Desarrollo de Sistemas de Análisis de Texto

Julio Castillo[1], Marina Cardenas[1], Adrián Curti[1], Osvaldo Casco[1]
Martin Navarro[1], Nicolás Hernandez[1], Melisa Velazco[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 proyecto se busca utilizar técnicas de aprendizaje automático para analizar y procesar textos que pueden estar en formato estructurado como no estructurado. Se han desarrollado un conjunto de herramientas que pueden ser utilizadas en el área de computación lingüística para diversos fines, entre los que se encuentran construcción de material de entrenamiento, procesamiento de datos estructurados y detección de similitudes entre fragmentos de textos.

Los problemas que se abordan en este proyecto son, entre otros, reconocimiento de implicación de textos e identificación de paráfrasis.

En este artículo se presenta la línea de investigación en la que se encuentra el proyecto, y se describen tres herramientas desarrolladas enel mismo.

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

Contexto

El presente proyecto denominado Análisis de Texto (ADT) es un proyecto homologado por la SCyT de la UTN, y se enmarca dentro del área de computación lingüística. El mismo es desarrollado 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/

A su vez, este proyecto se encuentra dentro del grupo de investigación denominado Grupo de Inteligencia Artificial (o GIA) de la UTN-FRC.

Este grupo nuclea proyectos de una línea de investigación relacionada al área de inteligencia artificial, redes neuronales artificiales, autómatas celulares, análisis y procesamiento de imágenes, minería de datos, y su aplicabilidad a la resolución de problemas de las ciencias naturales y de las ciencias sociales. El grupo se conforma por doctores, ingenieros, licenciados, becarios y pasantes.

En el mismo se investigan temas de ciencia básica, como puede ser el estudio de los momentos de aprendizaje de redes neuronales artificiales y su relación con autómatas celulares, como así también, se estudian aspectos de ciencia aplicada, como la estimación del cálculo del riesgo de la vivienda urbana para la salud, que se aplica concretamente en el campo de ciencias sociales.

1. Introducción

Mediante este proyecto se propone abordar el problema del análisis e interpretación de textos no estructurados, extracción de información [1] y minería de datos [2] basados en técnicas de aprendizaje por computadora [3][4][5], en especial las basadas en redes neuronales artificiales [6][7], máquinas kernel [8], y árboles de decisión entre otras [9]. Así, la línea de investigación de aprendizaje automático por computadora es otra de las líneas que intervienen y direccionan el proyecto de investigación.

En el marco de este proyecto se han desarrollado, y se están desarrollando actualmente, varios sistemas de análisis y procesamiento de texto, entre los que se destacan:

- Software de Asistente de Creación de Corpus: es un software que permite construir material de entrenamiento para aplicaciones de minería de datos sobre texto no estructurado.

- Sistema de Mapeo de Datos: 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.

- Sistema de detección de similitudes en archivos de código fuente: Es un sistema que permite analizar archivos de código fuente escritos en diferentes lenguajes de programación e informar el grado de similitud entre los mismos.

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 impactan directamente en la efectividad del algoritmo de clasificación. Adicionalmente, el programa permite acelerar el tiempo necesario para la confección del material de entrenamiento, como así también brinda trazabilidad respecto de los expertos humanos que contribuyeron a cada parte del corpus. Esto permite establecer métricas y calcular la confidencia del material de entrenamiento construido.

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

El otro sistema es denominado Sistema de Mapeo de Datos (SMD) y se desarrolla con el objetivo de realizar la manipulación y procesamiento desde diferentes fuentes y orígenes de datos y almacenarlos en un repositorio común y centralizado med iante una base de datos en un motor SQL Server.

De esta manera, es posible obtener datos estructurados desde diferentes orígenes y registrarlos en un nuevo repositorio normalizado que permite facilitar el proceso de análisis y búsqueda sobre textos.

Un último sistema que se está desarrollando se denomina sistema de detección de similitudes en archivos de código fuente (SDS), y tiene como objetivo proporcionar métricas de similitud textual entre programas escritos en diversos lenguajes de programación.

Estas métricas principalmente son realizadas a nivel léxico y sintáctico.

2. Líneas de Investigación, Desarrollo e Innovación

La línea de investigación en las que se enmarca el proyecto de análisis de texto es el área de inteligencia artificial, específicamente la sub-área de lingüística computacional en las que se desarrollan aproximaciones a las problemáticas de extracción de información, paráfrasis y minería de datos en textos.

Algunas de las técnicas utilizadas son basadas en redes neuronales artificiales, y otras en árboles de decisión, entre otros algoritmos de aprendizaje supervisado utilizados.

En este contexto, se han desarrollado varios productos software que abordan una de las principales problemáticas de esta línea de investigación, que incluyen la construcción de corpus para diversas aplicaciones de análisis de texto, análisis, clasificación de información almacenada en formato estructurado para su posterior consulta y procesamiento, y software de

detección de similitudes de fragmentos comunes en archivos detexto.

La innovación se ve reflejada en los nuevos algoritmos que se elaboran para realizar el procesamiento de textos, muchos de ellos son utilizados en las herramientas desarrolladas, mientras que otros podrán ser utilizados posteriormente gracias al uso de los sistemas de análisis de texto expuestos en este artículo. Actualmente, se están desarrollando algoritmos de reconocimiento de paráfrasis que hacen uso del etiquetado de la información lingüística que se registra mediante el software de asistente de creación de corpus.

