Aplicaciones de Análisis de Información Textual: Corpus Lingüísticos

Julio Castillo, Marina Cardenas

Laboratorio de Investigación de Software LIS, Dpto. Ingeniería en Sistemas de Información
Facultad Regional Córdoba, Uni versidad Tecnológica Nacional

{ jotacastillo, ing.marinacardenas }@gmail.com

Resumen

En este artículo se describe un proyecto de investigación relacionado al análisis y procesamiento de información textual, tal como el reconocimiento de paráfrasis o la implicación de textos.

En ese contexto se describe la creación de una herramienta para construir corpus lingüísticos que pueden ser utilizados como material de entrenamiento para sistemas de minería de datos y de extracción de información, en especial sobre texto no estructurado.

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

Contexto

El proyecto denominado análisis de texto no estructurado (ADT) 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

Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba.

Por otra parte, este proyecto se encuentra dentro del grupo de investigación denominado Grupo de Inteligencia Artificial (o GIA) de la UTN-FRC.

El grupo de investigación nuclea diferentes proyectos de investigación que se hayan todos en temáticas concernientes a la inteligencia artificial entre las que podemos destacar análisis de imágenes, algoritmos evolutivos, y su aplicabilidad en problemas de la ingeniería, de las ciencias naturales, y de las ciencias sociales.

Este grupo está compuesto de becarios, pasantes, docentes investigadores y doctores.

1. Introducción

Mediante este proyecto se propone abordar la detección de similitudes entre diferentes textos. Si bien se trata de un área de investigación compleja, se han llevado a cabo varias aproximaciones tendientes a clasificar si dos fragmentos de textos se relacionan entre sí, y en ese caso, qué grado de relación presentan entre sí [1][2].

A los efectos de poder abordar este tipo de problemas, es necesario contar con conjunto de datos (material lingüístico) apropiado para poder construir algoritmos de aprendizaje sobre los mismos [3][4][5].

Por ello, la creación de una herramienta que pueda ser utilizada por expertos humanos es de vital importancia para la construcción de un corpus. En el proyecto se ha desarrollado una herramienta que permite procesar y etiquetar diferentes conjuntos de datos.

Entre los objetivos generales de construcción de este asistente podemos enumerar tres principales:

- Proveer un mecanismo de clasificación de pares de texto con paráfrasis y

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

facilitar la lectura y estudio de otros corpus.

- Contar con una herramienta que permita automatizar la clasificación de fragmentos de textos en base a la presencia o ausencia de ciertos fenómenos lingüísticos que se mantengan en el mismo texto o que se mantenga entre dos fragmentos de textos.
- Generación de corpus etiquetado. El corpus permite incrementar la capacidad de reconocimiento y de clasificación de los algoritmos utilizados.

La construcción de material de entrenamiento involucra a anotadores humanos por lo cual el proceso de construcción es costoso, lento y a menudo es fuente de errores. La herramienta desarrollada permite automatizar la creación y acelerar el proceso de etiquetado, a la vez que posibilita realizar una trazabilidad entre los anotadores humanos y conocer su aporte en la construcción del corpus. Como consecuencia de agilizar el proceso de etiquetado es posible bajar los costos asociados en suconstrucción.

La información lingüística asociada a cada corpus creado dependerá del problema que se necesite abordar, pero el corpus debe ser consistente y cada elemento que lo constituya (texto o pares de fragmentos de textos) debe poseer el mismo tipo de anotaciones lingüísticas.

El tamaño del corpus que se puede crear puede ser tan grande como se requiera. La versión actual de la herramienta desarrollada soporta un máximo de archivos individuales de cuatro gigabytes de tamaño.

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, más concretamente una sub-especialidad que se denomina computación lingüística.

Los desarrollos de esta línea de investigación, lo constituyen, por un lado, las herramientas elaboradas para facilitar el análisis y procesamiento de archivos de texto, y por el otro, los sistemas de reconocimiento de implicación o de reconocimiento de paráfrasis entre dos fragmentos de textos.

La innovación del proyecto concierne a los nuevos métodos propuestos para el análisis y procesamiento de textos, como así también a los algoritmos creados para abordar las problemáticas anteriormente mencionadas. Los algoritmos diseñados aprovechan las diferentes características que se pueden aprender de los textos y que son recolectados y creados a partir de las herramientas de procesamiento de textos.

Son múltiples las posibles sub- disciplinas que podrían valerse de los resultados de este proyecto, entre las que podemos destacar a las tareas de recuperación de información, evaluación de las traducciones automáticas [6], evaluación de la calidad de las traducciones, reconocimiento de paráfrasis [7] e implicación de textos[8][9][10] . Adicionalmente, la creación de corpus es una actividad de relevancia y de impacto en las tareas relacionadas al procesamiento del lenguaje.

3. Resultados

Para poder construir una herramienta que permita construir corpus lingüísticos se investigaron y definieron el conjunto de fenómenos lingüísticos de interés. Se definieron cuatro tipos de fenómenos que son los que permite registrar el software asistente de creación de corpus.

