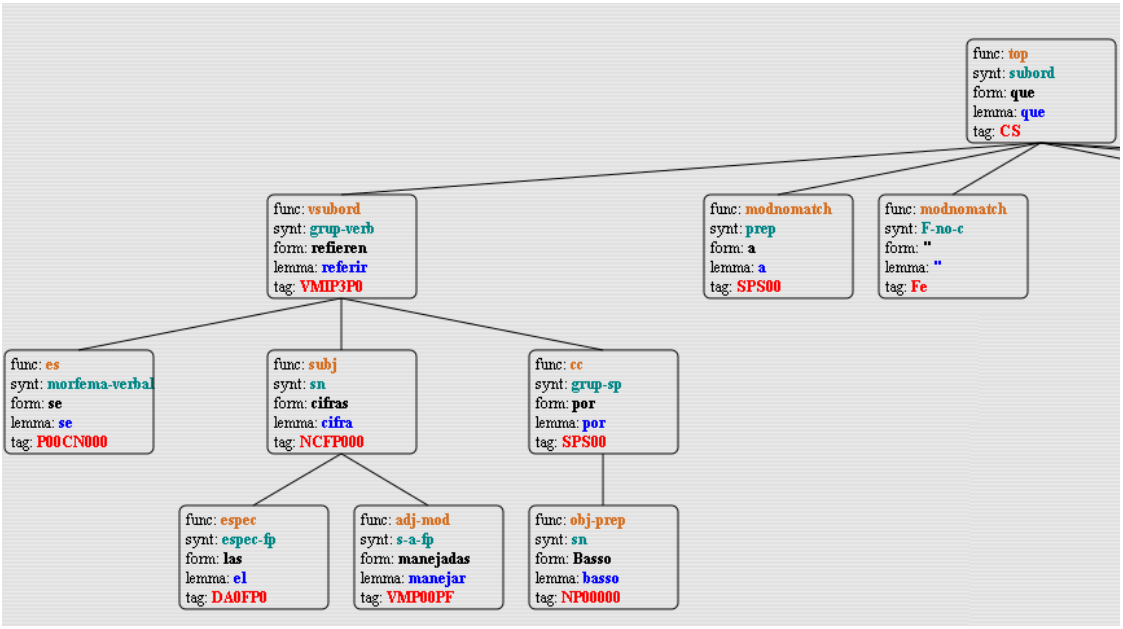
Como es sabido, en tareas de procesamiento de lenguaje natural el éxito rara vez cubre el 100% de los casos. Esto conlleva a que si bien teóricamente un módulo podría funcionar bien tiene como requisito recibir de otros módulos los datos correctamente procesados, cuando falla eso, el error se propaga. A continuación se muestra un caso donde FreeLing no devuelve un resultado correcto, lo que impediría reconocer el asunto

correctamente si nos basásemos en él[15].

Opinión 1.7

<opinion>
<fuente>Muñoz</fuente>
<predicado>precisó</predicado>
<mensaje>que **<asunto>las cifras manejadas por Basso</asunto>** se refieren a "los puestos centinelas del MSP" en distintas instituciones médicas</mensaje>
</opinion>

Suponiendo que en pasos previos se logró identificar la parte subrayada como asunto (grupo nominal al comienzo del mensaje), se buscaría ver si “por Basso” es dependiente del gn anterior, para extender el asunto. FreeLing recibiendo el texto del mensaje como entrada, devuelve el siguiente árbol:

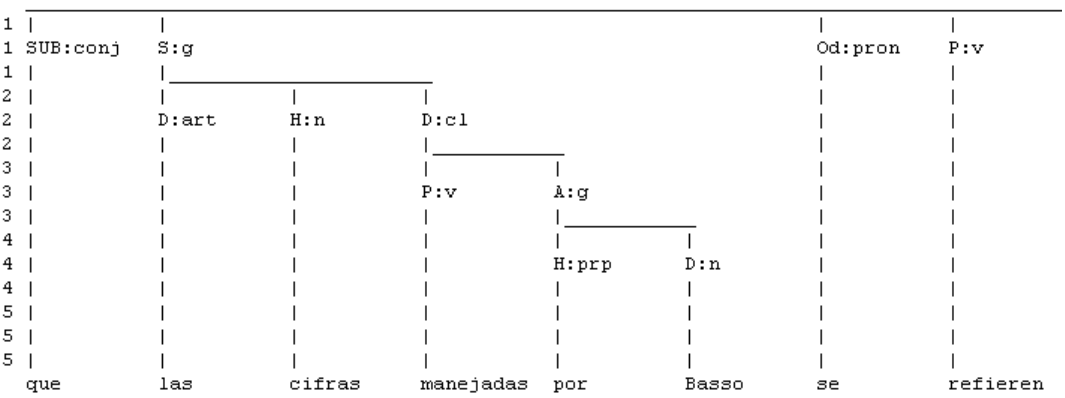


Lo cual no aporta en la solución pues se marca “por Basso” como un complemento circunstancial (y por tanto dependiente) de “refieren”, cuando debería tener una dependencia de “manejadas”, para así poder extender el asunto a palabras dependientes.

Sin embargo la herramienta VISL[16] sí hace un análisis mejor al marcar correctamente las dependencias:

15 Este asunto el sistema lo reconoce, por el uso de la regla <gn>+<S><gn>+

16 http://beta.visl.sdu.dk/visl/es/parsing/automatic/trees.php. VISL además obtiene árboles correctos en otros casos. No se usó porque solo se encontró accesible la versión online.



donde se ve que “por Basso” depende de manejadas.

Otra dificultad se presentó en el preprocesamiento de la entrada, cuando el ajuste de ciertas incongruencias resultó ser no trivial y terminaron realizándose con pérdida de información. Luego, se verificó que esta pérdida solo afectaba a pocas opiniones (menos de un 4%), por lo que se entendió como un problema menor.

4.10.2 Fundamentos básicos de lingüística

En el contexto formativo de un estudiante de ingeniería en computación, nociones básicas o de cierto nivel de elaboración del español, constituyen una limitante al momento de elaborar soluciones informáticas del área, por ejemplo en el siguiente texto:

> *El estudio*, que se basa en un análisis de los datos disponibles al cabo de ensayos comparativos de inhibidores de la neuraminidasa (enzima presente en el virus de la gripe) en los niños, subraya que **el Tamiflu** puede causar vómitos en algunos niños y que puede provocarles deshidratación y complicaciones

no resulta inmediato conceptualizar qué caso específico de *oración subordinada* es ni la función sintáctica que cumple la palabra *que* en ninguno de los tres casos, por ejemplo, si cumple una función de conjunción introduciendo una oración subordinada sustantiva en función sujeto, o si está como pronombre relativo. En el contexto de este proyecto poder realizar un análisis sintáctico de manera manual para una opinión puede llevar horas. Piense el lector también la dificultad que tiene el uso correcto de algo tan “sencillo” como la coma, repasando la normativa de la real academia[17]. Así como el cálculo es esencial a ciertas áreas de la ingeniería, las nociones de sintaxis de la gramática son herramientas esenciales para poder pensar tanto el problema que se trata como elaborar soluciones al mismo, siendo la conclusión obvia que este es un proyecto con cierta naturaleza interdisciplinaria.

4.10.3 Visualización de resultados

La visualización de los árboles obtenidos fue un obstáculo relevante, pues las visualizaciones por defecto (una humilde ventana con postscript) no son las mejores para hacer un análisis de los resultados ni tampoco son integrables trivialmente a una interfaz simple, por lo que construir una visualización más adecuada supuso un costo de tiempo que no se asumió y se trabajó a lo largo del proyecto con una interfaz de visualización algo pobre.

17 http://www.rae.es/dpd/?key=coma

4.10.4 Integración de sistemas, anotación manual

Si bien se trabajó con un corpus anotado manualmente, en principio se contaba con la salida de otro sistema que tenía la información morfosintáctica de los mismos ejemplos, pero existían pequeñas diferencias pues el corpus tuvo modificaciones de algunas etiquetas XML. A la larga, tratar las excepciones y ciertos corrimientos de numeración llevó decenas de horas. Una opción hubiese sido incorporar la información de FreeLing desde un principio, sin tener en cuenta la salida morfosintáctica del sistema anterior.

4.10.5 Administración/instalación del software

Por incompatibilidades varias entre varios sistemas operativos, la instalación de la versión 3 de FreeLing tomó unas 50 horas, llevándose una carga no menor de tiempo, y con un potente poder disuasivo para incorporar más herramientas.

4.11 Anidamiento del asunto en la opinión

En la medida que una opinión puede estar contenida en otra, se pueden dar complejidades al momento de determinar qué es un asunto. En este trabajo se asume la hipótesis más simple que es que, salvo excepciones, **cada opinión que contiene un mensaje y no tenga un asunto asociado fuera del mensaje debe tener un asunto contenido en el mensaje**. Por lo tanto, cada mensaje o bien tiene "su" asunto dentro de él, o al mismo nivel de profundidad por el procesamiento de alguna palabra clave (sobre, etc.). En la siguiente figura se ven 2 opiniones anidadas que no generan mayores problemas.

Opiniones 28.3, 28.4, 28.5 - Solapamiento que no genera conflictos.

<opinion>
<fuente>el diputado nacionalista Jorge Gandini</fuente>
<predicado>señaló</predicado>
<mensaje>que **<asunto>la interpelación</asunto>** se convocará formalmente el miércoles y que se tomó la decisión porque la semana pasada
<opinion>
<fuente>se</fuente>
<predicado>quiso</predicado>
<asunto>llamar a Comisión a las autoridades de ASSE</asunto>
</opinion> y
<opinion>
<asunto>la iniciativa</asunto>
<predicado>no contó con el apoyo</predicado> de
<fuente>los diputados del Frente Amplio</fuente>
</opinion>
</mensaje>
</opinion>

5 Pruebas y evaluaciones

La forma de medir en el área está bastante estandarizada, esto es, los resultados suelen medirse por medidas usuales (precisión y recuperación) por lo que es fácil en principio hacer una comparación entre distintos proyectos. Aunque como en todo, debe tenerse presente que si los conjuntos de datos son distintos, cada resultado refleja la aplicación de distintas técnicas a distintos datos. La sección comienza comentando cuantitativamente los recursos que se disponen así como el particionado del conjunto de datos. Los datos están separados en los usuales conjunto de desarrollo y conjunto de testeo. Se sigue con una introducción a los conceptos básicos de medidas de resultados.

En lo que respecta a los resultados en sí, primero se tiene en cuenta los resultados de nuestro proyecto buscando hallar una explicación de los resultados obtenidos. Para esto, se aprovecha el pequeño tamaño del corpus haciendo un análisis cualitativo para comprender mejor los resultados cuantitativos, buscando entender qué partes del sistema son las que producen los reconocimientos, o eventualmente las que dan problemas. Así como también señalando posibles categorías donde pueden agruparse ejemplos no reconocidos o reconocidos parcialmente.

Finalmente se ven los distintos resultados alcanzados en términos porcentuales poniendo el foco en los grupos de datos. Luego comparamos los resultados con otros proyectos relativamente parecidos usando las mediciones estándar.

5.1 Definición de los datos de prueba

De 38 noticias previamente extraídas de portales de noticias de la web, se tomaron 89 opiniones para toda la etapa de desarrollo y 89 para el testeo final en dos subgrupos distintos de 42 y 47 opiniones cada uno, es decir se trabajó con 178 noticias en total. El grupo de datos de testeo es útil para una medida más objetiva y más cercana al comportamiento en un entorno real. Los subgrupos de testeo no son los mejores posibles en la medida que el primero tiene muchas opiniones de las mismas noticias, que suelen presentar problemas parecidos y a su vez cuenta con noticias más difíciles de analizar[18]. Fue un error no haber distribuido desde un comienzo los grupos de desarrollo y de testeo. El segundo grupo tiene un etiquetado particular que no es fruto en su totalidad del sistema de [Rosá11], sino que se hace un procesamiento en nuestro sistema usando Freeling para el etiquetado de los grupos nominales y las etiquetas EAGLES. En particular los mensajes son obtenidos del corpus que ya está corregido manualmente.

