Agentes Inteligentes y Web Semántica: Preprocesamiento de Texto de Redes Sociales

Angel Addati Sandra Roger

email: angeladdati@gmail.com , roger@fi.uncoma.edu.ar

Grupo de Investigación en Lenguajes e Inteligencia Artificial
Departamento de Teoría de la Computación - Facultad de Informática
UNIVERSIDAD NACIONAL DEL COMAHUE

RESUMEN

El proyecto de investigación Agentes Inteligentes y Web Semántica, financiado por la Universidad Nacional del Comahue, tiene como objetivo general la generación de conocimiento especializado en el área de agentes inteligentes y en lo referente a la representación y el uso del conocimiento en sistemas computacionales basados en la Web, es decir, lo que se ha llamado la Web Semántica.

En particular, esta línea de investigación tiene como objetivo el desarrollo de una herramienta genérica y expansible para el prepocesamiento de lenguaje natural, para la normalización de texto. Para ello se ha definido su arquitectura básica, brindando la posibilidad de agregar nuevas funcionalidades de forma sencilla.

Palabras Clave: *Normalización de texto, Minería de opinión, Procesamiento de lenguaje natural, Preprocesamiento de texto, Tokenización de texto,*

CONTEXTO

Este trabajo está parcialmente financiado por la Universidad Nacional del Comahue, en el contexto del proyecto de investigación Agentes Inteligentes y Web Semántica, en el contexto de una beca interna doctoral. El proyecto de investigación tiene prevista una duración de cuatro años, desde enero del 2017 hasta diciembre del 2020.

1. INTRODUCCIÓN

Los cambios tecnológicos y el desarrollo de una cultura de las redes sociales, ha llevado a contar con bases de conocimientos globales de gran tamaño conformadas por las expresiones escritas de usuarios: sentimientos, gustos, personalidades, opiniones, tristezas, rencores, etc. Es decir, una base de datos enorme de textos de opinión.

El tema de estudio aparece naturalmente al investigar la literatura sobre minería de opinión. En estos desarrollos se utiliza generalmente como recursos los textos en redes sociales. Estos textos poseen sus propias particularidades de escritura. El trabajo estará orientado en esta dirección, abordando el preprocesamiento de texto por ser la primera etapa necesaria en la minería de opinión que se enfrenta a las dificultades del lenguaje en la redes sociales. Este preprocesamiento tiene como entrada el texto a analizar en lenguaje natural y genera como salida el texto ajustado para que sea más fácil de procesar.

En muchos de los trabajos de investigación, se invierte tiempo al estudio y desarrollo de esta primera etapa de preprocesamiento. Esta etapa es de vital importancia, ya que el texto proveniente de las redes sociales es de muy baja calidad: "*El crecimiento de los medios sociales ha ocasionado que la lingüística computacional se*

encuentre en contacto muy cercano con el lenguaje malo: texto que desafía nuestras expectativas sobre el vocabulario, el habla y la sintaxis" [1].

La ausencia de una herramienta de facto para el preprocesamiento y la carencia de una estandarización en esto, conlleva a que en cada desarrollo de minería de opinión se vuelva a programar un preprocesador que ataque los mismos temas, con distintas aproximaciones y eficiencias. Luego, tenemos distintas soluciones que resuelven con distintos grados de éxitos la misma problemática.

En la literatura existente se pueden apreciar algunos de los diferentes enfoques de cómo se ha abordado esta problemática

En [6,7] se realiza el preproceso de los tuits para tratar el uso de particular del lenguaje en la red como:

- Tratamiento de emoticones, cada emoticon se sustituye por una de estas cinco etiquetas: muy positivo, positivo, neutro, negativo y muy negativo.
- Normalización de URL's: Las direcciones son sustituidas por URL.
- Corrección de abreviaturas más frecuentes, sustituyéndolas por su forma reconocida.
- Normalización de risas.
- Tratamiento de elementos específicos de Twitter (#, @).

Por otro lado, en [12,8,9] se realizan un sistema para la resolución de la tarea de normalización de tweets basado en la concatenación de varios traductores. El trabajo describe un sistema de normalización de tweets en español. Buscando simplicidad y flexibilidad, su arquitectura es de pipeline ya que permite integrar, eliminar e intercambiar módulos de forma sencilla. Dichos módulos se comunican empleando un formato de representación intermedia codificado como texto de naturaleza estructurada y jerarquizada. Un tweet está formado por términos y para cada término existe una serie de candidatos para su normalización.

En [11] presenta una aproximación lingüística basada en traductores de estados finitos para la normalización léxica de mensajes de Twitter en español. El desarrollo de esta propuesta consiste de traductores que son aplicados a token fuera del vocabulario. Los traductores implementan modelos lingüísticos que genera un conjunto de candidatos. Mediante un modelo estadístico se obtiene la secuencia de palabra más probable.

