Introducción al Procesamiento de Lenguaje Natural

Laboratorio 2016

El objetivo del laboratorio 2016 del curso Introducción al Procesamiento de Lenguaje Natural (PLN) es normalizar y anotar un corpus en español de sentencias del Poder Judicial. Para esto, el estudiante deberá familiarizarse con diferentes herramientas de PLN de acceso libre e incorporarlas en un entorno unificado de programación (en este caso, sobre la plataforma Python).

Corpus

El corpus a utilizar son sentencias de la Base de Jurisprudencia Nacional Pública del Poder Judicial de Uruguay, accesibles para consulta pública en http://bjn.poderjudicial.gub.uy.

El corpus está compuesto por 20.046 sentencias y tiene la siguiente estructura:

- El texto completo de la sentencia, generalmente extenso.
- Un texto de resumen, de tamaño variable, que puede ser vacío. En particular, 2.240 sentencias tienen resumen.
- La importancia de la sentencia (ALTA, MEDIA, BAJA).
- El tipo de la sentencia (DEFINITIVA, INTERLOCUTORIA)
- La sede judicial
- El tipo de procedimiento (recurso de casación, proceso civil ordinario, etc.)
- La fecha de la sentencia (dd/mm/aaaa)
- Un número ordinal (n/aaaa)
- Una ficha (n1-n2/aaaa)

Descripción del trabajo

Al comienzo del laboratorio, se dejará disponible el corpus, así como un notebook de IPython con los comandos para importarlo.

El objetivo del laboratorio es obtener una nueva versión, normalizada y anotada, del corpus, en formato CoNLL-U. Este formato (una versión revisada del formato utilizado para los *shared tasks* de la conferencia CoNLL) provee una representación estándar de textos, incluyendo diferentes anotaciones lingüísticas. El formato preve 10 campos por cada palabra, de los cuales, para esta tarea, deberán completarse: ID, FORM, LEMMA, UPOSTAG, FEATS, MISC, según se especifica a continuación.

a) Para la separación en oraciones y tokens, se seguirá exactamente el formato definido en la referencia, indicando una palabra por línea y separando las oraciones con líneas en blanco. Para el caso de palabras identificadas como contracciones, deberá indicarse un rango de tokens para la

contracción, y entradas individuales para cada una de las palabras que la componen (véase en la referencia el ejemplo que utiliza “del”).

b) El lemma y el POS-tag se obtendrán procesando el texto con Freeling (u otra herramienta para el español), y convirtiendo el tag obtenido a un Universal POS tag, de acuerdo a un mapeo que deberá establecerse. Las características morfológicas deberán agregarse en el campo FEATS, siguiendo en lo posible el inventario y la guía para extensiones presentes en la referencia del formato, y detallando los problemas/soluciones encontradas.

c) Finalmente, el campo MISC se utilizará para la identificación de entidades con nombre, incluyendo organizaciones, personas y localidades, en el formato IOB. Este formato le asigna una B al primer token de la entidad con nombre, y una I a los siguientes tokens de la entidad, mientras que asigna una O a cualquier otro token. Para distinguir entre los distintos tipos de entidades con nombre, se deberá usar (B-ORG, I-ORG) para una organización, (B-PER, I-PER) para personas y (B-LOC, I-LOC) para localidades.

Las tareas se realizarán utilizando herramientas de código abierto disponibles para el español (e.g. Freeling o NLTK), debiendo implementarse las conversiones necesarias para llevar los resultados al formato solicitado. Todo el trabajo deberá estar integrado en la plataforma Python 3.

