Quick Start on Collaborative Development of NLP tools

Rodrigo Agerri
IXA NLP Group
UPV/EHU
rodrigo.agerri@ehu.es

Abstract

Looking at real NLP tools in a collaborative environment.

1 Introduction

Usar herramientas de PLN a través de la IXA pipeline:
http://github.com/ixa-ehu

IDEAS:

1. **Uso**: Bajar, compilar y usar las herramientas de IXA pipeline: ixa-pipe-tok, ixa-pipe-pos, ixa-pipe-nerc.
   (a) Baja cada uno de los módulos y compilar siguiendo las instrucciones de cada uno de los README.
   (b) Compila cada módulo y aprende su funcionamiento con los textos de ejemplo proporcionados en la siguiente sección.
   (c) Encadena las salidas de los diferentes módulos para obtener una visión general de la cadena de procesamiento.
   (d) Para cada uno de los textos, detecta errores en cada nivel de anotación. Por ejemplo, compara las entidades nombradas automáticamente detectadas por ixa-pipe-nerc con las que tú anotarías.
2. **Desarrollo**: Usar un gestor de proyectos y un sistema de control de versiones para realizar, compartir y gestionar proyectos de software: por ejemplo, maven.apache.org y github.com:
   (a) Crear un proyecto simple con Maven ponerlo en control de versiones de github e invitar a otro programador al repositorio.
   (b) Crear un parser de URLs (incluyendo tiny URL de tweets, etc.) en el proyecto recién creado.

3. **PLN:** Hacer un clone de los repos ixa-pipe-tok e ixa-pipe-pos e ixa-pipe-nerc y examinar su funcionamiento en detalle.

   (a) Ver el Lemmatizador basado en diccionario y añadir una nueva funcionalidad: En lugar de devolver la palabra sin lematizar cuando la palabra de entrada no se encuentra en el diccionario, intentar sacar el lemma quitando los sufijos. Por ejemplo, "-ing" para "singing".

   (b) Crear una pequeña función(es) para detectar fechas, precios y expresiones horarias.

   (c) Se podría explotar la información de una lista (larga) de nombres de cosas y personas?

2 Textos

(1) The disappearance of York University chef Claudia Lawrence is now being treated as suspected murder, North Yorkshire Police said. However detectives said they had not found any proof that the 35-year-old, who went missing on 18 March, was dead. Her father Peter Lawrence made a direct appeal to his daughter to contact him five weeks after she disappeared. His plea came at a news conference held shortly after a 10,000 reward was offered to help find Miss Lawrence. Crimestoppers said the sum they were offering was significantly higher than usual because of public interest in the case.

(Example 1 of RTE-4)

(2) American photojournalist James Nachtwey in a file photograph from May 18 2003 as he is awarded the Dan David prize in Tel Aviv for his out standing contribution to photography. It was announced by Time magazine on Thurs day, 11 December 2003 that Nachtwey was injured in Baghdad along with Time magazine senior correspondent Michael Weisskopf when a hand grenade was thrown into a Humvee they were traveling in with the US Army. Both journalists are reported in stable condition and are being evacuated to a US military hospital in Germany.

Caption 1470132 of the ImageCLEF-09 dataset (Paramita et al., 2009)

(3) Nothing special really. Comfortable and clean but very boring decor in comparison to other NH hotels. I stayed in NH in Brussels and Zurich and I really liked them because of their modern and stylish design and big rooms. This one was just like any other hotel. Basic rooms with basic and dull decor - bit disappointing. The customer service was average. The rate was very expensive and I still had to pay for Internet and 20 euros for breakfast!!! It was good but way overpriced! The best thing about the hotel was the location - city centre, 2min from a metro stop.

(review from OpeNER project)

References

Paramita, M., Sanderson, M., and Clough, P. (2009). Diversity in photo retrieval: overview of the imageclefphoto task 2009. In *CLEF Working Notes 2009*.