# Herramientas para el desarrollo de sistemas de análisis de textos no estructurados

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

## Resumen

En este artículo se describen las actividades desarrolladas y los subsistemas que conforman el proyecto de investigación que se denomina Desarrollo de Sistemas de Análisis de Texto, un proyecto de investigación homologado por la Secretaría de Ciencia y Técnica (SCyT) de la UTN.

En este proyecto se trabaja en el desarrollo de herramientas que permitan realizar el análisis de información textual de una manera más eficiente, e involucra la creación de material de entrenamiento de los sistemas de análisis de textos y el desarrollo de herramientas software que sirvan para el análisis y procesamientos de grandes volúmenes de textos.

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

## Contexto

Este artículo presenta el proyecto denominado Desarrollo de Sistemas de Análisis de Texto (ADT), que es un proyecto homologado por la SCyT de la UTN, que se enmarca dentro del área de lingüística computacional, y que tiene como objetivo el desarrollo de herramientas para análisis textual de diversas fuentes de información en formato no estructurado.

Actualmente, el proyecto se encuentra dentro del grupo de investigación denominado Grupo de Aprendizaje Automático, Lenguajes y Autómatas (GA$^2$LA) con fecha de creación del grupo en Octubre de 2018 en la UTN-FRC.

En el grupo GA$^2$LA se desarrollan proyectos relacionados con autómatas y lenguajes formales, procesamiento del lenguaje natural, y aprendizaje automático, y en especial proyectos orientados a la aplicación de la inteligencia artificial para resolver problemas de las ciencias sociales.

En este contexto, los recursos humanos con los que cuenta el grupo son ingenieros en sistemas de información, licenciados/doctores en ciencia de la computación, demógrafos y arquitectos, y se complementa con becarios y pasantes.

Los problemas que aborda este proyecto de investigación están relacionados a la extracción de información, detección y reconocimiento de paráfrasis, y de reconocimiento de implicación textual. Es decir, problemas relacionados a la identificación de oraciones (o párrafos) que tengan el mismo significado, o bien la identificación de oraciones-párrafos que estén semánticamente relacionados entre sí mediante una relación de implicación.

Físicamente, los integrantes del proyecto desarrollan sus actividades 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).

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

## 1. Introducción

Los sistemas de análisis de texto se enfrentan a problemas difíciles dentro del área de la ciencia de la computación, debido principalmente a la dificultad del análisis del lenguaje (derivada de la ambigüedad del lenguaje) relacionado a la etapa de análisis semántico, como así también, a los relativamente escasos materiales de entrenamiento y a la capacidad de cómputo necesaria para correr determinados algoritmos muy demandantes en recursos de hardware.

Como una manera de abordar el problema de la escasez del material de entrenamiento, en este proyecto se trabaja con la creación de corpus lingüísticos que puedan ser utilizados como material de entrenamiento en sistemas de aprendizaje supervisado o semi-supervisado. Una vez construidos, los conjuntos de entrenamiento pueden ser de utilidad en diversos problemas relacionados con el análisis de textos.

Adicionalmente, la correcta construcción y diseño del material de entrenamiento puede ayudar a abordar el problema de la ambigüedad en el texto, que son complemento de diversas técnicas de análisis del discurso.

En cuanto a la capacidad de cómputo, es un desafío pendiente, y trabajo futuro, el paralelizar algoritmos secuenciales que necesitan gran poder de cómputo y que son costos en cuanto al tiempo de ejecución.

Este proyecto aborda el problema del análisis e interpretación de textos no estructurados, extracción de información y minería de datos [1][2][3][4][5] basados en técnicas de aprendizaje automático por computadora, entre ellas destacamos las basadas en redes neuronales artificiales [6][7][8], máquinas kernel [9], deep learning [10][11], y árboles de decisión.

En el marco de este proyecto se están desarrollando sistemas de análisis y procesamiento de texto, entre los que podemos destacar:

- Software de Asistente de Creación de Corpus (ACC): es un software que permite construir material de entrenamiento para aplicaciones de minería de datos sobre texto no estructurado.
  Este software ha permitido la creación de corpus para sistemas de reconocimientos de paráfrasis.