5.2 Conceptos básicos

En recuperación de información existe un conjunto de métricas típicas para evaluar el desempeño de un sistema. En general se entiende que cuanto mayor sean ciertos valores, mejor es el sistema, sin hacer distinciones sobre qué cosas específicamente el sistema es capaz de clasificar correctamente. Para ello se utiliza la *precisión*, que es la cantidad de elementos relevantes recuperados dividido el total de elementos recuperados, también se utiliza la *recuperación*, que es la cantidad de elementos relevantes recuperados dividido el total de elementos relevantes. Una precisión alta, refleja que los datos que el sistema marcan como resultado, efectivamente son resultados correctos,

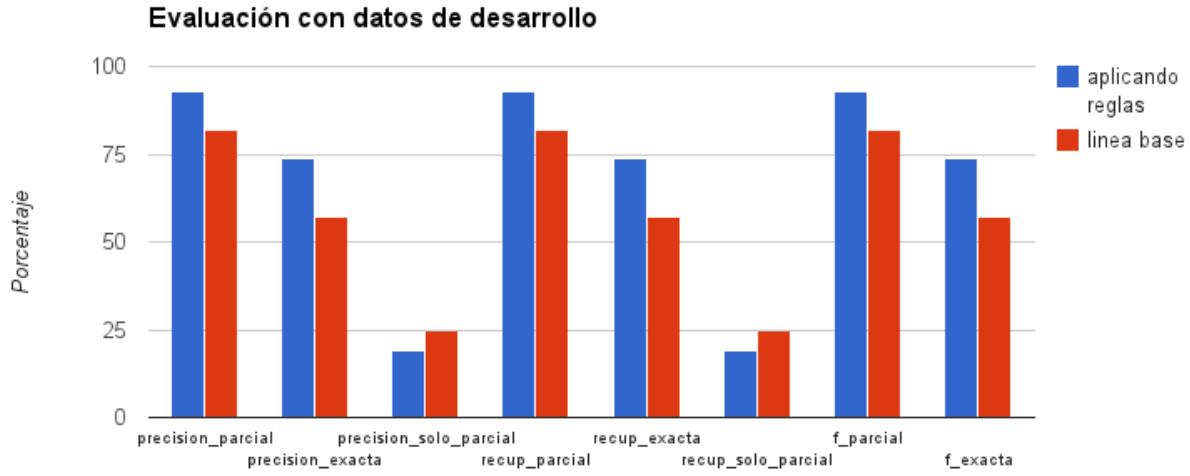
18 Curiosamente, para este trabajo resultaron ser las opiniones emitidas por jugadores de fútbol.

mientras que una recuperación alta, indica que de todos los documentos recuperables, se recuperó una cantidad sustanciosa. Si tal o cual porcentaje es bueno, depende del contexto y del estado del arte, para algunos sistemas un 30% de precisión será un resultado descollante, mientras que para otros un 80% de precisión puede ser un resultado desastroso. Piénsese el caso de Google, ya que cuando usamos un buscador en realidad solo nos interesa que haya buenos vínculos en los primeros lugares, sin importar demasiado cuál es la precisión entre los cientos de resultados devueltos, o cuál fue la recuperación de los potencialmente miles de documentos relevantes que existen en el planeta.

5.3 Resultados de nuestro trabajo

En este análisis la precisión exacta es la precisión de los asuntos que fueron identificados perfectamente, desde la primera a la última palabra, mientras que la precisión parcial, incluye los casos en que alguna parte del asunto fue identificada, incluso aunque se hayan marcado como parte del asunto palabras que no pertenecen al mismo. La precisión parcial importa, en la medida que si se hace un buscador, los reconocimientos parciales también indican lo que el sistema potencialmente puede encontrar, además del usuario poder recuperar información relevante del contexto en caso de devolver la información en su contexto. Por claridad en la gráfica se muestran los porcentajes de los asuntos que solo fueron reconocidos parcialmente de manera estricta (precisión_solo_parcial, recuperación_solo_parcial).

Para los datos de desarrollo los resultados son los siguientes:

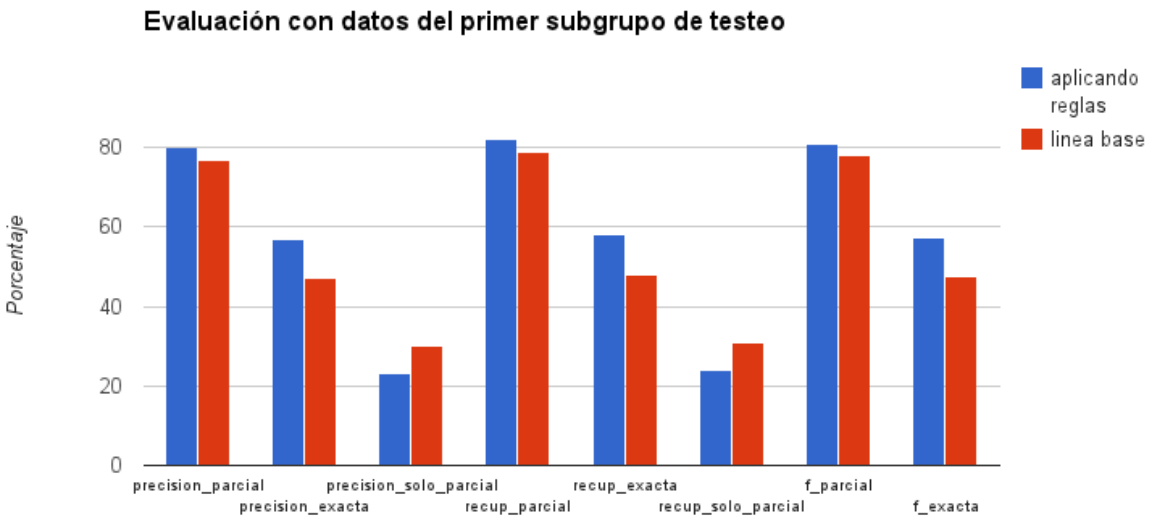


Es de notar que es considerable el incremento (17%) de un 57% a un total de 74% en medida F exacta aplicando reglas frente a la línea base, además la mayoría de los asuntos (57% recuperación exacta) son identificados trivialmente con los elementos usados en construir la línea base llegando a un 74% con reglas. En la precisión parcial se llega a un muy buen 93% con reglas y a un 82% en la línea base, teniendo presente que

se usó un criterio bastante inclusivo, pues podría haberse entendido el reconocimiento parcial de otra manera, por ejemplo que el asunto reconocido debe coincidir con el comienzo del asunto real. Se debe tener presente que la diferencia en precisión parcial entre el sistema y la línea base, no implica que se hayan mejorado todos los casos de precisión parcial a precisión exacta, podría darse el caso de que algunos asuntos se dejen de reconocer, aunque es claro que globalmente el sistema mejora al comparar las medidas F.

Se produce un fenómeno particular que es que la precisión es casi lo mismo que la recuperación. Ocurre que son muy pocos los casos en que el mensaje no tiene asunto, mientras que nuestro sistema siempre marca algún asunto en el mensaje, por lo que cuando se marca correctamente un asunto, incrementa tanto la precisión como la recuperación, mientras que si no se marca disminuirían los dos. En la sección 5.4 se verá en más detalle que ocurrió y algunos ejemplos de la aplicación del sistema.

Para el primer subgrupo de testeo la evaluación es la siguiente:



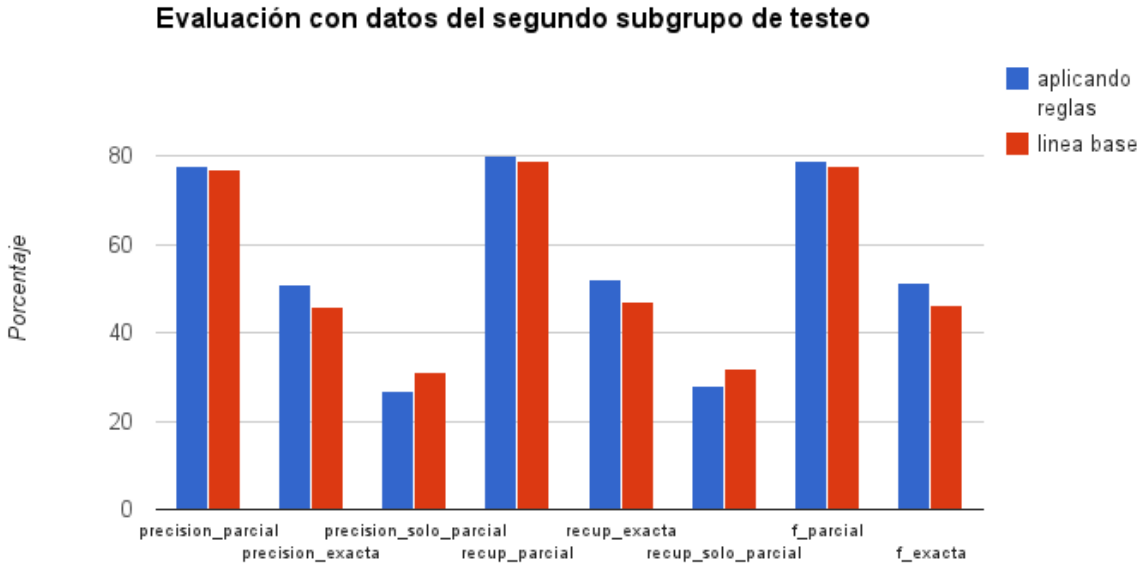
Se puede ver una mejora importante de 10 puntos porcentuales en la medida F exacta respecto a la línea base, mejora similar en la precisión exacta como en la recuperación exacta. La precisión parcial experimenta una leve mejora, de tan solo un 3%.

Hay que tener en cuenta que al estar trabajando sólo sobre 42 opiniones distintas difícilmente se dé la misma situación problemática que en los datos de elaboración de reglas. Tras una revisión no exhaustiva, se vio que gran parte de las mejoras se deben a la regla <gn><S><gn> que permite extender algunos asuntos para reconocerlos completamente.

También se debe tener presente que la mejora de 4 opiniones es lo que representa aproximadamente un 10% del total. Esto fomenta la idea que que no se pueden evaluar correctamente las reglas, pues al tener un corpus chico, no permite aventurar si algunas

reglas son demasiado específicas o incorrectas, tanto en la identificación de asuntos como en el incorrecto reconocimiento de otros.

El subgrupo 2 de testeo se hizo para atenuar el posible sesgo en los datos del subgrupo 1 por pertenecer en su mayoría a pocas noticias, ya que el subgrupo 2 sí tiene una selección variada de opiniones de distintas noticias, además que otro subgrupo permite incrementar la cantidad de opiniones de testeo. Tras la evaluación, se obtuvo el siguiente resultado:



En este subgrupo también se ven mejoras en la medida f exacta, aunque solo de un 5% y una mejora en la precisión parcial de un 1%. Este subgrupo de testeo en un comienzo no tenía cierta información del sistema en que nos basamos por lo cual se realizó un módulo que completa la información faltante con el etiquetado de FreeLing junto a los grupos nominales. Más allá de eso, analizaremos estos resultados en la sección 5.5, para entender mejor qué ocurrió, donde se verá qué reglas resultaron ser buenas y qué nuevos problemas permiten descubrir los datos.

5.4 Análisis de los resultados del grupo de datos de desarrollo

En el procesamiento de los datos de desarrollo, 17 asuntos fueron identificados parcialmente (estrictamente parcial), un 19% del total. A continuación se busca agrupar en categorías los problemas hallados listándolos por orden de importancia:

1. Verbos en el asunto
2. Sobresimplificación de gn
3. Dependencia no reflejada por FreeLing
4. Solucionable con correferencia
5. Dependencia reflejada por FreeLing pero no reconocida por la algoritmia

Tres solamente son los casos en que no se logra identificar nada de los datos de desarrollo, aproximadamente un 3% del total.

5.4.1 Detalle de algunas categorías

A continuación se profundiza en algunas de las categorías enumeradas anteriormente

5.4.1.1 Verbos en el asunto

En los siguientes recuadros con mensajes con etiquetado gramatical, se encontrará el asunto encontrado por nuestro sistema en negrita y resaltado en gris el asunto anotado manualmente.

En el siguiente caso:

```
[(u'que', u'C'), (u'aun', u'R'), (u'250', 'gn'), (u'personas', 'gn'),
(u'permanecen', u'V'), (u'detenidas', u'A'), (u',', u'F'), (u'entre',
u'S'), (u'ellas', 'gn'), (u'50', 'gn'), (u'personalidades', 'gn'),
(u'políticas', 'gn')]
```

resultó difícil encontrar un criterio por el cual extender el asunto e incluir los verbos, puesto que si se hace de manera trivial, (reconocer verbos después de un gn) se identificarían muchos *falsos positivos,* es decir, se darían por asuntos cosas que no lo son.

También son problemáticos los siguientes casos:

```
1.
[(u'"', u'F'), (u'Me', u'P'), (u'voy', u'V'), (u'a', u'S'), (u'retirar',
u'V'), (u'en', u'S'), (u'Nacional', 'gn'), (u'"', u'F')]

2.
[(u'que', u'C'), (u'no', u'R'), (u'sabe', u'V'), (u'nada', u'P'),
(u'del', u'S'), (u'asunto', 'gn'), (u'"', u'F')]
```

Donde el ejemplo 2 resulta especialmente difícil por contener una negación y un verbo en el asunto. El asunto también es mencionado de forma indirecta, justamente a través de la palabra *asunto*.