En la siguiente sección se presentan los productos software desarrollados hasta el momento.

3. Resultados

En esta sección se presenta una descripción de las herramientas realizadas explicando su motivación y resultados alcanzados hasta el momento.

3.1 Sistema de Mapeo de Datos:

Consiste de una aplicación de escritorio junto con una aplicación web, que permite manipular múltiples orígenes de datos estructurados, tales como archivos correspondientes a información contenida en diversos motores de bases de datos, y concentrarlos en un repositorio común (una base de datos sql-server), para poder aplicar posteriormente técnicas de análisis de textos.

Su construcción está motivada por la necesidad de procesar y analizar información distribuida en repositorios de datos diferentes. Como ventaja destacamos la centralización de la información, y como desventaja observamos que este procedimiento debe realizarse en diferentes instantes de tiempo a los efectos de mantener actualizada la información en el repositorio.

La Figura 1 muestra la interfaz principal de la aplicación de mapeo de datos.

Figura 1. Pantalla de Mapeo de Datos

3.2 Asistente de Creación de Corpus:

Este software permite la lectura y edición de corpus en los formatos provistos por el NIST (National Institute of Standards and Technology)[2] y por el CLEF (Cross Evaluation Language Forum)[3]. Al mismo tiempo, es posible realizar un etiquetado adicional sobre estos corpus agregando información lingüística.

También permite trabajar con corpus multilingües, es decir con pares de texto en idiomas diferentes.

También es posible 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.

En la figura 2 puede observarse la interfaz principal de esta herramienta que permite realizar una construcción semiautomática de corpus para facilitar la tarea de los anotadores humanos.

[2] www.nist.gov

[3] http://www.clef-initiative.eu/

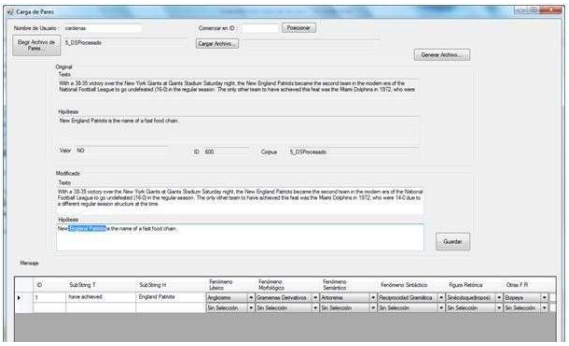

Figura 2. Asistente de Creación de Corpus

3.3. Sistema de detección de similitudes:

Se está desarrollando una herramienta software con el propósito de medir similitudes entre archivos de código fuente entre diferentes lenguajes de programación.

La misma consta de una interfaz gráfica que facilita la inspección de múltiples archivos, y ayuda a encontrar aquellos archivos con mayores similitudes. Por el momento las similitudes son evaluadas a nivel léxico-sintáctica.

Este software hace uso de diferentes herramientas de detección de similitudes y las resume en porcentajes. A su vez permite la comparación exhaustiva entre un conjunto de archivos y brinda un ranking de similitud que puede ser utilizado en la reutilización de códigos fuentes, y en la detección de uso de patrones en el código fuente, entre otras aplicaciones.

Se prevé continuar ampliando y mejorando este sistema, migrándolo a un entorno web y reduciendo el tiempo de procesamiento de los archivos utilizando técnicas de paralelismo.

La Figura 3 muestra la interfaz principal del sistema de detección de similitudes en archivos de código fuente.

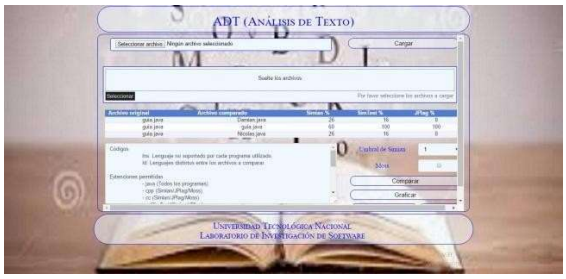

Figura 3. Software de detección de similitud entre archivos fuente

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:

- Actualmente el Dr. Julio Castillo está guiando a becarios de grado y de posgrado, como así también dirección de prácticas profesionales supervisadas y pasantías.
- Así mismo la Mg. Marina Cardenas está evaluando la posibilidad de desarrollar su tema de tesis de doctorado (en Ingeniería en Sistemas en la Universidad Tecnológica Nacional- FRC) en la misma temática con una variación del enfoque desde el punto de vista de los sistemas de Generación del Lenguaje Natural (NLG). Adicionalmente, realiza la guía de becarios de grado y de posgrado.
- También participan alumnos que realizan su práctica supervisada como parte de los requisitos para la obtención del grado de Ingeniero, haciendo aportes en el proyecto.
- Año tras año se capacita y forma a alumnos becarios que participan y aprenden desarrollando diversas tareas en el proyecto de investigación, lo que permite complementar su formación curricular desde el punto de vista científico.

Referencias

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

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

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