La clasificación se definió en base al tipo de fenómeno presente en un fragmento de texto. Se identificaron y clasificaron en Fenómenos a nivel Léxico, Morfológico, Semántico, y Sintáctico.

Algunos de los fenómenos Léxicos que se registran son anglicismos, arcaísmos, barbarismos, cultismos, eufemismos, neologismos, entre otros fenómenos.

En cuanto a los fenómenos morfológicos se pueden registrar lexemas, morfemas y grafemas.

Los fenómenos sintácticos que se registran son anáforas, flexiones, pronombres, y concondancias entre otros fenómenos. Como fenómenos semánticos se registran la antonimia, homonimia, polisemia y sinonimia.

La caracterización de estos fenómenos facilita el proceso de etiquetado a los anotadores humanos, al tiempo que provee información lingüística de textos que luego pueden ser utilizados por diversos algoritmos de clasificación.

En cuanto al proceso de desarrollo, se comenzó con una especificación de los requerimientos funcionales y no funcionales, y posteriormente, se realizó el análisis, diseño e implementación. En todas las etapas se utilizó UML como lenguaje de modelado. La creación de esta herramienta siguió el proceso de desarrollo unificado, pero esta es la excepción más que la regla, debido a que normalmente en el proyecto se utilizan metodologías agiles.

El software construido posee las siguientes funcionalidades:

- Lectura de corpus de implicación de textos y de paráfrasis.
- Carga de nuevos pares del corpus.
- Búsqueda y posicionamiento de un par dentro del corpus.
- Selección de substrings de fragmentos de texto en base a una clasificación lingüística elegida por el anotador humano.
- Clasificación de los fenómenos en categorías y subcategorías definidas previamente.
- Generación de un nuevo corpus de datos. Estos corpus generados conforman un corpora lingüístico que constituyen un recurso necesario para muchas aplicaciones de análisis y procesamiento del lenguaje [11][12].

Uno de los corpus generados consiste en corpus de implicación de textos ampliados.

Los corpus ampliados han demostrado ser de utilidad en diversas tareas y ya que permiten incrementar la efectividad en la clasificación de textos [13].

Actualmente se está trabajando en extender las funcionalidades de esta herramienta para que sea capaz de informar:

- Aquellos pares de textos en los cuales hay disidencia en cuanto a su clasificación.
- Porcentaje de fragmentos de textos en los cuales hay coincidencias y disidencias.
- Consistencia del material de entrenamiento.
- Sesiones de usuario, para que los anotadores puedan suspender y continuar el proceso de etiquetado en el momento que lo deseen.
- Extensión de los tipos de archivos y corpus que es capaz de reconocer la herramienta.

4. Formación de Recursos Humanos

El equipo de investigación está formado por docentes investigadores del Laboratorio de Investigación de Software LIS[2] del Dpto. de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba, se detallan a continuación los responsables del proyecto:

- El Dr. Julio Castillo desarrolló su tesis de doctorado en Ciencias de la Computación en la temática de implicación de textos y paráfrasis, y coordina las actividades del proyecto y dirige a los integrantes miembros del equipo.
- La Mg. Ing. Marina Cardenas está evaluando la posibilidad de desarrollar

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

su tema de tesis de doctorado en la misma temática con una variación del enfoque desde el punto de vista de los sistemas de Generación del Lenguaje Natural (NLG), y dirige a los integrantes miembros delequipo.

- Participan del proyecto alumnos que necesitan realizar su práctica supervisada que es uno de los requisitos para la obtención del grado de Ingeniero. Los alumnos que intervienen aprenden a realizar actividades de investigación, y cómo integrarse en un equipo existente. También generalmente participan por año uno o dos becarios alumnos a los que se les enseña como trabajar en un proyecto de investigación. Adicionalmente, se prevé que becarios de investigación de posgrado puedan realizar actividades en el marco del presente proyecto.

5. Bibliografía

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

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

[3] Judith K lavans and Philip Resnik. The Balancing Act. Combining Symbolic and Statistical Approaches to Language. MIT Press. 1996.

[4] C. Manning and H. Sch¨utze. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA, 1999.

[5] D. Lin and 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.

[6] Castillo, Julio and Estrella, Paula. Semantic textual similarity for MT evaluation. Proceedings of the Seventh Workshop on Statistical Machine Translation. WMT '12. 2012.

[7] Dolan, W., Brockett, C., Castillo, J. and Vanderwende, L. (2010). Mining phrase pairs from an structured resource. WO/2010/135204. 2010.

[8] C. Monz and M. de Rijke. Light-Weight Entailment Checking for Computational Semantics. In P. Blackburn and M. Kohlhase, editors, Inference in Computational Semantics (ICoS-3), pages 59–72, 2001.

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

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

[11] Lewis, D. 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, 1995), pp. 246254. 1995.

[12] M. Craven and J. Shavlik. Using Neural Networks for Data Mining. *Future Generation Computer Systems*, 13, pp. 211-229. 1997.

[13] Stefan. T; Stefanowitsch A. (2006). Corpora in Cognitive Linguistics. Corpus - Based Approaches to Syntax and Lexis, Berlin: Mouton, pág. 117. 2006.