Estos casos indican que tiene que haber una comprensión más profunda del lenguaje y en particular de la sintaxis y dependencias que puedan indicar un asunto en lo que respecta a los verbos.

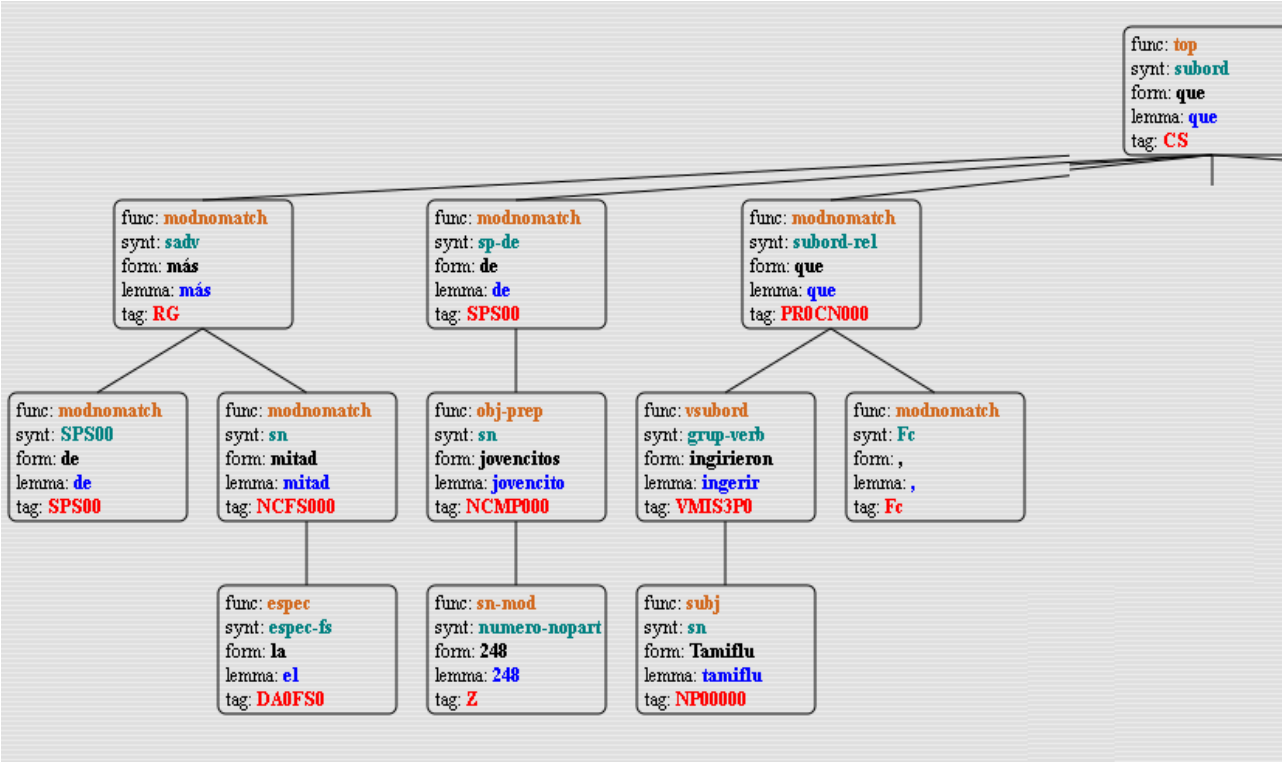
5.4.1.2 Dependencias no reflejadas por FreeLing

Para el siguiente mensaje (sin mostrar el etiquetado):

> que más de **la mitad** de 248 jovencitos que ingirieron Tamiflu, luego de que uno de sus compañeritos contrajera la gripe porcina, tuvieron efectos secundarios como náuseas, insomnios y pesadillas.

con el árbol de dependencia[19]:

19 Se observó que este caso al analizarlo con VISL se obtiene un análisis más correcto.



se puede ver que tanto la preposición del segundo “de” como la subordinada introducida por el segundo “que” no están dependiendo de lo que tienen que depender, por lo que se dificulta su identificación como asunto en caso de haber querido extender el grupo nominal reconocido como asunto por el sistema.

5.5 Análisis del subgrupo 2 de testeo

Como vimos anteriormente, los resultados para el subgrupo 2 de testeo fueron positivos, aunque como ya se señaló, mirar los porcentajes para una cantidad total muy baja (en este caso 47), puede ser engañoso. En particular, los cambios registrados se deben a 5 casos. Ya que al aplicar el sistema completo ocurre lo siguiente respecto a la línea base:

1. Tres asuntos reconocidos parcialmente por la línea base son reconocidos totalmente por el sistema
2. Un asunto se deja de reconocer al aplicar el sistema respecto a la línea base
3. Un asunto no reconocido por la línea base es reconocido parcialmente por el sistema

Respecto al ítem 1, los tres casos son similares a:

```
[(u'la', 'gn'), (u'primera', 'gn'), (u'búsqueda', 'gn'),
(u'policial', 'gn'), (u'en', 'S'), (u'la', 'gn'), (u'casa', 'gn'),
(u'no', 'R'), (u'tuvo', 'V'), (u'éxito', 'gn'), (u'debido', 'gn'),
(u'"', 'F'), (u'a', 'S'), (u'el', 'gn'), (u'enorme', 'gn'),
(u'volumen', 'gn'), (u'de', 'gn'), (u'pertenencias', 'gn'), (u'y',
'C'), (u'papeles', 'gn'), (u'que', 'P'), (u'habíaa', 'V'),
(u'dentro', 'R'), (u'"', 'F')]
```

donde la línea base identifica parcialmente el asunto por agarrar el primer grupo nominal. Mientras que el sistema completo por la regla <gn><S><gn> incorpora “en la casa” como parte del asunto obteniéndose el reconocimiento total.

Revisando los casos de desarrollo, se constata que la inclusión de la regla <gn><S><gn> para identificar asuntos provoca una mejora del 6% en la precisión exacta (respecto a no incluirla), mientras que para este subgrupo provoca una mejora del 7%. En la sección 5.3 también se observó que esta regla era útil en los datos del subgrupo 1 de testeo por lo que podemos aventurar que fue una regla bastante acertada.

Casos parecidos hay en los datos de desarrollo. Por ejemplo en el mensaje:

```
[(u'que', u'C'), (u'el', 'gn'), (u'99', 'gn'), (u'%', 'gn'), (u'de', 'gn'), (u'los', 'gn'), (u'casos', 'gn'), (u'de', 'gn'), (u'gripe', 'gn'), (u'en', u'S'), (u'Uruguay', 'gn'), (u'corresponden', u'V'), (u'al', u'S'), (u'virus', 'gn'), (u'H1N1', 'gn'), (u',', u'F'), (u'que', u'C'), ...]
```

también se incluye “el lugar” como parte del asunto. Este tipo de detalles impacta en los resultados finales, pues de haber anotado de forma distinta el corpus, la línea base hubiera obtenido reconocimiento total, mientras que el sistema hubiera reconocido parcialmente el caso, siendo peores los resultados.

En el ítem 2 se produce el infeliz caso en que una regla inducida por los datos de desarrollo en estos nuevos datos perjudica los resultados. La regla en cuestión es:

> Si el asunto marcado es un objeto directo dependiente de algo que no es el verbo principal y hay un objeto directo dependiente del verbo principal, se marca como asunto un grupo nominal asociado (o dependiente) a este último objeto directo.

y ocurre que en el siguiente mensaje:

```
[(u'que', 'C'), (u'no', 'R'), (u'hubo', 'V'), (u'sobrefacturación', 'gn'), (u',', 'F'), (u'sino', 'C'), (u'que', 'C'), (u'de', 'S'), (u'un', 'gn'), (u'año', 'gn'), (u'a', 'S'), (u'otro', 'gn'), (u'se', 'P'), (u'dejaron', 'V'), (u'de', 'S'), (u'realizar', 'V'), (u'gastos', 'gn'), (u'relacionados', 'gn'), (u'a', 'S'), (u'capacitación', 'gn'), (u'del', 'gn'), (u'personal', 'gn')]
```

FreeLing produce un árbol incorrecto donde “sobrefacturación” es un objeto directo de un verbo (“hubo”) que según FreeLing no es el principal. Si bien es el asunto real y además es reconocido por la línea base, nuestro sistema al aplicar la regla marca un objeto directo (“gastos relacionados...”) dependiente del “verbo principal” (que FreeLing marca erradamente como “realizar”), por lo que se llega a un asunto incorrecto.

El único caso que pasa de no reconocerse a parcial (ítem 3) es el siguiente:

```
[(u'"', 'F'), (u'yo', 'gn'), (u'no', 'R'), (u'vivo', 'V'), (u'de', 'S'), (u'eso', 'gn'), (u'"', 'F')]
```

donde se obtiene el reconocimiento parcial, debido a que la lista negra evita que el asunto sea el pronombre “yo”, identificándose como asunto el segundo grupo nominal. Notar que es un caso similar al presentado en la sección 5.4.1.1 pues es un asunto particularmente difícil de identificar por contener un verbo y una negación.

5.5.1 Asuntos no reconocidos por el sistema

Para el subgrupo en cuestión 10 casos (21%) no fueron reconocidos por lo que detallaremos algunos casos:

```
  [(u'que', 'C'), (u'el', 'gn'), (u'mes', 'gn'), (u'próximo', 'gn'),
(u'comienzan', 'V'), (u'las', 'gn'), (u'últimas', 'gn'), (u'audiencias',
'gn'), (u'en', 'S'), (u'La', 'gn'), (u'Haya', 'gn')]
```

Lo que ocurrió en este caso que el sistema identifica “próximo” como asunto, es que la técnica de lista negra permitió descartar hasta el token “mes” del asunto (inicialmente reconocido como “el mes próximo”), pero no es lo suficientemente buena como para también descartar próximo (que es un adjetivo vinculado a cuestiones de tiempo/lugar). De haberse descartado, la regla <gn><S><gn> hubiera identificado correctamente el asunto,

Un caso similar es:

```
  [(u'hace', 'V'), (u'miles', 'gn'), (u'de', 'gn'), (u'años', 'gn'),
(u'las', 'gn'), (u'mujeres', 'gn'), (u'tenían', 'V'), (u'que', 'C'),
(u'trabajar', 'V'), (u'sobre', 'S'), (u'todo', 'gn'), (u'en', 'S'),
(u'el', 'gn'), (u'espacio', 'gn'), ...]
```

donde la lista negra podría haber funcionado mejor si hubiese incluido la palabra “años”, para poder descartar “miles de años” e identificar el siguiente grupo nominal como asunto.

Más allá de estos casos particulares, el resto de casos que no se logra identificar el asunto pueden agruparse en las siguientes categorías:

1. El asunto no es el primer gn que aparece (varios casos)
2. El asunto contiene un verbo (2 casos)
3. No se anotó asunto en el corpus por ser un caso muy complejo (1 caso)
4. Las reglas que usan dependencias perjudican la identificación (1 caso)

Para el primer caso, a veces resulta difícil encontrar motivos para descartar el primer grupo nominal como asunto. Viendo un ejemplo:

```
  [(u'que', 'C'), (u'después', 'S'), (u'del', 'gn'), (u'Pato',
'gn'), (u'Aguilera', 'gn'), (u',', 'F'), (u'yo', 'gn'), (u'era',
'V'), (u'el', 'gn'), (u'mejor', 'gn'), (u'definidor', 'gn'),
(u'del', 'gn'), (u'mundo', 'gn')]
```

notar que también se podría haber optado por marcar como asunto el primer grupo nominal y considerar que el mensaje tenía más de un asunto.

Un ejemplo de 2 es:

```
  [(u'"', 'F'), (u'Si', 'C'), (u'después', 'R'), (u'eso', 'gn'),
(u'me', 'P'), (u'va', 'V'), (u'a', 'S'), (u'llevar', 'V'), (u'o',
'C'), (u'no', 'R'), (u'a', 'S'), (u'la', 'gn'),
(u'vicepresidencia', 'gn'), (u'depende', 'V'), (u'de', 'S'),
(u'las', 'gn'), (u'urnas', 'gn'), (u'.', 'F'), ...]
```

que como ya se mencionó es dificultoso para este trabajo el comprender verbos en el asuntos.

El ítem 3 es autoexplicativo y el 4 ya se mencionó.

