ARTÍCULO SELECCIONADO DEL CONAIISI

Creación de corpus para aplicaciones de análisis de texto no estructurado

Dr. Julio Castillo[1]
Ing. Marina Cardenas[2]
Ing. Adrian Curti

[1]Email: jotacastillo@gmail.com, jcastillo@sistemas.frc.utn.edu.ar .
[2]Email: ing.marinacardenas@gmail.com

Universidad Tecnológica Nacional
Facultad Regional Córdoba

RESUMEN

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

El desarrollo de dichas aplicaciones está motivado en la necesidad de generar y utilizar corpus que sirva en la fase de entrenamiento de técnicas de aprendizaje automático. En particular estas herramientas fueron concebidas para que puedan ser utilizadas en las tareas de traducción automática, y en generación de paráfrasis.

ABSTRACT

This paper describes two of the applications developed to support activities of building training materials for data mining systems over unstructured text.

The development of such applications is motivated by the need to generate corpus for the training phase of machine learning techniques. As special case, these tools can be used in the task of machine translation and paraphrase generation.

Palabras clave: creación de corpus, material de entrenamiento, aprendizaje automático

1. INTRODUCCIÓN

Las herramientas que se describen en este artículo han sido desarrolladas con el objetivo de proveer un mecanismo de automatización del proceso de construcción de material de entrenamiento que es necesario para los sistemas basados en aprendizaje por computadora.

Se pretende que los materiales de entrenamiento producidos permitan abordar problemas de extracción de información y minería de datos en textos no estructurados [1][2][3][4][5] mediante técnicas de aprendizaje por computadora (machine learning), en especial las basadas en redes neuronales artificiales [6][7][8]. En particular, estas herramientas fueron diseñadas para ser empleadas cuando se utilicen técnicas de aprendizaje supervisado.

La generación y análisis del material del entrenamiento para un sistema de análisis de texto no estructurado es una tarea muy ardua y artesanal para ser realizada por un humano, lo cual motiva la construcción de programas que permita ayudar a los anotadores humanos en la construcción (semiautomática) de corpus [9][10].

Por otra parte, contar con un adecuado y completo material de entrenamiento es de

vital importancia para sistemas basados en aprendizaje automático.

Por ello, se han desarrollado dos programas que pueden ser utilizados para proveer material de entrenamiento a aplicaciones de minería de datos sobre texto no estructurado. En las siguientes secciones se describen las principales características de los mismos.

Uno de los problemas que se pretende abordar con este material de entrenamiento es el problema de las evaluaciones de las traducciones automáticas. La traducción automática es el área que estudia el problema de realizar una traducción desde un idioma fuente a otro idioma distinto denominado idioma destino, con la ayuda de programas de computadora principalmente basados en traducción automática estadística [11][12][13]. En ese contexto, la evaluación de las traducciones automáticas consiste de desarrollar mecanismos que permitan realizar ranking entre diversos sistemas de traducción automática [14][15], con el objetivo de determinar cuál es el traductor más adecuado. Este ranking se realiza contrastando las traducciones obtenidas con las traducciones realizadas con un evaluador humano [16].

Este artículo presenta en la Sección 2 el programa asistente de creación de corpus, y posteriormente, la Sección 3 describe la herramienta de Mapeo de Datos. Finalmente, la Sección 4 resume las conclusiones.

2. PROGRAMA ASISTENTE DE CREACIÓN DE CORPUS

Para poder construir el programa Asistente de Creación de Corpus (ACC) 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:

• Proveer de un medio semiautomático que sirva de herramienta a los usuarios para sistematizar e identificar los diferentes fenómenos lingüísticos presentes en diversos textos.

• Permitir la clasificación de pares de texto con paráfrasis y facilitar la lectura y estudio del corpus.

• 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 en otras subáreas de la Inteligencia Artificial.

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) y por el CLEF (Cross Evaluation Language Forum) [11] para su posterior generación, tabulación, ordenamiento y etiquetado, como así también la traducción del material al español utilizando el traductor automático de GoogleTranslate y luego se refinaron las traducciones por traductores humanos que revisaron y corrigieron algunos detalles sintácticos y semánticos de las traducciones automáticas.

• Carga de pares (texto e hipótesis) 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.

• Almacenamiento en archivos de las salidas de este nuevo corpus etiquetado.

• Almacenamiento gradual. Esto significa que no es necesario que el anotador huma-

no realice la etiquetación del corpus completo durante una sesión de trabajo, lo puede hacer en sucesivas sesiones e ir guardando sus avances en forma progresiva.

