Software para asistencia en la creación de corpus para sistemas de análisis de texto no estructurado

Julio Castillo[1], Marina Cardenas[1], Adrián Curti[1], Osvaldo Casco[1]

[1] Laboratorio de Investigación de Software/Dpto. Ingeniería en Sistemas de Información/ Facultad Regional Córdoba/ Universidad Tecnológica Nacional
{ jotacastillo, ing.marinacardenas, adriancx84, casqui.159}@gmail.com

Resumen

En este proyecto se busca utilizar técnicas de aprendizaje automático (machine learning), especialmente utilizando Redes Neuronales Artificiales (RNA) para analizar texto en lenguaje natural (por ejemplo un artículo de diario) y en base a ello determinar la existencia de texto (oraciones o párrafos) que tengan el "mismo sentido" es decir que presenten la misma semántica, o bien oraciones/párrafos que estén semánticamente relacionadas entre sí. Estos problemas son comúnmente conocidos como identificación de paráfrasis e implicación de textos, respectivamente. El fenómeno de paráfrasis puede pensarse como un caso particular de la implicación, que ocurre cuando la misma es bidireccional.

En el presente trabajo se describen dos de las aplicaciones desarrolladas para dar soporte al proyecto en las actividades de elaboración de material de entrenamiento para sistemas de minería de datos sobre texto no estructurado.

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

Contexto

El presente proyecto se encuentra consolidado dentro de la línea de investigación relacionada con lingüística computacional y es llevado a cabo en el Laboratorio de Investigación de Software del Dpto. de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba (UTN-FRC), siendo acreditado por la Secretaría de Ciencia, Tecnología y Posgrado de la UTN.

Nuestro trabajo se enmarca dentro del Grupo GIA (Grupo de Inteligencia Artificial) de la UTN-FRC, el cual tiene como objetivo general el investigar técnicas, algoritmos de inteligencia artificial, entre los que se destacan el estudio de las redes neuronales, autómatas celulares, análisis y procesamiento de imágenes, minería de datos, y su aplicabilidad y resolución de problemas de las ciencias naturales y de las ciencias sociales. El grupo está integrado por doctores, ingenieros, licenciados, becarios y pasantes.

De esta manera, se puede observar que se investigan técnicas de IA (Inteligencia Artificial) tanto desde el punto de vista teórico, como desde el punto de vista práctico.

Introducción

Mediante este proyecto se propone abordar el problema de extracción de información y minería de datos en textos no estructurados [1][2][3][4][5] mediante técnicas aprendizaje por computadora (machine learning), en especial las basadas en redes neuronales artificiales [6][7][8], por lo cual abarca también la línea de investigación de Machine Learning, y más específicamente de Aprendizaje Supervisado.

En el marco de dicho proyecto se desarrolló el programa Asistente de Creación de Corpus (ACC) con el objetivo

de proveer de material de entrenamiento para aplicaciones de minería de datos sobre texto no estructurado y también un Programa de Mapeo de Datos (PMD).

La generación y análisis del material de entrenamiento para un sistema de análisis de texto no estructurado es una tarea muy ardua y artesanal para ser realizada por un humano, razón por la cual en este proyecto se ha propuesto construir un programa que permita ayudar en la construcción semiautomática de corpus a los anotadores humanos.

Para poder construir este software se investigaron diversos fenómenos lingüísticos y se los clasificaron en base al tipo de fenómeno presente en un fragmento de texto. En base a ello se identificaron y clasificaron en Fenómenos Léxicos, Morfológicos, Semánticos, y Sintácticos.

La caracterización de estos fenómenos ayudó al diseño del software permitiendo orientar la funcionalidad con el objetivo de facilitar la identificación y clasificación de los mismos por parte del experto anotador humano.

El ACC tiene como objetivos:

- Permitir la clasificación de pares de texto con paráfrasis y facilitar la lectura y estudio de corpus.
- Proveer de un medio semiautomático que sirva de herramienta a los usuarios para sistematizar e identificar los diferentes fenómenos lingüísticos presentes en los textos.
- Generar un corpus etiquetado. La utilidad de un nuevo corpus etiquetado es vital, ya que servirá como material de entrenamiento a algoritmos de aprendizajes supervisados implementados en el proyecto, y también servirá como material para su aplicación otras subáreas de la Inteligencia Artificial.

En cambio el PMD, tiene como objetivo realizar la manipulación de diferentes fuentes y orígenes de datos y almacenarlos en una estructura estándar en una base de datos estructurada que trabaja sobre SQL Server.