5.5.2 Asuntos reconocidos parcialmente por el sistema

Hay 13 casos que son reconocidos parcialmente por el sistema, que se pueden agrupar en las siguientes categorías:

1. Reconocimiento de un grupo nominal muy amplio (varios casos)
2. Coordinaciones
3. Verbo en el asunto
4. Lista negra
5. Enumeraciones

A la primer categoría pertenecen casos como el ya mostrado en la sección 3.2.1.4 (opinión 27.2)

La segunda categoría permitió observar un nuevo tipo de problema que no se había visto en los casos de desarrollo. Este nuevo problema es el tratamiento de coordinaciones. Ocurrieron 4 casos y son los siguientes:

```
  1.
  [(u'pagaron', 'V'), (u'la', 'gn'), (u'garantía', 'gn'), (u'que',
'gn'), (u'quedaba', 'gn'), (u'pendiente', 'gn'), (u'y', 'gn'),
(u'el', 'gn'), (u'contrato', 'gn'), (u'permanecerá', 'V')]

  2.
  [(u'"', 'F'), (u'(', 'F'), (u'Los', 'gn'), (u'partidos', 'gn'),
(u'Colorado', 'gn'), (u'y', 'C'), (u'Nacional', 'gn'), (u')',
'F'), (u'son', 'V'), (u'como', 'C'), (u'esos', 'gn'), (u'primos',
'gn'), (u'que', 'gn'), (u'se', 'gn'), (u'ven', 'gn'), (u'poco',
'gn'), (u'pero', 'C'), (u'se', 'P'), (u'quieren', 'V'), (u',',
'F'), (u'y', 'C'), (u'cuando', 'C'), (u'necesitan', 'V'), (u'uno',
'gn'), (u'del', 'gn'), (u'otro', 'gn'), (u'estén', 'V'), (u'"',
'F')]

  3.
  [(u'"', 'F'), (u'liberación', 'gn'), (u'inmediata', 'gn'),
(u'y', 'C'), (u'sin', 'S'), (u'condiciones', 'gn'), (u'"', 'F'),
(u'de', 'S'), (u'Suu', 'gn'), (u'Kyi', 'gn')]

  4.
  [(u'que', 'C'), (u'sabía', 'V'), (u'que', 'C'), (u'había', 'V'),
(u'una', 'gn'), (u'conspiración', 'gn'), (u'clara', 'gn'), (u'contra',
'S'), (u'el', 'gn'), (u'jerarca', 'gn'), (u'y', 'C'), (u'su', 'gn'),
(u'equipo', 'gn')]
```

para el primer caso se ve el problema de que a veces se incluye la coordinación dentro de un grupo nominal lo que induce a un reconocimiento demasiado extenso del asunto. El segundo y tercer caso inducen a pensar que es una buena idea agregar la regla <gn><C><gn> al módulo de reglas, aunque habría que comprobar si esta regla no extiende demasiado asuntos ya reconocidos correctamente. El caso 4 es un poco más complejo que el 2 y el 3, pues aparte de una coordinación, hay una preposición, donde es claro que la regla <gn><S><gn><C><gn> sería la adecuada.

En la tercera categoría aparecen casos como los ya mencionados en subsecciones anteriores.

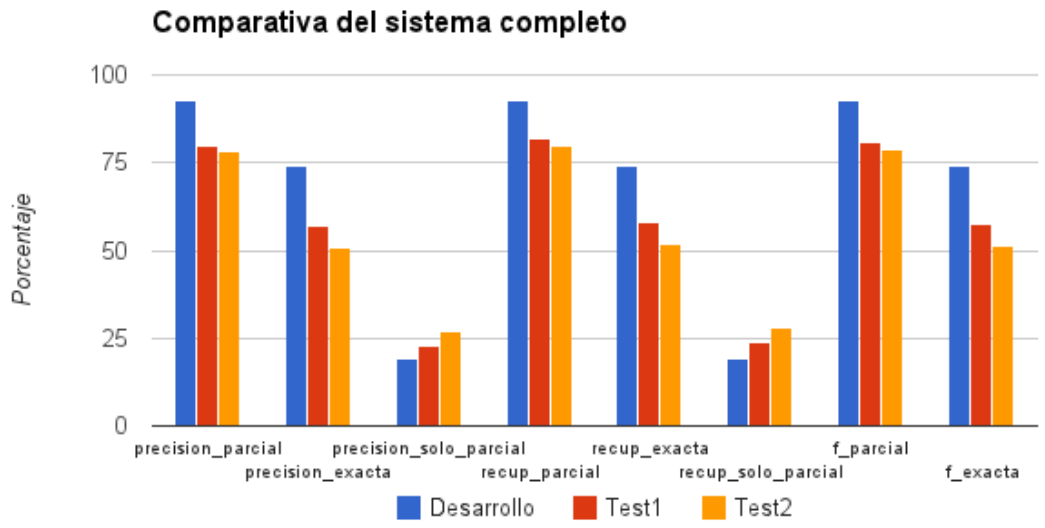
La cuarta categoría contiene casos parecidos a los analizados al comienzo de la sección 5.5.1, donde se descubre que la palabra “años” es relevante si se la incluye en la lista negra o no.

El caso de las enumeraciones (quinta categoría) es un tipo bastante concreto de problema:

```
[(u'las', 'gn'), (u'emisiones', 'gn'), (u',', 'gn'), (u'la', 'gn'), (u'calidad', 'gn'), (u'de', 'gn'), (u'aire', 'gn'), (u'y', 'C'), (u'del', 'gn'), (u'agua', 'gn'), (u'arrojan', 'V'), (u'parámetros', 'gn'), (u'buenos', 'gn'), (u'y', 'C'), (u'muy', 'R'), (u'buenos', 'A')]
```

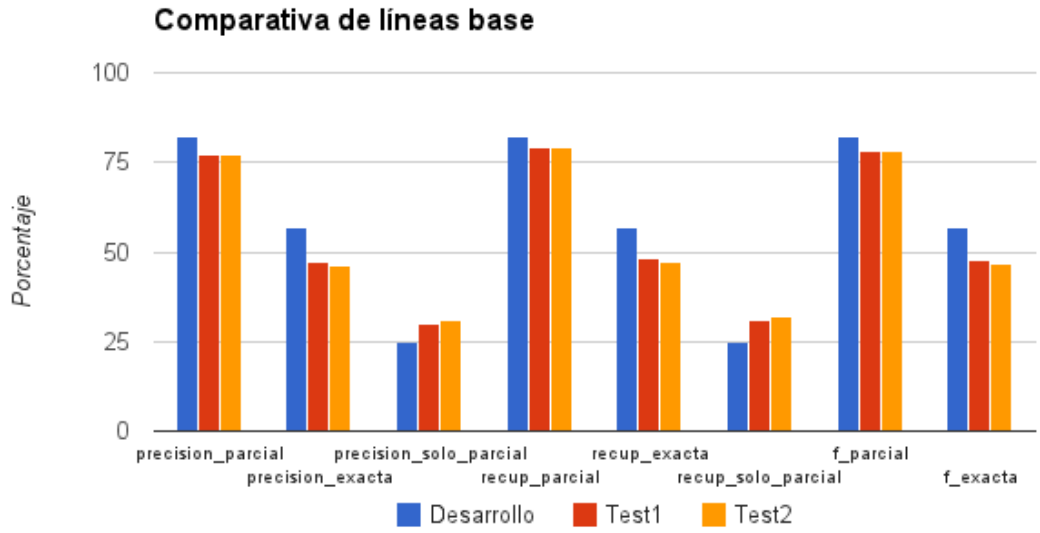
pues si bien se puede ver como un problema de reconocimiento de asunto con coordinación, sería de utilidad contar con un reconocedor de enumeraciones.

5.6 Resultados del sistema completo



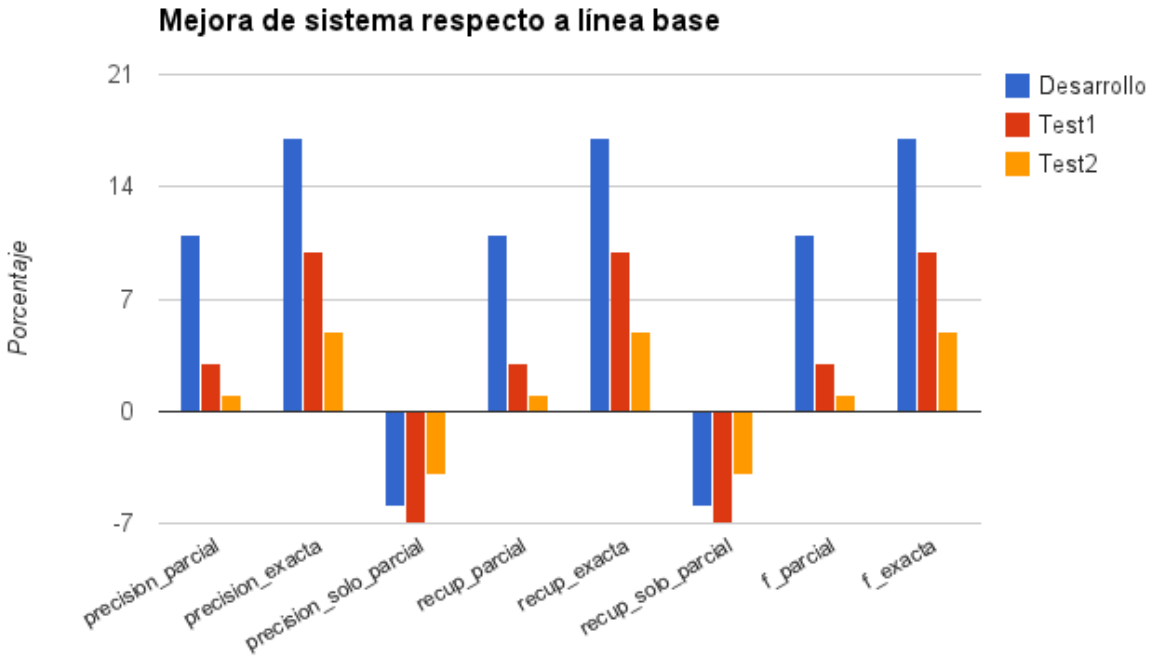
En esta gráfica es claro que en términos absolutos los datos de desarrollo de reglas dan mejores resultados que los datos de testeo, especialmente en la precisión/recuperación exacta donde la diferencia es de un 18%, aunque se debe tener presente que se parte de distintos niveles respecto a la línea base. Podríamos descartar que el primer subgrupo de testeo tenga demasiado ruido por pertenecer las opiniones a pocas noticias ya que los resultados son parecidos e incluso mejores que los del segundo grupo de testeo.

Si comparamos solo las líneas base:



se ve que es superior en los datos de desarrollo, lo que podría explicarse por un mayor cuidado al momento de anotar los datos de desarrollo o indicando un posible "sobreajuste" o sesgo al momento de anotar.

Finalmente queda analizar las mejoras en los distintos grupos respecto a la línea base:



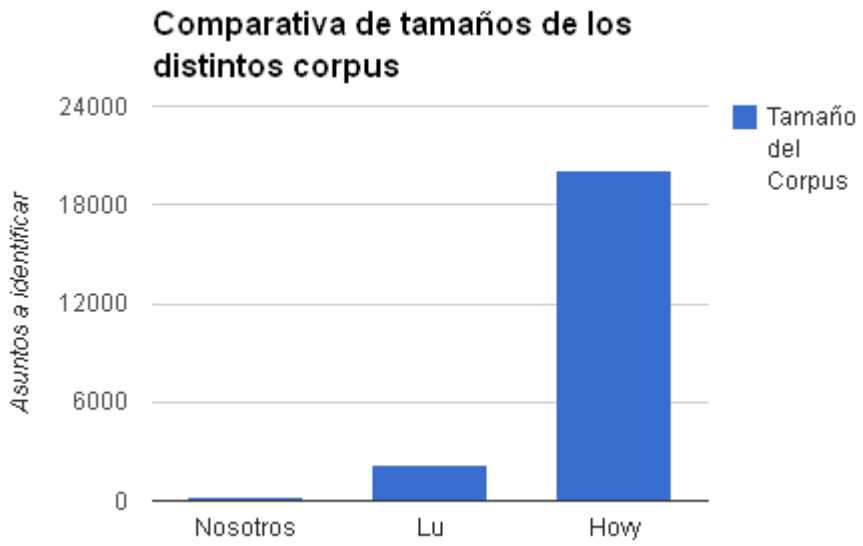
El gran tamaño de la línea de desarrollo indica también un posible sobreajuste de las reglas a los datos de desarrollo, en particular, la medida F exacta mejora un 17% para los datos de desarrollo, y un 10% y 5% para los datos de testeo, por lo que el sistema mejora los resultados para todos los subgrupos.

5.7 Comparación con otros trabajos

A continuación se presentan comparaciones con otros trabajos. Se debe tener en cuenta que cada autor tiene su propia definición de opinión y de asunto, que si bien son parecidas, no son lo mismo ya que el el procesamiento de datos no está estandarizado.

