Recursos en euskera para la herramienta NLTK para enseñanza de procesamiento del lenguaje natural

Basque resources for the Natural Language Toolkit (NLTK)

Iker Manterola
Elhuyar Fundazioa
Zelai Haundi kalea, 3 - 20170 Usurbil
i.manterola@ehu.es

Arantza Diaz de Ilarraza, Koldo Gojenola, Kepa Sarasola
Euskal Herriko Unibertsitatea
Manuel Lardizabal 1, -20018 Donostia
kepa.sarasola@ehu.es

Resumen: Presentamos los recursos que hemos definido para adaptar las herramientas del paquete Natural Language Toolkit al euskera.
Palabras clave: Natural Language Toolkit, euskera, corpus, chunker, etiquetador

Abstract: We present the resources we have adapted in order to enable NLTK package to deal with text in Basque.
Keywords: Natural Language Toolkit, Basque, corpus, chunker, tagger

Natural Language Toolkit[1] (NLTK) es un paquete de herramientas y recursos libres para un espectro muy amplio de tareas dentro de procesamiento del lenguaje natural (PLN). En principio fue creado dentro de una iniciativa didáctica en PLN, de hecho, el libro recientemente editado en 2009 persigue dos objetivos: (1) enseñar los fundamentos en recursos, herramientas y aplicaciones de PLN, y (2) enseñar a programar en Python. El libro, la documentación y el material complementario on-line fueron creados en inglés y para tratar la lengua inglesa. Pero, ya que esta iniciativa va de la mano del software libre, no es de extrañar que año tras año han sido más las lenguas que se han ido incluyendo en torno a este proyecto. En la actualidad NLTK incluye utilidades para más de 10 idiomas entre la que se encuentran el español, catalán, portugués, y euskera.

El paquete integra diferentes tipos de corpus: texto simple, texto etiquetado con su categoría, etiquetado con sintaxis superficial, etiquetado con sintaxis profunda, PropBank, e incluso simples listas de palabras o léxicos. Los diferentes corpora accesibles se pueden cargar utilizando el paquete nltk.corpus. Cada corpus ofrece una serie de métodos para leer sus datos palabra por palabra, oración por oración, párrafo por párrafo, o por unidades de etiquetado.

Aparte de corpora, en NLTK también existen otros recursos menores para cada lengua tales como pequeñas gramáticas independientes del contexto, expresiones regulares para el reconocimiento de chunks, etc. que permiten obtener material inicial interesante tanto para docencia del PLN como para investigación.

En esta demostración mostramos los recursos, de corpus o demás, que hemos definido para adecuar el uso de algunos paquetes de NLTK para euskera. El principal recurso que hemos añadido es una adaptación del corpus CoNLL Shared Task on Dependency Parsing 2007. Se ha adecuado la parte correspondiente al euskara ofertada en la conferencia internacional CoNLL[2] (Conference on Computational Natural Language Learning). El corpus proporcionado[3] contiene 50.128 palabras (3.175 oraciones) etiquetadas manualmente con información sintáctica de dependencias (treebank). La inclusión de este corpus en NLTK posibilita la construcción automática de varios etiquetadores; por ejemplo, el tagger basado en unigramas que alcanza una precisión del 74,98%, y otro tagger que utiliza unigramas y bigramas y que alcanza hasta un 78,23%.

[1] http://www.nltk.org/

[2] http://depparse.uvt.nl/depparsewiki/SharedTaskWebsite

[3] Los árboles correspondientes a la oraciones se pueden visualizar en la página del proyecto Ancora: http://clic.ub.edu/ancora/

```
>>> tokens
['Espero',   'dut',   'hau',   'ondo',
'ibiltzea.', 'Nie', 'lagunak', 'ondo',
'egin', 'zuen', 'bidea']

>>> t2.tag(tokens)
   [('Espero','IZE_IZB'),('dut','ADL'),
('hau','DET_ERKARR'),
('ondo','ADB_ARR'),
('ibiltzea.','IZE_IZB'),
('Nie','IZE_IZB'),
('lagunak','IZE_ARR'),
('ondo','ADB_ARR'),('egin','ADI_SIN'),
('zuen','ADL'), ('bidea','IZE_ARR')]
```

