Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 351-352 recibido 1-05-2009; aceptado 5-06-2009

Construcción de un sistema PLN usando el *framework* UIMA *

Building a NLP system using the UIMA framework

Fermín Cruz, José A. Troyano, Fernando Enríquez, Víctor J. Díaz
Dep. de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Avda. Reina Mercedes s/n
41012 Sevilla
fcruz@us.es

Resumen: Este trabajo resume una experiencia de uso de UIMA (*Unstructured Information Management Architecture*), una plataforma que permite la creación e integración de aplicaciones que procesen información no estructurada como textos en lenguaje natural, audio o vídeo. La característica más importante de UIMA es la interoperabilidad ya que permite la integración de distintas herramientas y recursos. Presentamos un sistema que realiza diversas tareas PLN como el análisis morfosintáctico, el reconocimiento de entidades o el análisis de dependencias. En este caso, UIMA nos ha permitido integrar de forma muy cómoda cinco herramientas distintas (OpenNLP, TnT, FreeLing, Yamcha y Minipar).
Palabras clave: Información no estructurada, multimodalidad, herramienta de desarrollo, integración de componentes

Abstract: This paper describes our experience in using UIMA (*Unstructured Information Management Architecture*), a framework to create and integrate applications to process unstructured information like natural language texts, audio and video. The most important feature of UIMA is its interoperability, allowing the integration of diverse tools and resources. We present a system that performs several NLP tasks like POS tagging, named entity recognition or dependency parsing. In this system, we have been able to easily integrate five different tools with the support of UIMA (OpenNLP, TnT, Freeling, Yamcha and Minipar).
Keywords: Unstructured information, multimodality, develpment tool, component integration

1. Introducción

El artículo en el que se presentó la arquitectura UIMA (Ferrucci, 2004) comienza su argumentación comentando que IBM dispone de un grupo de unas 200 personas a lo largo de todo el mundo trabajando en los más diversos temas relacionados con el Procesamiento del Lenguaje Natural. Acto seguido afirma lo siguiente:

> An analysis of these activities strongly suggested that improving the organization's ability to quickly discover each other's results and rapidly combine different technologies and approaches would accelerate scientific advance. Furthermore, the ability to reuse and combine results through a common architecture and a robust software framework would accelerate the transfer of research results in NLP into IBM's product platforms.

Es lo que se puede considerar como el argumento principal sobre el que descansa la idea de la arquitectura UIMA. Nuestro grupo de investigación ha comenzado a utilizar UIMA como arquitectura de referencia en los sistemas PLN que desarrollamos. Con este trabajo pretendemos compartir nuestras experiencias y animar a los miembros de la comunidad de la SEPLN a experimentar con

* Parcialmente financiado por el proyecto CICYT HUM2007-66607-C04-04.

esta prometedora plataforma que, tal y como afirman sus autores, estamos seguros que ayudará a acelerar el avance científico.

2. El modelo UIMA

UIMA consta de una arquitectura que especifica cómo se puede diseñar un sistema que maneje información no estructurada en base a componentes y de un *framework* que proporciona un entorno de ejecución en el que integrar dichos componentes. La figura 1 muestra de forma simplificada los elementos más importantes de la arquitectura UIMA.

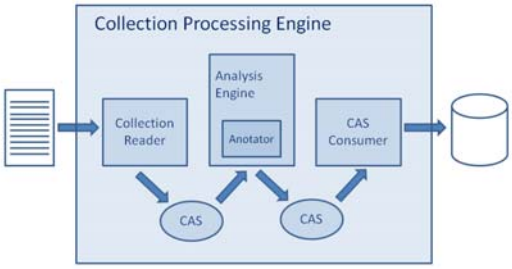

Figura 1: Arquitectura

Los tres conceptos más importantes del modelo UIMA son:

- *Analysis Engine (AE)*: son las piezas básicas de un sistema UIMA, ellas se encargan de procesar los documentos y de inferir información adicional.
- *Common Analysis Structure (CAS)*: establecen la manera en la que las AEs representan y comparten sus resultados.
- *Collection Processing Engine*: es el componente que establece cuál es el flujo de proceso de una aplicación.

3. Construcción de un *pipeline* básico

La aplicación más inmediata de UIMA es la implementación de componentes que realicen las tareas de análisis básicas en el PLN. Disponiendo de dichos componentes, resulta muy fácil construir un *pipeline* de procesamiento. Siempre que utilicemos un mismo sistema de tipos, con las mismas entradas y salidas para cada una de las tareas de análisis, podremos usar distintos componentes para cada una de estas tareas. Esto nos permitirá escoger en cada caso aquellas herramientas que se adapten mejor al problema que estemos resolviendo.

Las tareas básicas que hemos contemplado en nuestro *pipeline* son la tokenización y lematización, etiquetado morfosintáctico, detector de oraciones, reconocimiento de entidades con nombre, análisis sintáctico superficial y análisis de dependencias. En concreto, disponemos de los siguientes *Analysis Engines*:

- Detector de oraciones de OpenNLP.
- Tokenizador de OpenNLP.
- Etiquetador POS TnT.
- Analizador basado en FreeLing: incluye las tareas de tokenización, lematización, etiquetado POS y detección de oraciones.
- Reconocedor de entidades con nombre basado en Yamcha.
- Reconocedor de entidades con nombre de Stanford.
- Analizador sintáctico superficial basado en Yamcha.
- Analizador de dependencias basado en Minipar.

4. Conclusiones

Consideramos que la plataforma UIMA es una muy buena alternativa para el desarrollo de aplicaciones PLN ya que establece un modelo de integración de componentes que permite la combinación de soluciones propias con herramientas externas. Otro de los aspectos interesantes de esta herramienta es su aplicación a sistema multimodales, al estar diseñada para manejar cualquier tipo de información no estructurada ya sea texto, audio o vídeo. Por último es de destacar que esta herramienta ofrece un marco estupendo de colaboración a través del cual distintos grupos de investigación podrían compartir sus herramientas o desarrollar proyectos comunes.

Bibliografía

Ferrucci, D., Lally, A.: UIMA. An architectural approach to unstructured information processing in the corporate research environment. In *Natural Language Engineering*, vol 10, 327-348. Cambridge University Press (2004)