Trabajo	Línea Base	Sistema	Recursos	Método usado
Bin Lu, Identifying Opinion Holders and Targets with Dependency Parser in Chinese News Texts,2010	P: 11% R: 9% F: 9.9%	P: 29% R: 28% F: 28.5%	2174 oraciones	Parser de dependencias
Eduard Hovy, Soo-Min Kim, Extracting opinions, opinion holders, and topics expressed in online news media text, 2006	Conjunto1: P(verbos): 85.5% R(verbos): 18.5% F: 30.4% Conjunto2 (complicado) P: 12.5% R: 9.4% F: 10.7%	Conjunto1 (sencillo): P(verbos): 69,1% R(verbos): 67,5% F: 68.3% Conjunto2 (complicado): P: 64.7% R: 20.8% F: 31.5%	20133 oraciones	Etiquetado de roles semánticos (Semantic Role Labeling) Usando Framenet
Nosotros	Test1: P: 47% R: 48% F: 47,5% Test2: P: 46% R: 47% F: 46,5%	Test1: P: 57% R: 58% F: 57,5% Test2: P: 51% R: 52% F: 51,5%	192 opiniones	Reglas y dependencias

Es decir que en promedio tenemos un 54,5% de medida F para el sistema y un 47% para la línea base. Eso nos sitúa por encima de Hovy y de Lu, salvo para el conjunto “sencillo” de Hovy donde su sistema está más de 10 puntos porcentuales por encima del nuestro. Nuestra línea base es notablemente superior(17% - 35%) en todos los casos, por lo que podemos pensar que nuestro dominio es más fácil de identificar (noticias en general sobre política). También una diferencia importante es el tamaño de los corpus, ya que suponiendo que una oración de los otros trabajos equivale a una opinión nuestra, se ve que nuestro corpus es realmente pequeño, como muestra la siguiente figura:


Comparativa de tamaños de los distintos corpus
24000
18000
12000
6000
0
Asuntos a identificar
Nosotros
Lu
Howy
Tamaño del Corpus

6 Conclusiones

El resultado más visible y concreto de este trabajo es el sistema que se programó en Python para resolver el problema de anotar asuntos en mensajes en opiniones que se encuentran en textos extraídos de portales web de noticias en español.

En lo que respecta a la implementación, FreeLing resultó de ayuda para algunos pequeños avances pero en algunos textos mostró ser problemático por los errores en el análisis que produce para el nivel de especificidad que buscamos. Podrían incorporarse otras herramientas (como VISL) de apoyo para identificar dependencias. Esto agregaría la la dificultad de seleccionar cuándo utilizar cada herramienta en caso de no utilizar una única herramienta.

En lo que respecta a las técnicas se utilizó agrupamiento con reglas sencillas, lista negra, algoritmos para correcciones menores, información posicional y se incorporó la información de los árboles de dependencias de FreeLing conjuntamente con algunas reglas complejas.

Si pasamos a los recursos y sus implicancias, vemos como aporte que se deja un corpus con asuntos anotados, pues es una de las principales carencias en este tipo de proyecto la falta de recursos lingüísticos. Además, para poder buscar generalizar soluciones con cierta especificidad, tal vez se debería de contar con un conjunto de opiniones 10 o 100 veces más grande. Pues si tenemos en cuenta que el lenguaje es muy rico, podría darse que sean muchas las “clases de equivalencia” de los potenciales asuntos problemáticos y que unas 100 opiniones, en la práctica, nos estén dando un solo representante de cada clase. Por ejemplo el caso de las coordinaciones, que no aparecieron como problemáticas en la parte de desarrollo al identificar el asunto, pero si en la parte de testeo. Para dimensionar, sirve tener presente que en [Tink2007] se señala que para cubrir la gramática del español, fueron necesarias 4500 reglas. El asunto de un mensaje seguramente no sea algo tan complejo como la propia gramática del idioma español, pero tal vez sean necesarias cientos de reglas para llegar a una identificación de asuntos con altos niveles de reconocimiento.

Es importante ver que este es el único trabajo que se centra en el idioma español para el reconocimiento de asuntos con cierto nivel de elaboración.

En otro orden aportó a este trabajo el hecho de que los trabajos e informes del grupo de PLN estuvieran disponibles y accesibles. El acceder a material preexistente evita tener que explorar demasiado cuestiones de forma y permite centrarse en el problema a resolver, además del contenido en sí que aportan los demás trabajos.

Otra conclusión muy importante, es que el proyecto sirve para tener una aproximación a la investigación, que si bien en el imaginario de una persona puede esperarse como algo emocionante lleno de regocijo intelectual, hay una parte rutinaria que involucra mucho esfuerzo y corresponde a el trabajo de hormiga de ver ejemplos y más ejemplos. Esto último en este trabajo se manifestó tanto en la anotación del corpus, como en los cientos de comparaciones y dilucidaciones que deben hacerse para probar el sistema y vincular qué cosas se están reconociendo y por qué.

Como se explicó en la sección de “dificultades encontradas” la visualización de resultados debe ser tenida en cuenta para la planificación de proyectos que requieran de

un análisis cualitativo de ejemplos intensivo (como este), puesto que no encontramos mención en la literatura, y resultó un elemento relevante.

Debe tenerse en cuenta si el abordaje que se está haciendo es adecuado o no, porque el yerro más grande puede estar en la equivocación epistemológica. ¿Puede el asunto de una opinión identificarse correctamente sin una base teórica más sólida, desde la lingüística y la psicolingüística? ¿Puede captarse el asunto automáticamente si no se entiende con mayor claridad cómo es que hace un humano para identificar un asunto?, ¿Pueden las reglas ser pertinentes sin un enfoque sobre el discurso como el planteado en [Renk2004]? Pues algunas de las técnicas usadas de alguna manera son la programación de las heurísticas que tienen las personas para identificar asuntos. Por ejemplo la identificación de grupos nominales del comienzo de la oración como objeto probable de asunto de un mensaje.

Resultó un reto interesante el intentar empujar las fronteras del conocimiento, que resultaron ser algo inamovibles para nuestro caso. Son interesantes los proyectos de grado de la carrera pues permiten construir sobre el trabajo de los investigadores del área para intentar estar a la par y superar las investigaciones existentes en el mundo. Puesto que los resultados obtenidos (medida F) son superiores en general a otros trabajos, no parece un mal camino el intentar las mismas medidas con mayor profundidad. Es decir, el proyecto puede plantearse de vuelta con los mismos objetivos, pero yendo más a fondo en temas claves, como puede ser los recursos lingüísticos, así como en las técnicas vinculadas a la identificación de asuntos, como en el análisis de tipos de fenómenos lingüísticos a atacar para reconocer la mayor cantidad de asuntos posibles con una técnica dada. Los criterios para anotar asuntos manualmente podrían conservarse. Incluso podría desarrollarse una herramienta visual que permita comparar con facilidad los elementos identificados por el sistema contra el gold standard. Por ejemplo indicando tanto las etiquetas gramaticales como el árbol de dependencia asociado.

Desde el punto de vista de la carrera si bien la Facultad de Ingeniería de la Universidad de la República templa el espíritu por su alto nivel de exigencia ya desde los más tiernos comienzos para todo estudiante, es de tener presente lo que señala Vaz Ferreira [VazF1920] en su primer consejo como deber de cultura de los estudiantes, pues todavía al final de la carrera quedan hábitos y habilidades por aprender imposibles de lograr con cursos semestralizados, donde el ritmo vertiginoso poco espacio deja para las reflexiones que requieren tiempo. En pocas palabras: Vaz Ferreira se pronuncia a favor del estudio en profundidad, que es lo que permite hacer el proyecto de grado.

O en sus propias palabras:

> *Todo estudiante, ya en su bachillerato, en los estudios preparatorios, debe profundizar algunos temas; poco importa cuáles: esto realmente es secundario; que se tome un punto de historia o de literatura o de filosofía o de ciencia; que se estudie a Artigas, o el silogismo, o las costumbres de los diversos pueblos, o la teoría atómica o la constitución física del Sol, es secundario: lo fundamental, son los hábitos que se adquieren profundizando un punto cualquiera.*
>
> *Carlos Vaz Ferreira, 1920*

7 Bibliografía y Referencias

[AC2009] Jorge Aguirre, Raúl Carnota, *Historia de la Informática en Latinoamérica y el Caribe: Investigaciones y testimonios*, Argentina, Universidad Nacional de Río Cuarto, 2009.

[AHFB2005] James Allan, Stephen Harding, David Fisher, Alvaro Bolivar, Sergio Guzman-Lara, Peter Amstutz, *Taking Topic Detection From Evaluation to Practice*, System Sciences, 2005, Proceedings of the 38th Annual Hawaii International Conference on. IEEE, HICSS'05, 2005

[ACM2005] Jordi Atserias, Eli Comelles, Aingeru Mayor, *TXALA un analizador libre de dependencias para el castellano*, Procesamiento de lenguaje Natural, Vol. 35, pp. 455-456, 2005

[ARZ2010] Fernando Acerenza, Macarena Rabosto, Magdalena Zubizarreta, *Resolución de correferencias en expresiones de opinión*, Informe de proyecto de grado, Facultad de Ingeniería, Universidad de la República, Uruguay, 2010

[BYTH2004] Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, Dan Jurafsky, *Automatic Extraction of Opinion Propositions and their Holders,* en 2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text, p. 2224, 2004

[BKL2009] Steve Bird, Ewan Klein, Edward Loper, *Natural Language Processing With Python(disponible online)*, O'reilly, 2009

[Desc2010] Alan Descoins, *Reconocimiento automático de eventos en textos,* Informe de proyecto de grado, Facultad de Ingeniería, Universidad de la República, Uruguay, 2010

[Dint2002] Felipe Dintel, *Cómo se elabora un texto*, Buenos Aires, Alba Editorial, 2002

[DiTu1997] Ángela Di Tullio, *Manual de gramática del español,* Segunda Edición, Buenos Aires, EDICAL S.A.,1997

[Gimp2006] Kevin Gimpel, *Modeling Topics*, Information Retrieval, Vol. 5, 2006

[HK2006] Eduard Hovy, Soo-Min Kim, *Extracting opinions, opinion holders, and topics expressed in online news media text, en* Proceedings of the Workshop on Sentiment and Subjectivity in Text, SST '06, pp. 1-8, Stroudsburg, PA, USA, Association for Computational Linguistics, 2006

[Jack1998] Ray Jackendoff, *La conciencia y la mente computacional,* Madrid, Visor, 1998

[JM2008] Daniel Jurafsky, James H. Martin, *An introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition,* Upper Saddle River, Prentice Hall, 2008

[KKM2008] Youngho Kim, Seogchan Kim, Sung-Hyon Myaeng, *Extracting Topic-related Opinions and their Targets in NTCIR-7,* en Proceedings of NTCIR-7 Workshop, Tokyo, Japan, 2008

[Liu2010] Bing Liu, *Sentiment Analysis and Subjectivity, en Handbook of Natural Language Processing, Second Edition (Indurkhya y Damerau)*, Taylor and Francis Group, Boca, 2010

[Llor2009] Elena Lloret, *Topic Detection and Segmentation in Automatic Text Summarization,* http://www.dlsi.ua.es/~elloret/publications/SumTopics.pdf, 2009

[Lu2010] Bin Lu, *Identifying Opinion Holders and Targets with Dependency Parser in Chinese News Texts,* en Proceeding HLT-SRWS '10 Proceedings of the NAACL HLT 2010 Student Research Workshop, pp. 46-51, Stroudsburg, PA, USA, Association for Computational Linguistics, 2010

[MLWS2007] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai, *Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs*, en Proceedings of the 16th international conference on World Wide Web, pp. 171-180, New York, NY, USA, 2007

[Mins2010] Marvin Minsky, *La máquina de las emociones,* Debate, 2010

[NGB2011] *Nueva gramática básica de la lengua española. 1Ed,* Asociación de academias de la lengua española, Argentina, 2011.

