Subsistemas para Análisis de Textos

Marina Cardenas[1], Julio Castillo[1], Nicolas Hernandez[1]
[1] Laboratorio de Investigación de Software/Dpto. Ingeniería en Sistemas de Información/ Facultad Regional Córdoba/ Universidad Tecnológica Nacional
{jotacastillo, ing.marinacardenas}@gmail.com

Resumen

En este artículo se presenta el proyecto de investigación denominado *Desarrollo de Sistemas de Análisis de Texto*.

Este proyecto se centra en el estudio, análisis y procesamiento de información textual y en el desarrollo de herramientas software que permitan agilizar y facilitar su exploración e inferencia de información.

Se presenta desde un enfoque de sistemas y subsistemas, las partes principales proyecto y las herramientas que se están desarrollando para abordar el procesamiento de textos.

Las líneas de investigación en la que se encuadra el proyecto es dentro de las áreas de lingüística computacional y de aprendizaje automático, para atacar problemas con orígenes de texto con y sin estructura definida.

Palabras clave: análisis de texto, extracción de información, corpus, machine learning.

Contexto

En este artículo se describe el proyecto denominado Análisis de Texto (ADT), un proyecto homologado por la SCyT de la UTN [1], y que se enmarca dentro del área de lingüística computacional [2]. El proyecto físicamente se desarrolla en el Laboratorio de Investigación de Software LIS[1] del Dpto. de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba (UTN-FRC).

En el laboratorio convive el desarrollo de proyectos de muy variada envergadura, entre los temas abordados resaltamos el estudio y diseño de compiladores y metalenguajes de programación, proyectos relacionados con análisis de imágenes, autómatas y modelos de pronósticos, y un proyecto de educación en tecnologías de información [3].

El LIS está compuesto por doctores, doctorandos, ingenieros, docentes-investigadores, pasantes, becarios alumnos y de posgrado. También se trabaja con expertos en ciencias sociales de CONICET, con los cuales se colabora para desarrollar modelos computacionales que permitan abordar la problemática de la salud humana siguiendo una perspectiva holística y en base a la interacción con su entorno [4].

Los proyectos desarrollados en el LIS se enmarcan en varias líneas de investigación y desarrollo (que se mencionan en la siguiente sección) y que dieron origen o colaboraron en la creación de varios grupos UTN de investigación.

1. Introducción

El proyecto denominado Desarrollo de sistemas de análisis de texto (ADT) aborda el problema de análisis y procesamientos de textos los cuales pueden provenir de orígenes muy variados y disímiles [5].

En ese contexto, una de las tareas que se desarrollan es la orientada a la construcción de corpus lingüísticos, cuya necesidad varía de acuerdo al problema que se intenta abordar.

Para llevar a cabo esta tarea se ha desarrollado un subsistema que facilite la clasificación manual y que sirva para disminuir la tasa de errores de anotación de

[1] www.investigacion.frc.utn.edu.ar/mslabs/

los etiquetadores humanos, el cual hemos denominado Subsistema de Asistente de Creación de Corpus (ACC).

Nos concentramos en una serie de problemas de procesamientos de textos que detallamos a continuación:

- el problema de la normalización de diversas fuentes de información en un repositorio común, al cual lo hemos aproximado mediante el desarrollo de un subsistema denominado Subsistema de Mapeo de Datos
- el problema de evaluación y selección de métricas en base a la elección objetiva de un criterio, a través del Subsistema de Banco de Prueba de Algoritmos de Semejanza.
- el problema de detección de similitud en textos se ha abordado mediante el desarrollo del Subsistema de Análisis de Similitud
- y el problema de visualización de diferencias en el texto, se ha atacado a través del desarrollo de un Subsistema de Visualización de Diferencias en Documentos.

Los resultados proporcionados por estos subsistemas se pueden utilizar en el contexto de varias tareas de análisis de texto, ya que pueden ser útiles para problemas de extracción de información y minería de datos en textos no estructurados [6,7]. Otros problemas, como búsquedas de implicaciones en textos [8], implicaciones textuales [9, 10], extracción de información [11,12] o minería de datos [13, 14] también pueden hacer uso de estos subsistemas.

2. Líneas de Investigación, Desarrollo e Innovación

La línea de investigación en la que se enmarca este proyecto es computación lingüística desde el abordaje del aprendizaje por computadoras.

