View metadata, citation and similar papers at core.ac.uk
brought to you by CO
provided by SEDICI - Repositorio de la U

Sistemas de análisis textual en formato no estructurado

Julio Castillo[1], Marina Cardenas[1], Martin Navarro[1],
Nicolas 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 artículo se describen las actividades desarrolladas y los subsistemas que conforman el proyecto de investigación denominado *Desarrollo de Sistemas de Análisis de Texto*.

Este proyecto aborda la problemática del desarrollo de herramientas que permitan recolectar, tabular, y etiquetar, textos en diferentes formatos y de diferentes fuentes de información con el propósito de someterlos a un posterior análisis utilizando aprendizaje automático y técnicas de minería de datos. Además del desarrollo de estas herramientas, el proyecto contempla el desarrollo de sistemas de análisis de texto que puedan abordar problemas como el reconocimiento de paráfrasis, es decir identificar oraciones (o párrafos) que tengan el mismo significado, o bien identificar oraciones-párrafos que estén semánticamente relacionados entre sí mediante una relación de implicación.

Las líneas de investigación en la que se encuadra el proyecto es dentro de las áreas de lingüística computacional y de aprendizaje automático. En particular, el proyecto se enfoca en modelos que utilizan redes neuronales artificiales (RNA) para analizar y procesar textos no estructurados.

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

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 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).

Actualmente, el proyecto se encuentra dentro del grupo de investigación denominado Grupo de Inteligencia Artificial (o GIA) de la UTN-FRC.

Este grupo GIA nuclea proyectos de una línea de investigación relacionada al área de inteligencia artificial, redes neuronales artificiales, análisis y procesamiento de imágenes.

El grupo se conforma por doctores, ingenieros, licenciados, becarios y pasantes.

El proyecto de análisis de texto, junto con otros proyectos, la mayoría de los ellos surgidos en el laboratorio de investigación de software de la UTN-FRC, han dado origen a varias líneas de investigación consolidadas (teoría de autómatas y gramáticas formales, modelos de predicciones de fenómenos climatológicos, entre otros). Esta sinergia entre múltiples proyectos y líneas de investigación han llevado a la necesidad de creación de un nuevo grupo de investigación UTN el cual se encuentra en una etapa de formación, tal como se detalla en la sección de líneas de investigación, desarrollo e innovación.

1. Introducción

El proyecto denominado Desarrollo de Sistemas Análisis de Texto (ADT), aborda

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

dos grandes problemas claramente diferenciados. El primero, relacionado a la necesidad de obtener información para poder construir corpus lingüísticos. El segundo, relacionado con el desarrollo de sistemas que hagan uso de dichos corpus para encarar problemas complejos del lenguaje natural.

En este sentido, el 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, especialmente aquellas basadas en redes neuronales artificiales [6][7][8], máquinas kernel [9], deep learning [10][11], y árboles de decisión entre otras.

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

- 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.

- 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.

- Sistema de detección de similitudes en archivos de código fuente (SDS). Es un sistema que se está comenzando a desarrollar y que tiene como objetivo analizar archivos de código fuente escritos en diferentes lenguajes de programación e informar el grado de similitud entre los mismos.

2. Líneas de Investigación, Desarrollo e Innovación

La línea de investigación principal y específica de este proyecto es el abordaje de problemáticas de lingüística computacional utilizando aprendizaje automático.

De esto subyace la línea demarcada por la Lingüística de Corpus [12], entendida como el estudio empírico de la lengua a partir de los datos que proporcionan ejemplos reales de producciones lingüísticas (orales o escritas) almacenadas en una computadora.

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.

Muy emparentada a esta línea de investigación se encuentra otra línea de investigación relacionada a la construcción de modelos computacionales de predicción de incendios forestales y de fenómenos climatológicos. En esta línea también participan integrantes de este proyecto.

La confluencia del trabajo de varias líneas de investigación (teoría de autómatas y gramáticas formales, modelos de predicciones de fenómenos climatológicos, y el modelado de problemas del área de ciencias sociales), consolidadas en el tiempo han llevado a la creación formal de un nuevo grupo UTN de investigación. Actualmente, la aprobación de dicho grupo se encuentra en trámite dentro de la UTN.