[Orla2009] Virginia Orlando (coord.), Yamila Montenegro, Ana Clara Polakof, Carlos Hipogrosso, Carmen Lepre, Merces Costa, *Manual de gramática del español,* Montevideo, Departamento de Publicaciones, Unidad de Comunicación de la Universidad de la República, 2009

[Padr2011] Lluis Padró, *Analizadores multilingües en FreeLing,* Linguamática, Vol. 3, num 1, pp. 13-20, 2011

[PS2012] Lluís Padró, Evgeny Stanilovsky*, FreeLing 3.0: Towards Wider Multilinguality,* en Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC'12, Estanbul, Turquía, European Language Resources Association*,* 2012

[PL2008] Bo Pang, Lillian Lee, *Opinion Mining and Sentiment Analysis,* Foundations and Trends in Information Retrieval 2(1-2), Vol. 2, pp. 1-135, 2008

[Renk2004] Jan Renkema, *Introduction to discourse studies,* John Benjamins Publishing, 2004

[Rosá2011] Aiala Rosá, *Identificación de opiniones de diferentes fuentes en textos en español, Ph.D. Dissertation,* Programa de Desarrollo de las Ciencias Básicas - Universidad de la República, Montevideo, Uruguay - École Doctorale Connaissance, Langage, Modélisation Université Paris Ouest, Nanterre, La Défense, 2011

[RSW2008] Josef Ruppenhofer, Swapna Somasundaran, Janyce Wiebe, *Finding the sources and Targets of Subjective Expressions*, en Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC'08, Marrakech, Marruecos, 2008

[RN2004] Stuart Russell, Peter Norvig, *Inteligencia Artificial. Un enfoque moderno. 2da Ed.,* México, Prentice Hall, 2004

[RWM2010] Aiala Rosá, Dina Wonsever, Jean-Luc Minel, *Opinion Identification in Spanish Texts,* en Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, Los Angeles, California, Association for Computational Linguistics, 2010

[SW2009] Swapna Somasundaran, Janyce Wiebe, *Recognizing Stances in Ideological On-Line Debates,* en CAAGET'10, NAACL-HLT'10, pp. 116-124, 2009

[SC2008] Veselin Stoyanov, Claire Cardie, *Annotating Topics of Opinions,* en Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC'08, Marrakech, Marruecos, 2008

[SC2008b] Veselin Stoyanov, Claire Cardie, *Topic Identification for fine-grained opinion analysis*, en Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1, COLING '08, pp. 817-824, Stroudsburg, PA, USA, Association for Computational Linguistics, 2008

[Tink2007] Nevena Tinkova Tincheva, *A State-of-the-Art review on Automatic Parsing of Spanish*, Grial Research Report Nº 1/2007, University of Barcelona, Department of General Linguistics, 2007

[VAG2013] Daniel Vilares, Miguel A. Alonso, Carlos Gómez-Rodríguez, *Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias*, GRIAL RESEARCH REPORT Nº 1/2007, Procesamiento del Lenguaje Natural. N. 50, pp. 13-20, 2013

[VazF1920] Carlos Vaz Ferreira, *Moral para intelectuales*, Montevideo, Arca, 1969

[VC2012] G. Vinodhini, R. M. Chandrasekaran, *Sentiment Analysis and Opinion Mining: A Survey,* International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, issue 6, pp. 282-292, 2012

[WWC2005] Janyce Wiebe, Theresa Wilson, Claire Cardie, *Annotating expressions of opinions and emotions in language,* Language Resource and Evaluation, Vol. 29, issue 2-3, pp. 165-210, Kluwer Academic Publishers, 2005

Vínculos usados frecuentemente en el desarrollo:

Etiquetas EAGLES: http://www.lsi.upc.edu/~nlp/tools/parole-sp.html
FreeLing: http://nlp.lsi.upc.edu/freeling/index.php
VISL: http://beta.visl.sdu.dk/visl/es/parsing/automatic/trees.php

ANEXOS

ANEXO A: Glosario

aposición(RAE):

1. Construcción en la que un sustantivo o un grupo nominal sigue inmediatamente, con autonomía tonal, a otro elemento de esta misma clase para explicar algo relativo a él.
2. Construcción de dos elementos nominales unidos, el segundo de los cuales especifica al primero; p. ej., mi amigo el tendero; el rey Felipe II. Por ext., se aplica a construcciones del tipo La calle de Goya o el tonto de Rigoberto.

aprendizaje automático: Es una rama de la Inteligencia Artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender en base a datos. Es, por lo tanto, un proceso de inducción del conocimiento.

chunker: Herramienta que permite el análisis de una oración para separarla en constituyentes sintácticos.

complemento circunstancial: Complemento que expresa circunstancias de la acción verbal, como lugar, tiempo, modo, instrumento, etc.

concordancia(gramatical): Cómo los distintos elementos de una frase deben indicar su acuerdo mediante la uniformidad de su flexión, por ejemplo, en género y número.

constituyente sintáctico: Es una palabra, o secuencia de palabras, que funciona en conjunto como una unidad dentro de la estructura jerárquica de una oración.

corpus: Colección de material lingüístico.

deverbal (nombre): Nombre derivado de un verbo.

dialéctica: Técnica de la conversación. Es una rama de la filosofía cuyo ámbito y alcance ha variado significativamente a lo largo de la historia.

gold standard: Resultado predefinido por un evaluador que se entiende como “lo correcto”.

lema: Forma básica de una palabra que busca eliminar toda inflexión.

morfosintaxis: La morfosintaxis es una parte de la lingüística que estudia, concretamente, el conjunto de las reglas y los elementos que hacen de la oración un elemento con sentido y carente de ambigüedad. Para ello, el análisis morfosintáctico, entre otras cosas se ocupa de marcar las relaciones gramaticales que se dan dentro de una oración, las relaciones de concordancia, la estructura jerárquica de los principales constituyentes sintácticos.

nominalización: La sustantivación morfológica es una derivación léxica consecuencia de la formación de un sustantivo partiendo de otro tipo de palabra. La sustantivación sintáctica es el uso de una palabra que morfológicamente no es un sustantivo pero puede ser el núcleo de un sintagma nominal.

objeto directo:

1. El objeto directo es la parte de la oración que expresa la persona, animal o cosa sobre la cual recae directamente la acción del verbo.
2. (RAE): Nombre, pronombre, sintagma o proposición en función nominal, que completa el significado de un verbo transitivo.

preposición: Las preposiciones constituyen una clase cerrada de palabras, normalmente átonas y dotadas de valor relacional, que introducen un complemento que se denomina *término* con el que forman grupo sintáctico y al que pueden caracterizar semántica y sintácticamente.

preposición(RAE): Palabra invariable que introduce elementos nominales u oraciones subordinadas sustantivas haciéndolos depender de alguna palabra anterior. Varias de ellas coinciden en su forma con prefijos.

pronombre(personal): Designa a un participante en el discurso, poseen flexión de persona, además que puede tener otros rasgos gramaticales, género, número, caso y reflexividad.

ANEXO B: Contenido de los mensajes de las opiniones analizadas del corpus

A modo de ejemplo se enumeran 63 mensajes de distintas opiniones, no se muestra la opinión completa ni su contexto.

Nº	Texto del mensaje
0	"No es que haga desaparecer la común , sólo que aparece más la A en los puestos centinela"
1	que la pandemia es " leve a moderada"
2	que el 99 % de los casos de gripe en Uruguay corresponden al virus H1N1 , que " desplazó " a la gripe estacional
3	que "no encontramos a la gripe estacional" , atribuyendo a esto la "vigilancia intensa" de este nuevo virus así como la alta vacunación registrada contra la estacional
4	que las cifras manejadas por Basso se refieren a "los puestos centinelas del MSP" en distintas instituciones médicas
5	"Nosotros estudiamos, a través del Laboratorio de Higiene Pública, la vigilancia epidemiológica. Ese laboratorio no hace un diagnóstico para hacer tratamientos, hace diagnósticos para conocer la circulación de los virus dentro del país. Para eso todos los años, todos los inviernos, tiene puestos centinelas donde concurren muchos casos con cuadros respiratorios, esos estudios se envían al CDC de Atlanta y con ellos y los de otros países se elabora la vacuna de la gripe estacional para el año siguiente"
6	Hasta el momento el virus H1N1 tiene una predominancia mayor que la de los demás virus en esos estudios específicos
7	que la cifra del 99% refiere a esos casos
8	que "todavía no terminó el invierno, el invierno sigue hasta y los primeros días de setiembre"
9	que "los meses se analizan en conjunto" por lo cual las cifras actuales "no dan lugar a ninguna conclusión"
10	que existe la posibilidad de que "el ciudadano, que ahora tiene menos gripe, baje las manos en las medidas de prevención y no queremos eso"
11	que la prevención hasta el momento sirvió a la contención de la pandemia en el país
12	"No es que haga desaparecer la común , existe igual pero aparece más la A”
13	"Esas diferencias pueden ser fruto del legado evolucionario , es decir de nuestro pasado de cazadores-recolectores"
14	que la empresa de limpieza de la esposa de Eleuterio Fernández Huidobro, Alejandra De Melo, haya sobrefacturado el servicio que le brinda al Hospital
15	que "no es un tema de sobrefacturación. Lo que se hizo fue un cálculo estimativo de horarios y se les fue pagando en base a cálculos estimativos de horarios. A medida que se fueron ajustando los horarios y se fueron viendo las reales necesidades del hospital, se

	fueron ajustando los sistemas de pago. Nada más que eso"
16	que "la empresa reconoció que hubo un cálculo entre horas estimadas y horas reales y eso fue lo que paso. Nosotros, primero empezamos a controla a nuestro personal. Ese era un problema. No había control de personal. Cuando logramos controlar eso, logramos controlar adecuadamente a las empresa contratadas"
17	que hubo un acuerdo bilateral en hacer esa devolución en horas de trabajo, “porque como nosotros comenzamos a abrir más servicios y más salas, empezamos a requerir más horas de limpieza. Preferimos que nos devuelvan con más horas de trabajo"
18	que su detención fue tan absurda , que pensó que era una “cámara oculta para Marcelo Tinelli"
19	que los tres demandantes "fueron incorrectamente procesados con prisión en , en el marco del más mediático procedimiento judicial de los últimos tiempos"
20	que previo a su procesamiento, "sabía que había una conspiración clara contra nosotros. Cuando me secuestro personal de inteligencia y me llevan al juzgado, era tan absurdo todo, que nunca pensé que sería procesado con prisión. Pensaba que iba a se una más de la tantas que nos intentaron ensuciar"
21	que mantuvo una reunión con el subsecretario de dicho organismo financiero
22	"tuvimos una conversación sincera donde no faltó la autocrítica del Fondo y las fallas que tuvieron en el monitoreo de la crisis reciente. Conversamos acerca de la situación de la economía mundial y de las proyecciones vinculadsa al trabajo que está haciend el FMI"
23	que dos hombres intentaban ingresar por la fuerza a una vivienda en Cambridge , Boston (noreste)
24	que la policía había actuado "estúpidamente" por arrestar a su amigo, lo que elevó la polémica a nivel nacional
25	que Joan comenzó a salir de compras con frecuencia, luego de tener algunos altercados con sus vecinos
26	algunos jóvenes "hacían ruido constantemente" cerca de la vivienda de la anciana, hecho que la habría impulsado a pasar horas y horas en centros comerciales, para no permanecer en su hogar
27	que vio por última vez a su amiga en la navidad última
28	que dos productores le pidieron dinero a cambio de evitar que un informe saliera al aire
29	que si les entregaba 500.000 pesos todo quedaba archivado
30	"Botnia en Uruguay está teniendo un comportamiento excelente"
31	que el control se realiza sobre más de 30 parámetros en efluentes líquidos, más de 10 en emisiones gaseosas y más de 15 a nivel de residuos líquidos
32	que desde la instalación de la pastera hubo una mejora en algunos aspectos medioambientales del departamento
33	"Nos sentimos satisfechos con los resultados obtenidos, satisfechos porque tenemos la evidencia material de que la planta de Botnia no produce ningún efecto engañoso desde el punto de vista ambiental y esta va a ser la última reunión de monitoreo que se celebre

	antes de las audiencias ante La Corte Internacional de Justicia que va a tener lugar en La Haya"
34	que "no hubo un pronunciamiento oficial" para dejar en la libertad de acción a los votantes del PN ante el plebiscito por la anulación de la ley de caducidad
35	"Yo mismo cuando voté en contra reclamé libertad de acción"
36	que ésta "es una cuestión de conciencia"
37	que en el plebiscito de 1986 "algunos legisladores del PN votaron a favor y otros en contra" de anular la ley de caducidad
38	que hay que respetar la diversidad de opinión que existe a la interna del partido sobre el asunto
39	"Yo mismo cuando voté en contra y junté firmas para el plebiscito , reclamé que haya libertad de acción"
40	que haya libertad de acción"
41	que en ese momento contrajo "un compromiso público" pues "como legislador tenía una postura tomada" que fue "la misma como ciudadano"
42	Más de 2.0 personas
43	Unas treinta personas murieron por la represión a las manifestaciones que estallaron tras la elección
44	Varias personas fueron detenidas
45	que Musavi logró salir de su coche y emprender el camino hacia la tumba de Neda Agha Soltan , la joven que murió baleada el pasado , convirtiendo se en símbolo de las protestas contra el resultado de las elecciones
46	"no fue autorizado a recitar los versos del Corán que se dicen en estas ocasiones e inmediatamente fue rodeado por la policía antidisturbios que lo llevó hasta su coche"
47	que aun 250 personas permanecen detenidas , entre ellas 50 personalidades políticas
48	que el PBI se contraerá un 4,8 % en 2009 y un 0,3 % el año próximo
49	que la Eurozona registrará una "modesta recuperación" económica durante el primer semestre de 2010, pero "lenta" y rodeada de "incertidumbres" por lo que cerrará el año con una contracción del 0,3%
50	que el Producto Bruto Interno (PBI) de la zona euro, integrada por 16 países, se contraerá un 4,8% en 2009 y un 0,3% el año próximo
51	"La Eurozona está en recesión, con fuertes señales de mejora que todavía deben transformar se en recuperación"
52	que este panorama "permanecerá rodeado de incertidumbres importantes" y que "la recuperación será lenta y sujeta a riesgos considerables"
53	Este fenómeno podría dar se
54	que lo recaudado será destinado al hospital
55	que esta experiencia piloto es "muy enriquecedora", no solo desde el punto de vista de la educación cívica, sino como parte del fortalecimiento de la democracia