Una versión del corpus CESS_eu adaptada al formato IOB, permite su uso en el desarollo progresivo y evaluación de *chunkers*. En unos pocos pasos se puede llegar a definir una gramática con una precisión apreciable (IOB accuracy: 47.3%; Precision: 30.0%; Recall: 23.9%; F-Measure: 26.6%).

```
>>> grammar = r"""
   PP:{<.*(GEN|GEL)>*<IZE><ADJ.*>*
        <DET.*>*<.*(ALA|INE|SOZ|ABZ)>}
   NP:{<.*(GEN|GEL)>*<IZE><ADJ.*>*
        <DET.*>*<.*(ABS|ERG|DAT)>}
      {<IZE.*>+} # chunk2
   VP: {<ADI>* <ADL>?}
       {<ADT>}
   """
```

Además, se han definido otros 51 recursos menores que posibilitan la ejecución en euskara de otros tantos ejercicios o experimentos del paquete didáctico NLTK[4]; por ejemplo concordancias, frecuencias, gramáticas, etc. El resultado obtenido en la adaptación al euskera de estas herramientas ha sido muy satisfactorio. Solo ha habido cinco excepciones en las que la calidad ha disminuido al adecuarlas al euskera, esto ha ocurrido cuando las herramientas NLTK no contemplaban que puede ser necesario el proceso de lematización de las palabras.

Paralelamente, el mismo trabajo también fue hecho para castellano, en este caso utilizando el corpus CoNLL 2007 en castellano (95.028 palabras y 3.512 oraciones). El corpus CoNLL 2007 de NLTK está dividido en cuatro subconjuntos: esp.test, esp.train, eus.test y eus.train.

```
>>> cl=nltk.corpus.conll2007
>>> cl.fileids()
['esp.test','esp.train','eus.test','eus
.train']
```

[4]http://www.nltk.org/

Usando este corpus se puede obtener, por ejemplo, un etiquetador para español que utiliza unigramas y bigramas y que alcanza hasta un 80,70% de precisión.

```
>>> cls[100]
['La',   'campaña',   'oficial',   'de',
'doce',  'décadas',  'comenzará',  'el',
'13_de_junio', '.']

>>> t2.tag(cls[100])
[('La',   'da'),   ('campaña',   'nc'),
('oficial',   'aq'),   ('de',   'sp'),
('doce',   'dn'),   ('décadas',   'nc'),
('comenzará',   'vm'),   ('el',   'da'),
('13_de_junio', 'w'), ('.', 'Fp')]
```

La adaptación a NLTK que hemos hecho con los corpus CoNLL 2007 en euskara y castellano se puede descargar desde el sitio web de NLTK[5]. El conjunto de los demás recursos no se distribuye con NLTK, pero puede ser descargado desde la página de productos del Grupo Ixa[6].

Bibliografía

Aduriz I., Aranzabe M., Arriola J., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., Urizar R. 2006 Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. *Corpus Linguistics Around the World. Book series: Language and Computers. Vol 56 (pag 1- 15). ISBN 90-420-1836-4 Ed. A. Wilson, P. Rayson, and D. Archer. Rodopi. Netherlands.*

Bird S., E. Klein, & E. Loper. 2009 Natural Language Processing with Python--- Analyzing Text with the Natural Language Toolkit O'Reilly Media. http://www.nltk.org/book

Manterola I. 2008 Lengoaia naturala irakasteko NLTK aplikazioa euskara landu ahal izateko zabaltzea. Karrera bukaerako Proiektua. Informatika Fakultatea. Euskal Herriko Unibertsitatea. Donostia.
https://ixa.si.ehu.es/Ixa/Produktuak/1235483470

[5]La página sobre corpus de NLTK ha tenido un error y presentaba este corpus como “corpus de catalán”.

[6]http://ixa.si.ehu.es/Ixa/Produktuak/1235483470