• Identificación de usuarios. Esto posibilita la identificación de qué tipo de fenómeno fue seleccionado por un usuario determinado. Esta información es útil para poder determinar la eficiencia de clasificación de los anotadores, permite conocer la trazabilidad de los mismos.

Se definieron cuatro categorías de fenómenos y, posteriormente, se reclasificaron permitiendo la aparición de subcategorías en cada uno de estos fenómenos lingüísticos. Por ejemplo, dentro de la categoría de Fenómenos Léxicos identificamos los siguientes subfenómenos: Anglicismos, Arcaísmos, Barbarismos, Cultismos, Eufemismos, Galicismos, Jerga (o algarabía), Neologismo, Tecnicismo, y Vulgarismo.

De la misma manera se procedió con las otras categorías de fenómenos.

En la Figura 1 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.

El uso de esta herramienta ha permitido ahorrar notable tiempo de procesamiento manual.

Experimentalmente se puso a prueba la velocidad en la clasificación del ACC en comparación con el uso de una planilla de cálculos para clasificar pares RTE. Para ello, se seleccionaron dos anotadores humanos y un subconjunto del corpus RTE4 provisto por el NIST de 100 pares.

En comparación con una planilla de cálculo, se ha podido comprobar que un mismo anotador humano pudo incrementar su velocidad de clasificación en un 40%, y al mismo tiempo, el ACC ayudo a eliminar los errores en los que incurrían los evaluadores humanos al registrar la clasificación dentro de una celda de la planilla de cálculo.

Esto es debido a que esta herramienta permite que el anotador se enfoque en un par de textos por vez, y así no incurra en errores de clasificación, y ayuda a que el anotador no se equivoque y asigne un valor de clasificación a otro par. Por otra parte, y como ventaja adicional del asistente de creación de corpus, se lleva registro de cada par clasificado por un anotador, y esto permite computar fácilmente estadísticos sobre el corpus, tales como el acuerdo inter-anotador para determinar el nivel de confidencia de clasificación de los anotadores humanos.

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.

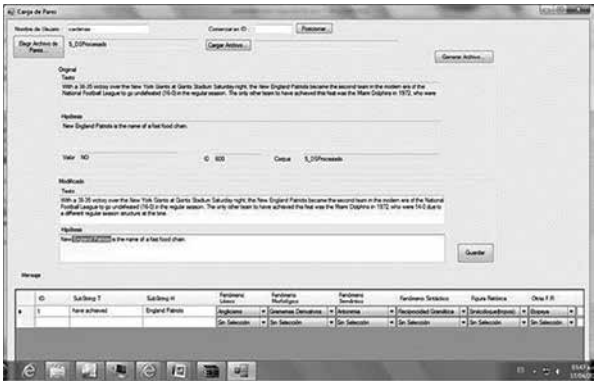

Figura 1. Interfaz principal de carga y clasificación de pares de un corpus

3. PROGRAMA DE MAPEO DE DATOS

El Programa de Mapeo de Datos (PMD) permite realizar un análisis exploratorio y una primera aproximación al análisis de textos sobre textos estructurados y 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.

El PMD está basado en una aplicación web que trabaja sobre Flash y almacena los datos normalizados en una estructura estándar de una base de datos SQL Server facilitando la búsqueda y análisis de textos.

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

En la Figura 2 puede observarse la selección del origen y destino de datos estructurados, mientras que en la Figura 3 se visualiza 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.

Figura 2. Pantalla de Conexión con Bases de Datos

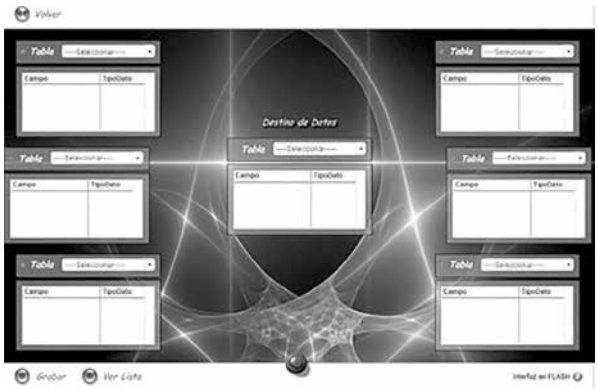

Figura 3. Pantalla de Mapeo de Datos

De esta manera, esta herramienta permite acceder a diferentes bases de datos (y orígenes de datos) y permite unir la información en un repositorio común. Este repositorio es en sí mismo, una base de datos relacional normalizada sobre la que se podrán efectuar consultas y realizar minería de datos. Dado que se almacena información de diferentes fuentes, es necesario realizar una actualización cuando la información cambia en los repositorios originales. Así, la manera de funcionar de este repositorio es análoga a la estructura con la que se diseña un cubo OLAP [17], lo cual permite posteriormente, su consulta, análisis y exploración.