56	Las contraindicaciones son superiores a los beneficios
57	que la actual política de prescribir Tamiflu para una enfermedad relativamente benigna es una "estrategia inadecuada"
58	que el Tamiflu puede causar vómitos en algunos niños y que puede provocar les deshidratación y complicaciones
59	que los niños que de forma preventiva recibieron Tamiflu tuvieron efectos secundarios como náuseas y pesadillas
60	que más de la mitad de 248 jovencitos que ingirieron Tamiflu, luego de que uno de sus compañeritos contrajera la gripe porcina, tuvieron efectos secundarios como náuseas, insomnios y pesadillas
61	"Parecía que una flagrante ola de flores hubiese invadido todo para que yo no pudiese dar un solo paso sin estar sobre ella"
62	que "la sencillez de Juana puede ser aparente , pero hay muchas capas de lectura en las que si se va ahondando se descubren muchas sombras y abismos"
63	que la obra de Ibarbourou " no conoció límites de edad ni de intereses "

ANEXO C: Algunas noticias completas del corpus

Extracto del corpus donde se muestran 18 opiniones de la siguiente noticia:

Noticia - Los casos en su sitio. 30.07.2009 13:42.

"No es que haga desaparecer la común, sólo que aparece más la A en los puestos centinela", dijo a Montevideo Portal María Julia Muñoz, aclarando datos que señalan que el 99 % de los casos corresponden al nuevo virus.

La ministra destacó la prevención y reiteró que la pandemia es "leve a moderada".

La ministra de Salud Pública realizó algunas precisiones sobre la información que recorrió este jueves la prensa local e internacional, señalando que el 99 % de los casos de gripe en Uruguay corresponden al virus H1N1, que "desplazó" a la gripe estacional.

El director de Salud, Jorge Basso, había mencionado en la jornada del miércoles que "no encontramos a la gripe estacional", atribuyendo a esto la "vigilancia intensa" de este nuevo virus así como la alta vacunación registrada contra la estacional.

Muñoz precisó que las cifras manejadas por Basso se refieren a "los puestos centinelas del MSP" en distintas instituciones médicas y recalcó que "no quiere decir que no haya" casos de gripe común, porque, recordó, "la gripe no es una enfermedad de denuncia obligatoria, ni la común ni la H1N1".

"Nosotros estudiamos, a través del Laboratorio de Higiene Pública, la vigilancia epidemiológica. Ese laboratorio no hace un diagnóstico para hacer tratamientos, hace diagnósticos para conocer la circulación de los virus dentro del país. Para eso todos los años, todos los inviernos, tiene puestos centinelas donde concurren muchos casos con cuadros respiratorios, esos estudios se envían al CDC de Atlanta y con ellos y los de otros países se elabora la vacuna de la gripe estacional para el año siguiente",

detalló.

Hasta el momento el virus H1N1 tiene una predominancia mayor que la de los demás virus en esos estudios específicos, precisó la ministra, que dijo que la cifra del 99% refiere a esos casos. Asimismo, Muñoz dijo que "todavía no terminó el invierno, el invierno sigue hasta agosto y los primeros días de setiembre" y agregó que "los meses se analizan en conjunto" por lo cual las cifras actuales "no dan lugar a ninguna conclusión".

La ministra señaló que existe la posibilidad de que "el ciudadano, que ahora tiene menos gripe, baje las manos en las medidas de prevención y no queremos eso", destacando que la prevención hasta el momento sirvió a la contención de la pandemia en el país.

Las medidas, según Muñoz, "tuvieron una buena respuesta de la población. Sin duda la gente ha consultado más a domicilio, ha tomado precauciones correctas, se lavan las manos, tienen alcohol en gel en la cartera, los ómnibus limpian las unidades en varias ocasiones, se ventila el transporte colectivo, se han minimizado los riesgos sin necesidad de medidas extremas", apuntó.

"No es que haga desaparecer la común, existe igual pero aparece más la A", reiteró la jerarca, que recordó que según la OMS "esta gripe ha sido leve a moderada, con muy pocos caso mortales".

Opinión	Contenido
OP 1.1	<opinion> <mensaje>**"No es que haga desaparecer** <asunto>**la común**</asunto>**, sólo que aparece más** <asunto>**la A**</asunto> **en los puestos centinela"**</mensaje>**,** <predicado>**dijo**</predicado> **a Montevideo Portal** <fuente> **María Julia Muñoz**</fuente> </opinion>

	, aclarando datos que señalan que el 99 % de los casos corresponden al nuevo virus.
OP 1.2	<opinion> <fuente>**La ministra**</fuente> <predicado>**destacó**</predicado> <asunto>**la prevención**</asunto> </opinion> **y**
OP 1.3	<opinion> <predicado>**reiteró**</predicado> <mensaje>**que** <asunto>**la pandemia**</asunto> **es "leve a moderada"**</mensaje> </opinion>**.**
OP 1.4	<opinion> <fuente>**La ministra de Salud Pública**</fuente> **realizó algunas** <predicado>**precisiones**</predicado> **sobre** <asunto>**la información que recorrió este jueves la prensa local e internacional**</asunto> </opinion>**,**
OP 1.5	<opinion> <predicado>**señalando**</predicado> <mensaje>**que el 99 % de** <asunto>**los casos de gripe en Uruguay**</asunto> **corresponden al virus H1N1, que "desplazó" a la gripe estacional**</mensaje> </opinion>**.**
OP 1.6	<opinion> <fuente>**El director de Salud, Jorge Basso**</fuente> **,** <predicado>**había mencionado**</predicado> **en la jornada del miércoles** <mensaje>**que "no encontramos a** <asunto>**la gripe estacional**</asunto>**", atribuyendo a esto la "vigilancia intensa" de este nuevo virus así como la alta vacunación registrada contra la estacional**</mensaje> </opinion>
OP 1.7	<opinion> <fuente>**Muñoz**</fuente> <predicado>**precisó**</predicado> <mensaje>**que** <asunto>**las cifras manejadas por Basso**</asunto> **se refieren a "los puestos centinelas del MSP" en distintas instituciones médicas**</mensaje> </opinion> **y**
OP 1.8	<opinion> <predicado>**recalcó**</predicado> <mensaje>**que "no quiere decir que no haya" casos de gripe común, porque,** <predicado>**recordó** </predicado>**, "**<asunto>**la gripe**</asunto> **no es una enfermedad de denuncia obligatoria, ni la común ni la H1N1"**</mensaje> </opinion> **.**
OP 1.9	<opinion> <mensaje>**" Nosotros estudiamos, a través del Laboratorio de Higiene Pública,** <asunto>**la vigilancia epidemiológica**</asunto> **. Ese laboratorio no hace un diagnóstico para hacer tratamientos, hace diagnósticos para conocer la circulación de los virus dentro del país . Para eso todos los años, todos los inviernos, tiene puestos centinelas donde concurren muchos casos con cuadros respiratorios, esos estudios se envían al CDC de Atlanta y con ellos y los de otros países se elabora la vacuna de la gripe estacional para el año siguiente "** </mensaje>**,** <predicado>**detalló**</predicado> </opinion>**.**
OP 1.10	<opinion> <mensaje>**Hasta el momento** <asunto>**el virus H1N1**</asunto> **tiene una predominancia mayor que la de los demás virus en esos estudios específicos**</mensaje>**,**

	<predicado>**precisó**</predicado> <fuente>**la ministra**</fuente> </opinion> **, que**
OP 1.11	<opinion> <predicado>**dijo**</predicado> <mensaje>**que** <asunto>**la cifra del 99 %** </asunto> **refiere a esos casos**</mensaje> </opinion>**. Asimismo,**
OP 1.12	<opinion> <fuente>**Muñoz**</fuente> <predicado>**dijo**</predicado> <mensaje>**que "todavía no terminó** <asunto>**el invierno**</asunto>**, el invierno sigue hasta agosto y los primeros días de setiembre"**</mensaje> </opinion> **y**
OP 1.13	<opinion> <predicado>**agregó**</predicado> <mensaje>**que "los meses se analizan en conjunto" por lo cual** <asunto>**las cifras actuales**</asunto> **"no dan lugar a ninguna conclusión"**</mensaje> </opinion>**.**
OP 1.14	<opinion> <fuente>**La ministra**</fuente> <predicado>**señaló**</predicado> <mensaje>**que existe** <asunto>**la posibilidad de que "el ciudadano, que ahora tiene menos gripe, baje las manos en las medidas de prevención**</asunto> **y** <predicado>**no queremos**</predicado> **eso"**</mensaje> </opinion>**,**
OP 1.15	<opinion> <predicado>**destacando**</predicado> <mensaje>**que** <asunto>**la prevención**</asunto> **hasta el momento sirvió a** <asunto>**la contención de la pandemia en el país**</asunto></mensaje> </opinion>**.**
OP 1.16	<opinion> <mensaje><asunto>**Las medidas**</asunto></mensaje>**,** **según** <fuente>**Muñoz,**</fuente> <mensaje>**" tuvieron una** <opinion> <predicado>**buena respuesta**</predicado> **de** <fuente>**la población**</fuente> </opinion>**.** **Sin duda la gente ha consultado más a domicilio, ha tomado precauciones correctas, se lavan las manos, tienen alcohol en gel en la cartera, los ómnibus limpian las unidades en varias ocasiones, se ventila el transporte colectivo, se han minimizado los riesgos sin necesidad de medidas extremas "** </mensaje>**,** <predicado>**apuntó**</predicado> </opinion>**.**
OP 1.17	<opinion> <predicado>**buena respuesta**</predicado> **de** <fuente>**la población**</fuente> </opinion>
OP 1.18	<opinion> <mensaje>**"No es que haga desaparecer** <asunto>**la común**</asunto>**, existe igual pero aparece más** <asunto>**la A**</asunto> **"**</mensaje>**,**

	<predicado>**reiteró**</predicado> <fuente>**la jerarca**</fuente> </opinion>

ANEXO D: Trabajos relacionados

A continuación se presenta el análisis de artículos de la literatura con observaciones sobre lo que puede aportar a nuestro trabajo o sobre los distintos enfoques que utilizan.

Introducción al estudio del discurso. Jan Renkema [Renk2004]

Es un libro de texto en el cual el tema en cuestión es el discurso en sí. Los temas de las opiniones constituyen un apartado donde se hila fino, por ejemplo en la oración *¿has escuchado esa extraña historia acerca del borracho que decidió jugar al barbero y le cortó la oreja a su amigo?* Se destaca que hay una opinión embebida, marcada por el adjetivo "extraña", pues indica que la persona no solo está haciendo una pregunta, sino que opina que la historia en cuestión es extraña. El texto plantea muchos otros análisis interesantes, que pueden servir para profundizar este trabajo, pero por el grado de precisión en las "sutilezas" que abarca, se decidió que excedían el alcance de este trabajo como para poder hacer reglas que capturen los criterios del texto.

Topic detection and segmengtation in automatic text summarization. Elena Lloret [Llor2009]

Elena hace una observación interesante sobre los temas, referenciando a [Renk2004] en donde indica ciertos "piques" como :

1. Tendiente a ser *definite*[20] antes que *indefinite*
2. El primer pronombre antes que el nombre
3. El primer sujeto antes que el objeto.

Su análisis se centra más que nada en los pasos previos a poder resumir documentos completos. Por eso se centra en la "progresión temática" y en métodos más macro como el TextTilling, cuya idea es segmentar el texto (sin etiquetar) la estructura de subtemas usando la repetición de términos (medida *tf-idf*)

Extracting Topic-related Opinions and their Target in NTCIR-7. Youngho Kim, Seongchan Kim, Sung-Hyon Myaeng [KKM2008]

Youngho Kim y otros extraen el tema a nivel de sentencia mediante 3 mecanismos: modelos probabilísticos con el insumo de palabras clave, basado en modelos de lenguaje con ayuda de *web-snippets* para atacar el "*sparseness*" y luego usan aprendizaje estadístico que explota la sintaxis ("syntactic path and dependency") además de las entidades con nombre. Los autores basan su motivación en el crecimiento de blogs y noticias online. Una de sus observaciones principales es que la mayoría de las opiniones que hay en la web son poco útiles y por lo tanto es importante centrarse en las opiniones interesantes. Señala (en 2008) que identificar asuntos es una tarea reciente.