Esta aplicación permite tomar datos de orígenes de datos estructurados y registrarlos en nuestro origen de datos que se encuentra normalizado para facilitar la búsqueda y análisis de textos.

Líneas de Investigación, Desarrollo e Innovación

En este proyecto se aborda la problemática de extracción de información y minería de datos en textos no estructurados mediante técnicas basadas en redes neuronales artificiales, por lo cual uno de los ejes de la línea de investigación se centra en Machine Learning, y más específicamente en Aprendizaje Supervisado.

Sin embargo, la principal línea de investigación dentro de la cual se encuentra inmerso el proyecto descripto en el presente, es básicamente la Lingüística Computacional. La misma abarca un campo científico interdisciplinar relativamente nuevo cuyo principal objetivo radica en la incorporación de la capacidad de reconocer y comprender el lenguaje natural humano en computadoras a través de modelos computacionales.

Sin embargo el software desarrollado que se describe en el presente, pretende abordar una de las principales problemáticas de la línea de investigación de la Lingüística Computacional que es la disposición de material que pueda ser utilizado para el aprendizaje de una computadora, ya sea como material de entrenamiento o como material de pruebas. De esto subyace la línea demarcada por la Lingüística de Corpus [9], 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.

Resultados

Con el objeto de realizar un análisis exploratorio y una primera aproximación al análisis de textos sobre textos

estructurados, se desarrolló una aplicación web basada en Flash que permite manipular diferentes fuentes y orígenes de datos y almacenarlos en una estructura estándar en una base de datos SQL Server.

Esta aplicación permite tomar datos de orígenes de datos estructurados y registrarlos en un único origen de datos estructurado, normalizado para facilitar la búsqueda y análisis de textos.

A continuación se pueden observar algunas de las interfaces principales de la aplicación:

Figura 1. Pantalla de Conexión con Bases de Datos

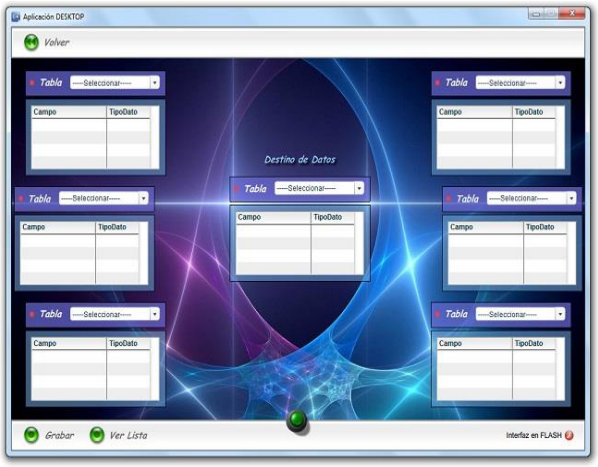

Figura 2. Pantalla de Mapeo de Datos

En la Fig1. puede observarse la selección del origen y destino de datos estructurados, mientras que en la Fig.2. puede observarse la interfaz diseñada para realizar el mapeo y transformación de datos de manera interactiva.

El fin último de este módulo es contar con un repositorio de información estructurada de modo tal que facilite y permita el adecuado procesamiento de la información y poder utilizar las técnicas que se vienen desarrollando en el proyecto para texto no estructurado.

El análisis y la identificación de los fenómenos lingüísticos descriptos en la introducción se utilizaron para la creación del software "Asistente para la creación de corpus".

Concretamente, el software desarrollado permite:

- Lectura de corpus: Para la obtención del corpus se realizó un módulo que permitió tomar como base corpus provisto por el 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 y luego se refinó las traducciones por traductores humanos que revisaron y corrigieron algunos detalles sintácticos y semánticos de las traducciones automáticas.

- Carga de pares del corpus.

- Búsqueda y posicionamiento de un par dentro del corpus.

- Selección de subcadenas de fragmentos de texto con el objeto de 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 definidas previamente.

- Almacenar en archivos las salidas de este nuevo c etiquetado.