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 (SMD), y un Sistema de detección de similitudes en

archivos de código fuente (SDS). Estos sistemas se describen con más detalle a continuación.

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.

Adicionalmente, el ACC permite registrar diversos fenómenos lingüísticos a nivel léxico, sintáctico, morfológico y semántico. Se trata de un software que ya está desarrollado pero al cual se le continúan agregando nuevas funcionalidades. Entre los resultados logrados por este asistente podemos destacar:

- Lectura de corpus: Se realiza la lectura de corpus del NIST (National Institute of Standards and Technology) para su posterior generación, tabulación, ordenamiento y etiquetado, como así también la traducción del material al español utilizando el traductor automático de Google Translate.

- Carga de pares del corpus.

- Búsqueda y posicionamiento de un par dentro del corpus.

- Selección de subcadenas de fragmentos de texto para someterlos a una posterior clasificación: esto permite seleccionar partes de un texto y visualizarlas gráficamente a través de una tabla para su posterior modificación.

- Clasificación de los fenómenos en categorías y subcategorías.

- Creación de un nuevo corpus etiquetado almacenado en formato .xml .

Por otra parte, 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 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.

A la fecha se han creado tres corpus monolingües. Un corpus consta de 50 pares en ingles, los otros dos corpus están en español, cada uno presenta 100 pares de elementos etiquetados con información lingüística, de acuerdo a en los que se describe en [13] y con las clasificaciones enumeradas en [14].

Como herramienta, el ACC ha contribuido a los objetivos del proyecto proveyendo de material de entrenamiento tanto en el idioma español, como en el inglés. Esto ha facilitado y mejorado el funcionamiento de los Sistemas de RTE (Implicación Textual).

Actualmente, el ACC se sigue utilizando en la generación de corpus, y se está estudiando la posibilidad de dejarla disponible para el acceso libre de otros investigadores que deseen hacer uso de la misma en sus trabajos.

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.

Finalmente, mencionamos el sistema de detección de similitudes en códigos fuente (SDS). Este software está en sus primeras etapas y presenta como objetivo el permitir cuantificar el grado de similitud entre dos archivos de texto plano, en particular dos archivos de código fuente escritos en el mismo lenguaje de programación.

Hasta el momento, se ha desarrollado un prototipo que incluye una simple interfaz gráfica en la que es posible seleccionar dos archivos de códigos fuentes para aplicarles medidas principalmente de similitud léxica, y se está integrando una medida de similitud sintáctica.

El SDS también permite la comparación del tipo 1-N, en la cual un código fuente es comparado con N-elementos de un conjunto. Esta operación demanda un alto costo computacional, por lo que se están estudiando y desarrollando técnicas basadas en paralelismo para poder encontrar las similitudes de todos los elementos de un conjunto e informarlas de manera ordenada en función del grado de similitud.
En este caso, una detección de similitud que contemple a todos los elementos de un conjunto de N-elementos requerirá (N*N)/2 comparaciones entre los archivos fuente. Es por ello, la necesidad de contar con algoritmos eficientes de detección de similitud.

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, como así también realiza la dirección de prácticas profesionales supervisadas y pasantías.
- Una magister en ingeniería en sistemas de información que está en la etapa de escritura del plan de tesis para iniciar su doctorado en ingeniería en las temáticas de la línea de investigación mencionada. La unidad académica de radicación del doctorando sería la Universidad Tecnológica Nacional, Facultad Regional Córdoba.
  También realiza la dirección de becarios de posgrado y de becarios de grado en el contexto del presente proyecto.
- Anualmente participan en el proyecto, entre dos y cuatro alumnos realizando su práctica supervisada, la cual es necesaria como parte de los requisitos para la obtención del grado de Ingeniero.
- El proyecto además posee investigadores en formación y en proceso de categorización.
- Año tras año se capacita y forma a alumnos becarios que participan en el proyecto y que realizan actividades de investigación, complementando de esta manera su formación curricular desde el punto de vista científico.
- 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] 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. Corpus-Based Approaches to Syntax and Lexis, Berlin: Mouton, pág. 117, 2006.

[13] 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.

[14] 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.