Esta línea abarca un campo científico interdisciplinar cuyo principal objetivo es el de desarrollar sistemas con la capacidad de reconocer y comprender el lenguaje natural humano a través de modelos computacionales. Más aún, esto da origen a otras sublíneas de investigación en lingüística, como la denominada Lingüística de Corpus [15], que es aquella que aborda los problemas del lenguajes a partir de ejemplos reales de producciones lingüísticas (orales o escritas) que se almacenan en un computador y a partir de las cuales es posible inferir conocimientos.

A su vez, hay otras líneas de investigación, que se desarrollan en el mismo laboratorio (LIS), como lo son: la línea de investigación en teoría de autómatas y gramáticas formales, la línea de investigación de construcción de modelos de pronósticos, y otra de modelado de problemas de ciencias sociales utilizando técnicas de inteligencia artificial.

Con estas líneas de investigación, y en particular con los proyectos que los componen, se forma una interesante sinergia que permite fortalecer a cada línea y proyecto de investigación en base a la fructífera interacción y colaboración entre los investigadores que la integran.

Estas líneas de investigación se han plasmado concretamente en la creación de un grupo UTN de investigación.

3. Resultados

En el proyecto se han desarrollado varios sistemas de análisis y procesamiento de texto, entre los más importantes mencionaremos a un sistema Software de Asistente de Creación de Corpus (ACC), un Sistema de Mapeo de Datos (PMD), un Sistema de detección de similitudes en archivos de código fuente (SDS), un subsistema de Banco de Pruebas de Algoritmos de Semejanza (BPAS), y un subsistema de Visualización de Diferencias en Documentos (VDD). Estos sistemas se describen con más detalle a continuación.

El ACC permite realizar la clasificación de un conjunto de fenómenos de origen léxico, sintáctico, semánticos y morfológicos, mediante la selección de

subcadenas entre las que se sostiene un determinado fenómeno lingüístico. Esto es especialmente útil en tareas como la implicación de textos o en la detección de paráfrasis. Una funcionalidad destacable del ACC es permitir trabajar con subcadenas de texto, y anotar el fenómeno lingüístico que entre ellos se sostiene.

Este subsistema continúa en ampliación de sus funcionalidades, y se está trabajando con un módulo que genere reportes estadísticos que permitan calcular automáticamente la consistencia del material de entrenamiento, midiendo el acuerdo inter-anotador.

El subsistema de Mapeo de Datos (PMD) permite procesar información originada en múltiples fuentes de datos estructurados, tales como archivos de textos, tablas de bases de datos diferentes, y armar con ellos un repositorio de información centralizada, sobre la cual puedan aplicarse posteriormente técnicas de recuperación de información. La característica positiva es que concentra la información de manera centralizada, pero existe un aspecto negativo, ya que la información debe mantenerse actualizada periódicamente a los efectos de reflejar los últimos cambios de la información en origen. En este sentido, es similar a lo que ocurre con bases de datos multidimensionales y con cubos OLAP.

El subsistema de Banco de Pruebas de Algoritmos de Semejanza (BPAS), permite trabajar con archivos de configuraciones en los cuales se encuentran un conjunto de criterios y un peso relativo de importancia. Luego, en base a estos valores se emplea el método de Jerarquía Analítica (AHP), el cual permite encontrar la opción óptima en base a los criterios previamente establecidos en los archivos de configuraciones. El método AHP [16] es utilizado cuando se deben tomar decisiones en escenarios complejos [17] y provee una manera de encontrar la solución óptima al problema de seleccionar una opción entre un conjunto de opciones posibles.

Esta herramienta dota de un criterio más preciso y riguroso para justificar la elección objetiva de una alternativa ante un conjunto de opciones.

El subsistema de Análisis de Similitud (SAS) realiza el procesamiento de archivos de códigos fuentes en varios lenguajes de programación e informa un valor porcentual con el grado de similitud de los mismos. Esta herramienta permite la comparación de un archivo contra un conjunto de archivos a los efectos de poder encontrar los k-archivos más similares. Es útil en la detección de patrones en códigos fuentes (que pueden estar escritos en varios lenguajes de programación), en la reutilización de código, entre otras potenciales aplicaciones.

Por último, el subsistema de Visualización de Diferencias en Documentos (VDD) cuenta con una interfaz gráfica que permite contrastar visualmente la diferencia entre dos archivos de texto. Para ello, se usan diferentes colores a los efectos de resaltar la información nueva, faltante o modificada, en relación a los documentos origen-destino.

Los subsistemas mencionados anteriormente se encuentran en etapa de desarrollo, ampliación e integración en un pipeline de trabajo común.