Ejemplo de salida

```
Entrada: La sentencia dictada por la Sra. Juez Letrado de Primera Instancia del
Chuy:

Salida:

1       La              el              DET     Gender=Fem|Number=Sing  _   _   _   _   O
2       sentencia       sentencia       NOUN    Gender=Fem|Number=Sing  _   _   _   _   O
3       dictada         dictar          VERB    VerbForm=Participle     _   _   _   _   O
4       por             por             ADP     _                       _   _   _   _   O
5       la              el              DET     Gender=Fem|Number=Sing  _   _   _   _   O
6       Sra.            sra.            NOUN    Gender=Fem|Number=Sing  _   _   _   _   B-PER
7       Juez            juez            NOUN    Gender=Masc|Number=Sing _   _   _   _   I-PER
8       Letrado         letrado         ADJ     Gender=Masc|Number=Sing _   _   _   _   I-PER
9       de              de              ADP     _                       _   _   _   _   I-PER
10      Primera         1               ADJ     Gender=Fem|Number=Sing  _   _   _   _   I-PER
11      Instancia       instancia       NOUN    Gender=Fem|Number=Sing  _   _   _   _   I-PER
12-13   del             _               _       _                       _   _   _   _   _
12      de              de              ADP     _                       _   _   _   _   O
13      el              el              DET     Gender=Masc|Number=Sing _   _   _   _   O
14      Chuy            chuy            PROPN   Gender=Masc|Number=Sing _   _   _   _   B-LOC
```

Metodología de Evaluación

Una vez obtenido el corpus deberá diseñarse una metodología para la evaluación de los métodos de POS-tagging e identificación de entidades con nombre. Esto deberá incluir:

- Procedimiento para etiquetado manual del corpus.
- Definición de corpus de entrenamiento, testeo, y eventualmente desarrollo.
- Técnicas y medidas para la evaluación.

Herramientas

- Se utilizará la plataforma Python para el desarrollo, en su versión 3.3 o superior. En particular, los resultados se probarán sobre una distribución Anaconda, por lo que se sugiere instalarla.
- La biblioteca para procesamiento de lenguaje natural sobre Python a utilizar es NLTK
- Para la tokenización, tagging y cualquier tarea para el idioma español, se tiene la herramienta FreeLing, la que deberá integrarse al ambiente.
- Podrán utilizarse bibliotecas adicionales de Python, tales como: NumPy y SciPy, pandas o scikit-learn.

Podrán utilizarse herramientas y bibliotecas adicionales, siempre y cuando se integren al entorno Python de trabajo.

Formato de entrega

La entrega deberá realizarse utilizando un notebook IPython donde se incluirá tanto la documentación como el código Python a ejecutar. Asimismo, deberá entregarse una versión anotada del corpus (en un archivo comprimido).

Las tareas a realizar y documentar en el notebook son las siguientes:

- Importación del corpus.
- Análisis del corpus: deberán definirse propiedades y calcularse estadísticas de utilidad para "entender" el corpus, tanto desde un punto de vista cuantitativo como cualitativo.
- Tokenización, separación en oraciones y tagging, incluyendo las herramientas utilizadas y los principales problemas resueltos en el proceso.
- Identificación de entidades con nombre.
- Procedimiento para seleccionar una sentencia entre las 20.046 y se muestre interactivamente en el notebook la anotación completa.
- Descripción de la metodología de evaluación.
- Análisis cualitativo de los resultados obtenidos y principales desafíos a futuro.

Evaluación

Para la evaluación del laboratorio, se tendrá en cuenta:

- El cumplimiento de las tareas pedidas, incluyendo el código solicitado.
- La claridad de la documentación, en particular la justificación de las decisiones tomadas, así como el análisis de los resultados obtenidos.

Deberá presentarse en la entrega toda la información de configuración necesaria para poder reproducir el proceso de anotación automática. Esto será condición necesaria para la aprobación del curso.

Insumos

Se proveerán los siguientes archivos para la realización del laboratorio:

- corpus_pj.csv : corpus de las sentencias del Poder Judicial.
- Laboratorio_IntroPLN.ipynb : notebook IPython con una implementación inicial que carga el corpus de sentencias judiciales.

Referencias

- Bird, Klein, Lopper: “Natural Language Processing with Python”
- Referencia de NLTK
- The IPython notebook
- Freeling 4.0, sitio oficial