En [13] se utiliza un clasificador para detectar palabras mal formadas y generar posteriormente, candidatos de corrección basados en su similitud morfológica. Para la selección del candidato, se realiza los siguiente: sobre la corrección más probable de la palabra se utiliza tanto la similitud de palabras como el contexto.

Apoyándonos en las investigaciones existentes, dentro del campo del procesamiento de lenguaje natural (PLN), vamos a definir el preprocesamiento como la primera transformación del texto de entrada. Su finalidad es generar un texto de mejor calidad para las etapas posteriores de trabajo. Por un texto de mejor calidad, entendemos a un texto segmentado, corregido (o con la menor cantidad posible de errores de entrada) y optimizado.

2. LÍNEA DE INVESTIGACIÓN Y DESARROLLO

El proyecto de investigación Agentes Inteligentes y Web Semántica tiene como objetivo general generar conocimiento especializado en el área de agentes inteligentes y en lo referente a la representación y el uso del conocimiento en sistemas computacionales basados en la web, es decir lo que se ha llamado la Web Semántica.

Específicamente, esta línea se centra en el estudio de un sistema multiagente para la mejora en la calidad de los textos de opinión pública sobre un determinado tema de interés,

más precisamente sobre textos escritos en tweets.

Figura 1: Arquitectura básica del preprocesador

El trabajo se centrará en dos ejes: uno conceptual y otro práctico. Primero, se intentará definir cuál es la funcionalidad que debe tener el preprocesador. Aquí el foco estará en delinear la funcionalidad sin importar el uso que se le dará con posterioridad. Es decir, hallar una estandarización de lo que debe hacer el preprocesador. Segundo, se desarrollará una herramienta que pueda ser utilizada por cualquier investigador para preprocesar sus datos. Para ello debe cubrir con lo especificado y debe tener la flexibilidad para ser expandida con nuevas características.

Basándonos en las distintas soluciones implementadas en otras investigaciones, el preprocesador se dividió en la ejecución de 6 etapas progresivas. En cada una de las etapas, se abordará una agrupación de problemas comunes que este debe resolver. La etapa contendrá módulos en donde estará la lógica de resolución de un problema. Una vez que el texto de entrada pasa por cada una de las etapas, se lo dará por preprocesado. La Figura 1 muestra la arquitectura básica para la normalización de texto planteada.

Cada etapa contiene sus propios desafíos, estrategias, definiciones e investigaciones asociadas. A continuación, se enumeran cada una de las etapas y se dará una breve descripción de las mismas:

ETAPA 1: Tokenizacion.

Aquí entrará el texto informal en formato plano y saldrá segmentado en tokens. El token será una palabra o un signo de puntuación. Para ello, es importante una correcta definición de los sig- nos de puntuación y espacios para una apropia- da separación en palabras

ETAPA 2: Clasificación en IV y OOV.

A esta fase entrará el texto segmentado en palabras y saldrán las palabras clasificadas en IV y OOV. Por IV se significa "*In Vocabulary*" cuya traducción es dentro del lenguaje. Las palabras dentro del lenguaje serán las palabras consideradas válidas, correctas y sin errores. Es decir, las palabras pertenecientes a nuestro lenguaje, bien escritas y reconocidas por los diccionarios. Por otro lado, OOV significa "*Out Of Vocabulary*". La traducción es fuera del lenguaje. Son las palabras desconocidas o no reconocidas como co-

rrectas. Estás palabras son las candidatas a corregir si es posible.

ETAPA 3: Normalización.

Aquí entrará el texto segmentado en palabras clasificadas en IV y OOV. Por cada palabra OOV, se intentará corregirla y encontrar su palabra IV semánticamente equivalente. Aquí se aplicará diferentes heurísticas con el fin de buscar palabras IV candidatas para cada OOV. No necesariamente toda OOV tendrá una IV candidata, y los candidatos para una OOV pueden ser más de uno. Al finalizar, al texto segmentado en palabras clasificadas en IV y OOV, se le agregarán las palabras IV candidatas.

ETAPA 4: Selección.

En esta fase se deberá seleccionar por cada OOV una de las palabras IV candidatas. El algo- ritmo de selección podrá tener varios criterios, por ejemplo: el orden, la probabilidad, algún peso asociado al método aplicado para encontrar la palabra candidata, etc. La salida del proceso será la entrada agregando la palabra IV candida- ta elegida.

ETAPA 5: Capitalización.

Esta fase generará como salida la apropiada capitalización de todas las palabras, tanto las corregidas en la etapa anterior como las no corregidas. En los mensajes escritos en los medios sociales, la correcta capitalización es un problema, ya que los usuarios cometen errores, no respetan ortografías y utilizan la mayúscula para expresar sentimientos o énfasis. La finalidad de esta etapa es aplicar un criterio estandarizado al uso de mayúsculas y minúsculas.