20 No encontramos una traducción precisa al español. Definite se refiere a palabras que sean particulares y no amplias y generales.

Identifying Opinion Holders and Targets with Dependency Parser in Chinese News Texts. Bin Lu [Lu2010]

Lu hace su análisis para noticias en chino. Parte de que una opinión incorpora fuente (Opinion Holder) y asunto (Target), señala que la fuente es una entidad y la definición de asunto es: “sobre lo que trata la opinión”, tomándola de [KH2006].

Remarca la importancia de identificar la fuente, porque ayuda a detectar el asunto, simplemente porque en caso de marcar un asunto, que se solape con la fuente (holder conflict) el asunto se descarta, ya que usualmente la fuente está correctamente identificada. Esto en nuestro trabajo se incorpora por la propia construcción de los insumos, donde el mensaje es distinto de la fuente y naturalmente la fuente nunca está contenida dentro del mensaje. Igualmente la fuente está separada del asunto de una opinión.

Indica que para “product review” identificar a la fuente es más fácil, por ser usualmente el autor de la review, y los asuntos son más simples porque se refieren al producto o a sus características.

Destaca que un asunto puede ser un grupo nominal, un grupo verbal o incluso oraciones citando a otros autores, mientras que marca como una limitación que [KH2006] solo tiene en cuenta adjetivos y verbos.

Asunto marcado en un ejemplo:

(1) Russian Foreig. Min. Ivanov *said* that <asunto>NATO's eastward expansion</asunto> was "towards the wrong direction".
said: es un “reporting verb” que indica un speech event expresando una opinión

Para llegar a la identificación de fuente/asunto se tiene en cuenta:

1. **verbos de reporte (REPORTING VERBS)**: que son verbos indicando un “evento de opinión” (speech event)
2. **palabras que indican opinión (OPINION-BEARING WORDS)**: palabras o frases con polaridad (positiva, negativa, neutra)

Algo a destacar es que los asuntos pueden ser: agentes, objetos concretos, acciones, eventos, ideas abstractas, grupos nominales, incluso grupos verbales, oraciones embebidas.

Lu sigue la siguiente heurística ("heuristic rules" HR), para identificar al asunto

1) Si hay FUENTE gracias a un verbo de reporte:
   a) asunto := el sujeto del objeto del verbo que enmarca el verbo de reporte
   b) Sino asunto := pasando el verbo de reporte, tomar el verbo o las palabras que indican opinión.

2) Sino y no hay FUENTE
asunto := sujeto de la oración
3) Sino asunto := objeto de la oración

Luego se aplica al asunto la expansión de candidato (EP) que toma el asunto más grande posible, tomando todos los modificadores que se le apliquen al "asunto base".

En nuestro trabajo se sigue una idea parecida pues los adjetivos de un sujeto se toman como pertenecientes al asunto, como en el siguiente ejemplo que tiene el adjetivo subrayado:

Opinión 4.1
<opinion>
<fuente>Alfredo Bruno, ex asesor del ex director Nacional de Aduanas, Víctor Lissidini,</fuente>
<predicado>dijo</predicado>
a Montevideo Portal
<mensaje>que sabía que había <asunto>una conspiración clara contra el jerarca y su equipo</asunto> </mensaje>
</opinion>

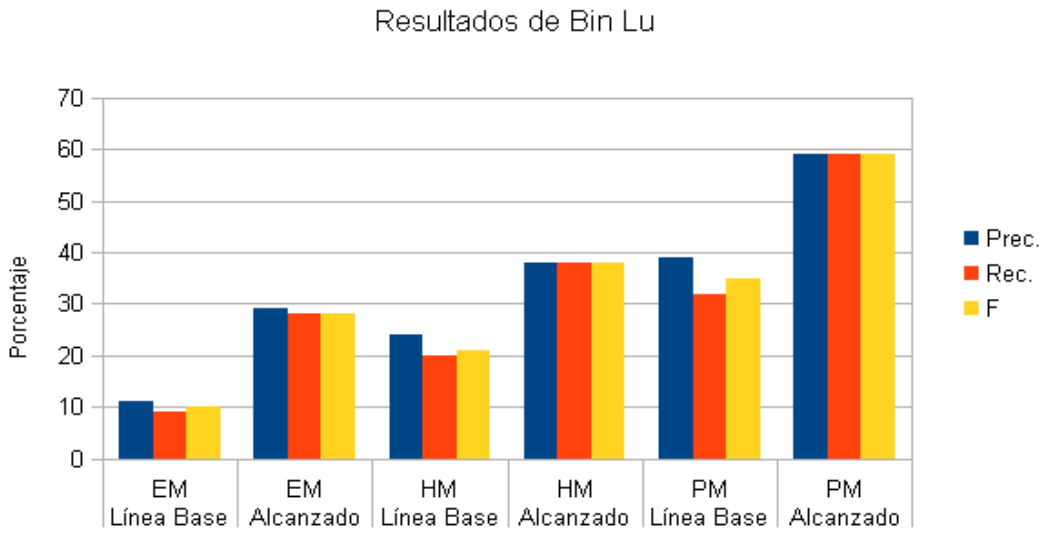
además, las oraciones subordinadas se buscarán incorporar al asunto siempre que se pueda.

En cuanto a la evaluación usa los datos de prueba de NTCIR-7 MOAT como *gold standard*.

Usa medidas de coincidencia exacta (exact match - EM), coincidencia al principio (head match - HM) y coincidencia parcial (partial match - PM).

Para su línea base simplemente toman el sujeto de las *opinion-bearing words*, o en su defecto el objeto, si no hay sujeto.

En particular los resultados que obtuvo son:



Señala que en el dominio de las noticias es muy común tener verbos de reporte (84 y 94% de los casos)

Como se destaca en [Rosá11] también identifica el asunto dentro de lo que sería nuestro mensaje, si bien no lo definen explícitamente.

Concluyen que es difícil identificar asunto y que el área de procesamiento de lenguaje natural debe prestarle atención, planteado el deseo que el asunto se pueda identificar dentro de grupos verbales y oraciones, no solo en grupos nominales.

Annotating Topics of Opinions. Veselin Stoyanov, Claire Cardie [SC2008]

Quienes escriben el artículo describen una metodología manual para anotar asuntos. Señalan la carencia que en los corpora no hay asuntos anotados. Este trabajo se enfrentó a la misma problemática de escasez de recursos por lo que hubo que hacer anotaciones manuales para generar un corpus contra el cual medirse.

Una de las observaciones que hacen es que identificar asuntos pertenece al área de *fine-grained subjectivity analysis*.

Definen asunto como: “El objeto del mundo real, evento, o entidad abstracta que es el sujeto primario de la opinión según la intención de la fuente”. Es interesante que son los únicos que señalan la cuestión de la intencionalidad de la fuente.

Señalan que en:

(1) John believes that there will be a question about Malaria on the midterm.

El asunto puede ser “Malaria” o “Midterm” por lo que se enfrentan al problema de la potencialidad de asuntos múltiples.

Otra definición importante es la de *Target Span:* “Cubre la superficie sintáctica de los contenidos de la opinión”, que se aproxima a nuestra definición de mensaje.

Señalan que si el dominio fuera la crítica de productos la técnica se reduce a hacer una búsqueda en el lexicón por lo que la dificultad es aprender el lexicón.

Comentan el método de Kim-Hovy para adjetivos y verbos y remarcan que su evaluación es limitada porque no tienen asuntos manualmente anotados.

En:

(2) “President X has on many occasions expressed goodwill toward <a>mainland China”

señalan que es fácil de identificar el asunto cuando hay un grupo nominal en el mensaje.

En:

(3) “It all depends on how <a> mainland China interprets <a>President Chen's

latetst remarks on <a>cross-strait relations and how the two sides cultivate on enviroment favorable for <a> resumption of their long stalled dialogue.

Señalan que hay muchos asuntos potenciales y que dependen críticamente del contexto, luego dan otro ejemplo y muestran cómo influye en el asunto de un texto, un fragmento de texto que lo precede, es decir vinculan la problemática a la progresión temática.

Otra dificultad que señalan es que al depender del contexto el asunto de una oración puede cambiar cuando se termina de leer el documento, obtienen un criterio muy claro, pero difícil de determinar: **que el asunto constituye la meta de información primaria de la fuente**. Esto nos llevaría un a el análisis de comunicaciones, que es analizar la intención a partir de la síntesis del mensaje, que como señala [Russ2004] es etapa previa a la generación y a la síntesis del mensaje.

Introducen la noción de opinión correferente en asunto, que hace que un conjunto de oraciones se agrupen bajo un mismo asunto, centrándose un poco en "el cambio de asuntos" y asociando las partes a la globalidad del documento. Finalmente el trabajo se centra en anotar *clusters* de opiniones con una etiqueta creada por el anotador.

Topic Identification for fine-grained opinion analysis. Veselin Stoyanov, Claire Cardie. [SC2008b]

Señalan que dan una *definición operacional*[21] de asunto. Las posibles áreas de aplicación podrían ser QA, resúmenes automáticos, extracción de información y soporte a motores de búsqueda para consultas de ciertas características. Repiten una definición de asunto de otros trabajos:

> *… el **asunto** de una opinión es la entidad del mundo real que es el sujeto de una opinión según la intención de la fuente* **basada en el contexto discursivo**

Vuelven a resaltar la importancia de que haya recursos anotados porque suelen colaborar a el progreso de un área. Señalan el MPQA [Wiebe05] como el mejor, sin embargo indican que originalmente el MPQA iba a tener anotaciones de asuntos pero fue abandonada la tarea por ser demasiado difícil, por lo que extienden el MPQA con asuntos anotados manualmente. Así como que la técnica de look-up no sirve para textos generales.

Establece diferencias con la segmentación de asuntos (topic segmentation), que es particionar texto en una secuencia lineal de temas porque no necesariamente el asunto de las opiniones es espacialmente coherente, es decir, puede haber dos opiniones en una oración.

Observan que hay múltiples asuntos en las opiniones sobretodo cuando tienen polaridad neutra.

21 No se encontró una definición precisa de qué es una definición operacional.

Otra limitación es que el asunto a veces no se menciona en la opinión (se menciona en otra) e incluso no se menciona en todo el documento.

El algoritmo lo centran en identificar los asuntos correferentes, pero la anotación del asunto la dejan para “trabajo futuro” teniendo en cuenta frecuencia de términos en los clusters. Luego muestran las características (features) que eligieron para hacer eso (posicionales, semánticas, opinión), por lo que tal vez un título más adecuado para el artículo hubiera sido “Towards Topic Identification for fine-grained opinion analysis” o “Topic Correferences in Clusters”

Extraction Opinions, Opinions Holders, and Topics Expressed in Online News Media Text. Hovy Kim [Kim2006]

En este trabajo se centran en explotar la estructura semántica usando como insumo a FrameNet, además de enfocarse en opiniones identificadas por adjetivos y por verbos.

Una de las motivaciones originales que plantean es de entender asuntos políticos y sociales, al tener información sobre las relaciones entre distintos actores, ya sean países u organizaciones.

Utilizan etiquetado de roles semánticos (Semantic Role Labeling) y postulan que es crucial investigar las relaciones semánticas para poder identificar el asunto (y la fuente) de una opinión.

Por ejemplo, dado el texto:

On Dec. 7, the Islamic Conference Organization (ICO) denounced the affar as a crime.

Texto	Rol Semántico	
On Dec. 7	Tiempo	
The Islamic (…) (ICO)	Comunicador	Fuente
denounced		
The affair	Evaluee	Asunto
As a crime	razón	

donde se ve dónde está la identificación de fuente y asunto, fruto de una asociación manual para saber cuándo corresponde la fuente o el asunto a determinado rol semántico.

La línea base la determinan de la siguiente manera:

1. Para oraciones cuando el predicado de opinión es un verbo
    1. El sujeto es la fuente y el objeto es el asunto
2. Para oraciones cuando el predicado de opinión es un adjetivo
    1. El sujeto de un adjetivo predicativo es la fuente
    2. La palabra modificada por el adjetivo es el asunto, o en caso de que el adjetivo

sea un predicado el asunto es el sujeto.

Señalan también la dificultad de delimitar claramente un asunto en algunos casos.

En las conclusiones señalan que para identificar asuntos (y fuentes) se requiere una gran cantidad de datos anotados (tener en cuenta que usan técnicas de aprendizaje automático).