- Sistema de Mapeo de Datos (SMD): Software que permite manipular orígenes de datos estructurados y centralizarlos para un posterior análisis con técnicas de recuperación de información o de minería de datos.
  Este software permite centralizar en un repositorio común, la información dispersa en grandes bases de datos.

- Sistema de detección de similitudes en archivos de código fuente (SDS). Es un sistema que tiene como objetivo determinar la similitud de archivos de código fuente escritos en diferentes lenguajes de programación. Se está trabajando a nivel de granularidad de archivos de código fuente y a nivel de funciones o módulos.

## 2. Líneas de Investigación, Desarrollo e Innovación

La línea de investigación de este proyecto es la lingüística computacional abordada con técnicas de aprendizaje automático.

Como se ha mencionado anteriormente, una de las herramientas desarrolladas es un software Asistente de Creación de Corpus, el cual forma parte de la línea de investigación denominada Lingüística de Corpus [12]. Esta rama de la lingüística se caracteriza porque los resultados de las investigaciones se realizan en base a las

entidades lingüísticas obtenidas de textos tabulados y clasificados.

El equipo de investigación también trabaja en el desarrollo de modelos computaciones en general. Uno de los modelos desarrollados permite modelar la ocurrencia de incendios forestales, y forma parte de otro proyecto de investigación que se denomina Modelos para la Predicción de Incendios Forestales, un proyecto homologado por la SCyT de la UTN.

La línea de innovación está relacionado a la integración de los resultados obtenidos de la lingüística de corpus, de la construcción de modelos y del uso de técnicas de aprendizaje automático, en el marco del grupo GA$^2$LA [13].

## 3. Resultados

Los subsistemas que se mencionan en la Sección de Introducción se están desarrollando en paralelo, y son los siguientes:

- Asistente de Creación de Corpus (ACC),
- Sistema de Mapeo de Datos (SMD), y
- Sistema de Detección de Similitudes en archivos de código fuente (SDS).

El Software de Asistente de Creación de Corpus (ACC) se desarrolla con el objetivo de facilitar la construcción de material de entrenamiento que se necesita en los algoritmos de aprendizaje supervisado. La calidad y el tamaño del conjunto de entrenamiento impacta directamente en la efectividad de los algoritmos de clasificación, es por ello que se necesita un tamaño adecuado del material de entrenamiento y que el mismo sea consistente.

El ACC es una herramienta semiautomática que permite a los usuarios sistematizar e identificar los fenómenos lingüísticos presentes en diversos textos. Además, permite clasificar pares de texto con paráfrasis, a la vez que facilita la lectura y el estudio de los corpus generados.

Como resultado se generan corpus etiquetados que son necesarios en la etapa de entrenamiento en sistemas de aprendizaje supervisado.

Esta herramienta permite registrar diversos fenómenos lingüísticos a nivel léxico, sintáctico, morfológico y semántico. El material de entrenamiento construido es para el idioma español e inglés, y es utilizado en sistemas de RTE (Implicación Textual).

Entre las aplicaciones que potencialmente podrían utilizar este material de entrenamiento podemos citar a traducción automática asistida por computador, creación de corpus de paráfrasis, creación de corpus para implicación de textos, resumen automático, entre otras posibles aplicaciones.

Se han creado dos corpus monolingües en español con 100 pares de elementos [14], clasificados según se describe en [15]. Además, se está terminado de construir un tercer corpus de 100 pares monolingüe en inglés.

El software de Sistema de Mapeo de Datos (SMD) se plantea con el objetivo de realizar una manipulación, procesamiento (desde diferentes fuentes y orígenes de datos) y almacenamiento de la información en un repositorio común centralizado (una base de datos en SQL Server). Se pretende entonces, explotar el repositorio con diversas técnicas del área de minería de datos y técnicas de recuperación de la información.

Hay que notar, que este sistema necesita mantenerse actualizado para que la información del repositorio sea correcta y fiable. El lapso de tiempo necesario entre cada actualización dependerá de la aplicación que se esté desarrollando.