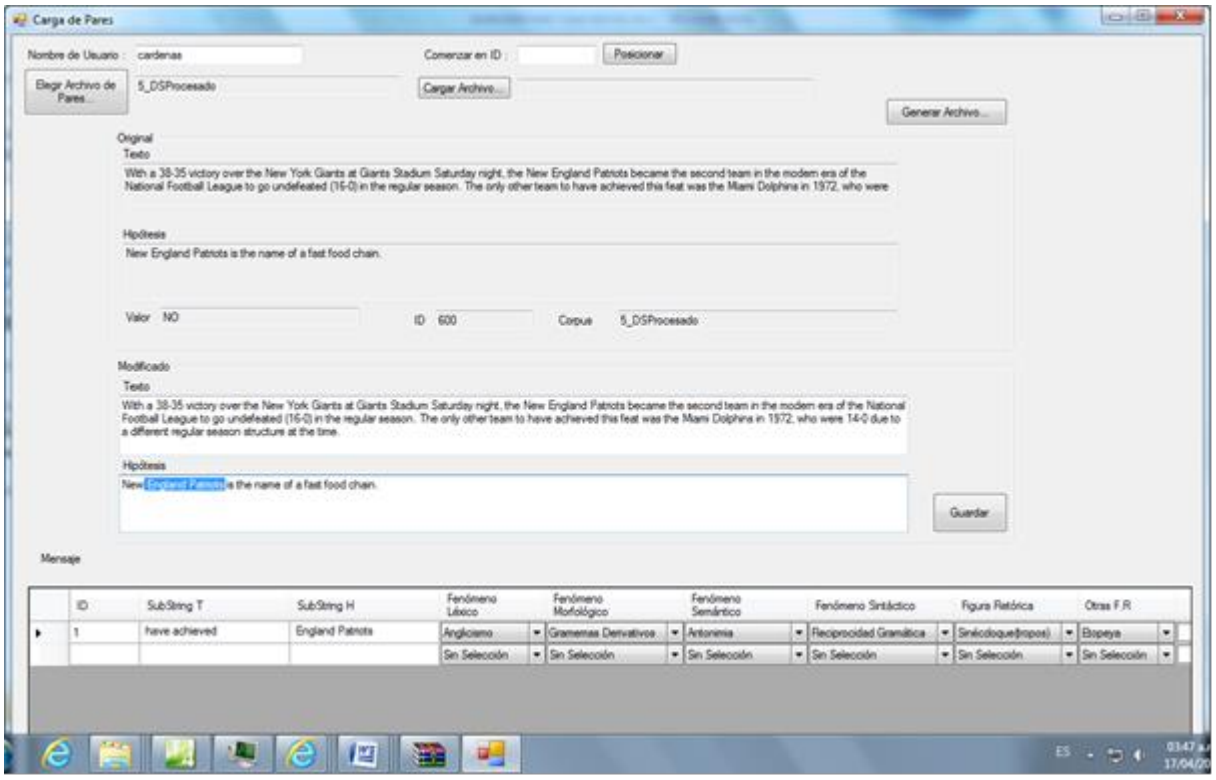

Figura 3. Interfaz principal de carga y clasificación de pares de un corpus.

En la Fig. 3 se muestra la interfaz principal del sistema asistente.

Como resultado del desarrollo del proyecto se ha obtenido un programa que permite ayudar en la construcción semiautomática de corpus para los anotadores humanos.

El desarrollo de esta herramienta contribuye a los objetivos del proyecto en el sentido que provee de material de entrenamiento tanto en el idioma español, como en el inglés. Esto facilita y mejora el funcionamiento de los Sistemas de RTE (Implicación Textual) en la medida que se entrena con mejor material de entrenamiento para el idioma español.

Se prevé continuar usando la herramienta para la generación de mejores corpus, contemplando la posibilidad de dejarla disponible para el acceso libre de otros investigadores del área que deseen hacer uso de la misma para sus trabajos.

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 Ing. Julio Castillo está desarrollando su tesis de doctorado en Ciencias de la Computación en la Universidad Nacional de Córdoba en la temática en la cual se encuadra el presente proyecto, lo que esperamos que contribuya un aporte tanto a nivel académico como curricular en su formación de posgrado.
- Así mismo la Ing. 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).
- También participan alumnos que realizan su práctica supervisada como parte de los requisitos para la obtención del grado de Ingeniero, haciendo aporte 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 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] D. Lin y P. Pantel. DIRT - Discovery of Inference Rules from Text. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 323–328, 2001.

[4] C. Monz y M. de Rijke. Light-Weight Entailment Checking for Computational Semantics. Inference in Computational Semantics (ICoS-3), págs. 59–72, 2001.

[5] Appelt, Douglas E., Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama y Mabry Tyson. The SRI MUC-5 JV-FASTUS Information Extraction System', Proceedings, Fifth Message Understanding Conference (MUC-5), Baltimore, Maryland, 1993.

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