4. Formación de Recursos Humanos

El equipo de investigación y desarrollo de software, está formado por docentes investigadores de la Universidad Tecnológica Nacional, Facultad Regional Córdoba, que a continuación se detallan:

- Un doctor en ciencias de la computación, quién guía a becarios de grado y de posgrado, dirige prácticas profesionales supervisadas y pasantías.
- Un doctorando en ingeniería con mención en sistemas de información de la Universidad Tecnológica Nacional,

Facultad Regional Córdoba que trabaja en temáticas similares a las de este proyecto. Colabora también en la dirección de becarios de grado y posgrado realizadas en el proyecto.

- Un grupo de dos o tres becarios alumnos participan cada año realizando actividades de investigación en el proyecto, complementando así su formación curricular.
- Eventualmente se desarrollan prácticas supervisadas que constituyen uno de los requisitos para la obtención del grado de Ingeniero, y son realizadas en tareas específicas en el proyecto, acotadas en tiempo y con alcances claramente definidos.
- El proyecto además posee investigadores en formación y en proceso de categorización.
- Finalmente, se han realizado charlas de difusión y jornadas de capacitación a alumnos y a docentes de ingeniería en sistemas de información en las líneas temáticas enumeradas anteriormente.

5. BIBLIOGRAFÍA

[1] Marina Cardenas, Julio Castillo, Martín Navarro, Nicolás Hernández, Melisa Velazco. (2019). "Herramientas para el desarrollo de sistemas de análisis de textos no estructurados". XXI Edición de Workshop de Investigadores en Ciencias de la Computación. WICC 2019, UN de San Juan, Argentina, 25 y 26 de abril de 2019.

[2] Judith Klavans y Philip Resnik. The Balancing Act. Combining Symbolic and Statistical Approaches to Language. MIT Press, 1996.

[3] Castillo, J., Cárdenas, M., Serrano, D. (2011). "Experiencias en el Desarrollo de Competencias de Programación en UTN-FRC". VI Congreso de Tecnología en Educación y Educación en Tecnología. Teyet 2011.

[4] Rojas MC, Meichtry NC, Ciuffolini MB, Vásquez JC, Castillo J. (2008). Repensando de manera holística el riesgo de la vivienda urbana precaria para la salud: un análisis desde el enfoque e la vulnerabilidad sociodemográfica. Salud Colectiva. 2008; 4(2): 187- 201.

[5] Castillo Julio J., Cardenas Marina E., Curti Adrián, Casco Osvaldo, Navarro Martín, Hernández Nicolás A., Velazco Melisa. Desarrollo de sistemas de análisis de texto. XIX Workshop de Investigadores en Ciencias de la Computación (WICC 2017). 2017.

[6] I. Goodfellow, Y. Bengio y A. Courville. Deep Learning. MIT Press. 2016.

[7] N. Buduma. *F*undamentals of Deep Learning*:* Designing Next-Generation Artificial Intelligence Algorithms. O' Really book. 2015.

[8] Castillo J. Sagan in TAC2009: Using Support Vector Machines in Recognizing Textual Entailment and TE Search Pilot task. TAC, 2009.

[9] Castillo J., Cardenas M. Using Sentence Semantic Similarity Based on WordNet in Recognizing Textual Entailment. Iberamia 2010, LNCS, vol. 6433, pp. 366-375, 2010.

[10] Castillo J. Using Machine Translation Systems to Expand a Corpus in Textual Entailment. Proceedings of the Icetal 2010, LNCS, vol. 6233, pp.97-102, 2010.

[11] FeldmanR. y Hirsh H.. Exploiting Background Information in Knowledge Discovery from Text. Journal of Intelligent Information Systems, 1996.

[12] Lewis, D.. Evaluating and optmizing autonomous text classification systems. In Proceedings of SIGIR-95, 18th

ACM International Conference on Research and Development in Information Retrieval. Seattle, US, págs. 246-254, 1995.

[13] M. Craven y J. Shavlik. Using Neural Networks for Data Mining. Future Generation Computer Systems, 13, págs. 211-229, 1997.

[14] Gaikwad, S., Chaugule, A., & Patil, P.B. (2014). Text Mining Methods and Techniques. International Journal of Computer Applications, 85, 42-45.

[15] Stefan Th. (2006). Y Anatol Stefanowitsch. Corpora in Cognitive Linguistics. CorpusBased Approaches to Syntax and Lexis, Berlin: Mouton, pág. 1-17.

[16] Bhushan N., Kanwal R. (2004). "Strategic Decision Making: Applying the Analytic Hierarchy Process". London: Springer-Verlag.

[17] Jiandong Zhang, Jin Yang, Xuan Ma and Juanjuan Gong. (2013). "Research and Application on Improved group decision making AHP algorithm". Applied Mechanics and Materials Vols. 336-338.