El SMD está basado en una aplicación web que almacena los datos normalizados en una estructura estándar de una base de

datos SQL Server facilitando la búsqueda y análisis de textos.

La aplicación permite la selección del origen y destino de datos estructurados, y para realizar el mapeo (transformación) de datos de manera interactiva.

Una vez construido un repositorio para un dominio de problema dado, es posible realizar minería de datos sobre el mismo.

El tercer subsistema se denomina Sistema de Detección de Similitudes en códigos fuente (SDS). Es el subsistema de desarrollo más reciente, y hasta el momento, permite la integración de diversas medidas de similitud léxica aplicadas sobre archivos de códigos fuentes escritos en el mismo lenguaje de programación (C o Java) [16].

La identificación de similitudes de código puede servir para varios propósitos, como la trazabilidad en proyectos de desarrollo de software, detección de reutilización de código, identificación de vulnerabilidades en el código, y detección de plagio.

El sistema SDS permite efectuar una comparación de un archivo de código fuente contra un conjunto de archivos. Debe notarse que la identificación de dos archivos potencialmente similares requiere un costo cuadrático en cantidad de comparaciones. Estas operaciones son muy costosas cuando el tamaño del conjunto es grande. Es por esta razón que se necesita técnicas de multiprocesamiento para obtener resultados en tiempos aceptables. Esto es parte de nuestro trabajo futuro.

## 4. Formación de Recursos Humanos

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

- Una doctorando en ingeniería con mención en sistemas de información en la UTN-FRC, que está trabajando específicamente en el subsistema de detección de similitudes en archivos de código fuente (SDS). Además, realiza la dirección de becarios de posgrado y de becarios de grado en el contexto del proyecto.
- Un doctor en ciencias de la computación, cuya tarea principal es la dirección del proyecto, y la dirección de becarios y prácticas supervisadas.
- Un maestrando en Ingeniería en Sistemas de Información de la UTN-FRC.
- Anualmente participan en el proyecto, alumnos que realizan su práctica supervisada. La cantidad de alumnos depende de la necesidades del proyecto y de los alumnos dispuestos a trabajar en el proyecto. Las prácticas supervisadas son un requisito necesario para la obtención del grado de Ingeniero. Las mismas pueden realizarse en la industria o en el ámbito académico en el contexto de proyectos de investigación homologados.
- En el proyecto participan dos alumnos becarios anualmente, complementando así su formación curricular desde el punto de vista científico.
- Adicionalmente, se realizan 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] Judith Klavans y Philip Resnik. The Balancing Act. Combining Symbolic and Statistical Approaches to Language. MIT Press, 1996.

[2] C. Manning y H. Schutze. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA, 1999.

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

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

[5] 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.

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

[7] 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.

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

[9] Castillo J. An approach to Recognizing Textual Entailment and TE Search Task using SVM. Procesamiento del Lenguaje Natural 44, 139-145, 2010. 4, 2010.

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

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

[12] Stefan Th. Y Anatol Stefanowitsch. Corpora in Cognitive Linguistics. CorpusBased Approaches to Syntax and Lexis, Berlin: Mouton, pág. 117, 2006.

[13] Vázquez, Juan C., Castillo, Julio J., Constable, Leticia, Cardenas, Marina E. GA²LA: Grupo de Aprendizaje Automático, Lenguajes y Autómatas. XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018). 2018.

[14] Cardenas Marina E., Castillo Julio J., Navarro Martín, Hernández Nicolás A., Velazco Melisa. Sistemas de análisis textual en formato no estructurado. XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018). 2018.

[15] Castillo Julio, Cardenas Marina, Curti Adrian, Velazco Melisa, Casco Osvaldo, Navarro Martin. Herramientas para Aplicaciones de Análisis de Textos. Congreso Nacional de Ingeniería Informática / Sistemas de Información. CONAIISI 2017.

[16] Cardenas, Marina E., Castillo, Julio J. Procesamiento de textos estructurados. XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018). 2018.