ETAPA 6: Optimización.

Esta es probablemente la etapa más difusa de las 6 y la más compleja de acotar. Por optimización se va a entender a toda transformación de la entrada, no incluida en etapas anteriores, que la simplifiqué. Luego, aquí el foco estará en definir módulos con optimizaciones básicas o comunes y brindarle al desarrollador un marco de trabajo sencillo para expandir la funcionalidad. Algunos ejemplos de optimización podrán ser: la eliminación de palabras con poco contenido semántico, la lematización de palabras, la eliminación de signos de puntuación, la eliminación de palabras repetidas, etc.

3. RESULTADOS OBTENIDOS Y TRABAJOS FUTUROS

Inicialmente, se hizo un relevamiento de las diferentes estrategias de normalización de texto para poder crear un marco comparativo y poder evaluarlas. Del mismo surgió la arquitectura planteada anteriormente.

Se está trabajando en terminar de desarrollar el diseño profundizando el detalle de cada etapa para estudiar más a fondo cada problemática en particular. El objetivo es crear un preprocesador más potente y configurable para el usuario.

Además, se plantea generar un software funcional dedicado exclusivamente a la tarea de preprocesamiento de lenguaje natural orientado a textos escritos en redes sociales.

Finalmente, teniendo como partida el preprocesador desarrollado como primer componente, continuar en el desarrollo de los componentes subsiguientes para tener una iteración completa de un software de análisis de opinión.

4. FORMACIÓN DE RECURSOS HUMANOS

Durante la realización de este sistema se espera lograr, como mínimo, la culminación de 2 tesis de grado dirigidas y/o codirigidas por los integrantes del proyecto.

Finalmente, es constante la búsqueda hacia la consolidación como investigadores de los miembros más recientes del grupo.

5. BIBLIOGRAFÍA

[1] Eisenstein, J. (2013). *What to do about bad language on the internet.* Atlanta, Georgia: Association for Computational Linguistics.

[2] Graña Gil, J., Rodríguez, B., Mario, F., & Vilares Ferro, J. (2001). *Etiquetación robusta del lenguaje natural: preprocesamiento y segmentación.* Sociedad Española para el Procesamiento del Lenguaje Natural.

[3] Padilla, A. P. (2004). *Técnicas lingüísticas aplicadas a la búsqueda textual multilingüe: ambigüedad, variación terminológica y multilingüismo.* Sociedad Española para el Procesamiento del Lenguaje Natural.

[4] Ramírez Bustamante, F., Sánchez León, F., & Declerck, T. (1997). *Correccción gramatical y preprocesamiento.* Sociedad Española para el Procesamiento del Lenguaje Natural.

[5] Dubiau, L., & Ale, J. M. (2013). *Análisis de Sentimientos sobre un Corpus en Español: Experimentación con un Caso de Estudio.* 14th Argentine Symposium on Artificial Intelligence, ASAI2013.

[6] Vilares, D., Alonso, M A., Gómez-Rodríguez, Carlos. *Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico*. Procesamiento del Lenguaje Natural, Vol. 51. 2013, pp 127-134.

[7] Vilares, J., Alonso, M. A., & Vilares:, D. (2013). Prototipado Rápido de un Sistema de Normalización de Tuits: Una Aproximación Léxica.

[8] Alegria, I., Aranberri, N., Fresno, V., Gamallo, P., Padró, L., Vicente, I. S., . . . Zubiaga, A. (2013). Introducción a la Tarea Compartida Tweet-Norm 2013: Normalización Léxica de Tuits en Español.

[9] Alegria, I., Etxeberria, I., & Labaka, G. (2013). Prototipado Rápido de un Sistema de Normalización de Tuits: Una Aproximación Léxica. CEUR-WS.org.

[10] Orquín, A. F., Rodríguez, K. V., Amable, A. C., Martín, R. P., Echarte, Á. L., & Morera, D. C. (2009). *Sistema para el pre-procesamiento de textos para el Procesamiento del Lenguaje Natural.*

[11] Porta, J., & Sancho, J. L. (2013). *Word normalization in Twitter using finite-state transducers.*

[12] Alegria, I., Etxeberria, I., Labaka, G. *Una Cascada de Transductores Simples para Normalizar Tweets*. CEUR Workshop Proceedings. Vol. 1086, 2013. pp. 15-19.

[13] Han, B, Baldwin, T. *Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter.* Proceedings of the 49th Annual Meeting of the Associationfor Computational Linguistics: Human Language Technologies – Vol. 1. 2011 pp. 368-378. Association for Computational Linguistics