4. CONCLUSIONES Y TRABAJOS FUTUROS

En este artículo presentamos dos herramientas para la manipulación de información no estructurada. La primera herramienta permite automatizar la creación de un corpus, mediante un programa de carga y clasificación.

La segunda herramienta, posibilita trabajar con múltiples orígenes de datos y registrarlas en una base de dato relacional, para poder realizar posteriormente tareas de minería de datos.

Ambas herramientas están siendo utilizadas en la creación de corpus, y permiten incrementar la productividad de anotadores humanos reduciendo el tiempo de clasificación y errores incurridos por los anotadores al momento de registrar una clasificación.

Como trabajo futuro se prevén extensiones de esta herramienta para poder manipular otros corpus y diversos orígenes de datos. Asimismo, se planea adaptar estas herramientas para que puedan ser utilizadas en la creación de corpus útiles para la tarea de evaluación de la calidad en las traducciones automáticas.

AGRADECIMIENTOS

Este proyecto de investigación es llevado a cabo en el Laboratorio de Investigación de Software (LIS) del Dpto. de Ingeniería en

Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba (UTN-FRC), con financiamiento de la Secretaría de Ciencia, Tecnología y Posgrado de la UTN.

REFERENCIAS BIBLIOGRÁFICAS

[1] KLALAVANSY, J; RESNIK, P. (1996). The Balancing Act. Combining Symbolic and Statistical Approaches to Language. MIT Press.

[2] MANNING, C; SCHUTZE, H. (1999). Foundations of Statistical NaturalLanguage Processing.The MIT Press, Cambridge, MA.

[3] CASTILLO J. (2009). 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] FELDMAN R.; HIRSH, H. (1996). Exploiting Background Information in Knowledge Discovery from Text.Journal of Intelligent Information Systems.

[7] LEWIS, D. (1995). Evaluating and optimizing autonomous text classification systems. In Proceedings of SIGIR-95, 18th ACMInternationalConference onResearch and Development in Information Retrieval. Seattle, US, págs. 246-254.

[8] CRAVEN, M.; SHAVLIK, J. (1997). Using Neural Networks for Data Mining. Future Generation Computer Systems, 13, págs. 211-229.

[9] CASTILLO, J.; CARDENAS, M.; CURTI, A; CASCO O. (2015). Software para asistencia en la creación de c orpus para sistemas de análisis de texto no estructurado. WICC 2015. San Luis, Argentina.

[10] STEFAN, T; STEFANOWITSCH, A. (2006). Corporain Cognitive Linguistics. Corpus Based Approaches to Syntax and Lexis, Berlin: Mouton, pág. 117.

[11] HE ,Y.; Du J.; Way, A. and Van J. (2010). The DCU dependency-based metric in WMT-MetricsMATR 2010. In: WMT 2010 - Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, ACL, Uppsala, Sweden.

[12] XIONG, D., LIU, Q., and LIN, S. (2006). Maximum entropy based phrase reordering model for statistical machine translation. In Proceedings of ACL-COLING. 521–528.

[13] CALLISON BURCH, C.; KOEHN, P; MONZ, C.; ZAIDAN, O. (2011).Findings of the 2011Workshop on Statistical Machine Translation. WMT 2011.

[14] COUGHLIN, D. (2003). Correlating Automated and Human Assessments of Machine Translation Quality. In MT Summit IX, New Orleans, USA pp. 23–27.

[15] CASTILLO J. (2008). The Contribution of FaMAF at QA@CLEF 2008. Answer Validation Exercise. 2008. In proceeding of CLEF 2008. September, Aarhus, Denmark.

[16] TURIAN, J.; SHEN, L. and MELAMED, I. D. (2003). Evaluation of Machine Translation and its Evaluation. Proceedings of the MT Summit IX, New Orleans, USA, 2003pp. 386–393.

[17] CODD, F; CODD, S; SALLEY, C. (1993). Technical Report. San Jose, Calif: Codd EF & Associates; 1993. Providing OLAP (Online Analytical Processing) to User-Analysts: An IT Mandate.