Ciencia de Datos y Lenguaje Natural

Clase 2.1 - Etiquetado gramatical

Grupo PLN - INCO

Universidad de la República

16 de agosto de 2022

Tradicionalmente 8 clases de palabras:

- clases léxicas
  - nombre *mesa azúcar chileno Uruguay*
  - verbo *está corrió comer*
  - adjetivo *suave chileno*
  - adverbio *rápido ayer*
- clases funcionales
  - conjunción *que y*
  - pronombre *que qué donde ella le la mi*
  - determinante *todo el esta*
  - preposición *a desde de*

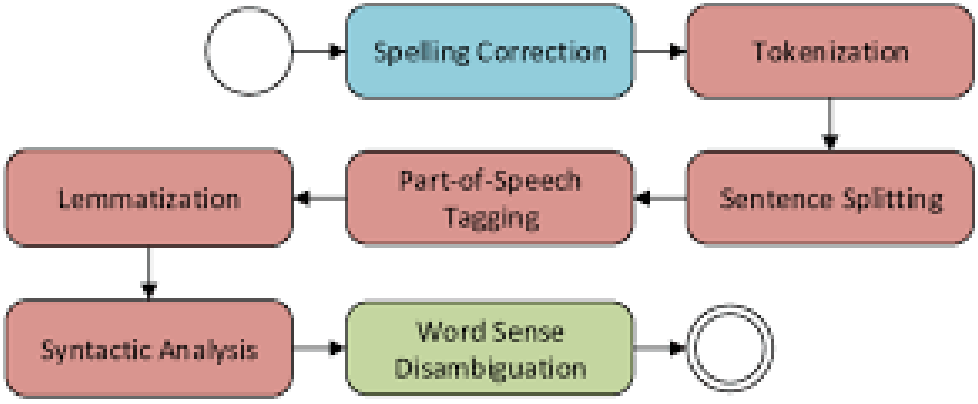
- Se han definido diversas subclasificaciones para algunas de estas clases
  - conjunción de subordinación *que,quien*
  - conjunción de coordinación *y, o, pero*
  - nombre contable *perro, banco, niño*
  - nombre de masa *azúcar, agua*
  - nombre común vs. nombre propio *río Andes*
- El reconocimiento automático de la clase de cada palabra de un texto es una tarea clásica en PLN

POS (Part Of Speech) tagging

Reconocer la categoría (el POS tag) puede ser útil en PLN

- Como base para un proceso posterior de reconocimiento sintáctico.
- Los nombres propios constituyen piezas de información importante en extracción de información.
- los verbos suelen asociarse al predicado central de un evento
- los adverbios dan muchas veces indicaciones de temporalidad

Pipeline clásico en PLN



Aplicaciones que se han resuelto 'a palabra'

- Análisis de sentimientos
- *bag-of-words*
- aprendizaje supervisado, *Naïf-Bayes*
- aprendizaje supervisado, con una red neuronal

Aplicaciones que se resuelven 'a carácter'

- Suena extraño, se pierde mucha información
- Se gana tambien: las palabras desconocidas no son un problema
- Se gana tambien: modelos mucho más pequeños
- En redes neuronales suele haber combinaciones: tokens + subwords embeddings

Etiquetado con categoría *(Part of Speech Tagging)*

Dada una oración, se trata de asignar la categoría gramatical (única en ese contexto) a cada una de las palabras de la oración.

El	gato	toma	leche	.
Det	Nom	Verb	Nom	Punt

- Es una tarea que requiere una salida palabra a palabra.
- Se dice que es una tarea secuencial
- hacerla en 2 pasos

¿Es difícil el tagging?

- ¿Cómo lo haríamos?
- Parece que sería necesario consultar un diccionario.
- O sea, tenemos un texto segmentado en palabras y le ponemos a cada palabra la categoría indicada por el diccionario.

¿Es difícil el tagging?

El	gato	toma	leche	.
el *DA0MS0* 1	gato *NCMS000* 1	toma *NCFS000* 0.551913	leche *NCFS000* 1	. *Fp* 1
		tomar *VMIP3S0* 0.442623		
		tomar *VMM02S0*		

¿Es difícil el tagging?

- El gran problema es la desambiguación.
- Ej.; Rushdie ya respira por sus propios medios y ha iniciado “la senda de la recuperación”, dice su agente.
- Muchas palabras tienen varias entradas de diccionario.
- Pero suele no haber ambiguedad: dado el contexto, una sola opción es válida

¿Es difícil el tagging?

Rushdie ya respira por sus propios medios y ha iniciado “la senda de la recuperación”, dice su agente.

Rushdie	ya	respira	por	sus	propios	medios	y
rushdie *NP00SP0* 1	ya *RG* 0.999785	respirar *VMIP3S0* 0.98924	por *SP* 1	su *DP3CPN* 0.999903	propio *AQ0MP00* 0.956349	medio *NCMP000* 0.997344	y *CC* 0.999989
	ya *CS* 0.000214961	respirar *VMM02S0* 0.0107595		sus *I* 9.65065e-05	propio *NCMP000* 0.0436508	medio *AQ0MP00* 0.00132802	y *NCFS000* 1.07213e-05
						medir+os *VMM02P0+PP2CP00* 0.00132802	

¿Es difícil el tagging?

Rushdie ya respira por sus propios medios y ha iniciado “la senda de la recuperación”, dice su agente.

ha	iniciado	“	la	senda	de	la	recuperación
haber *VAIP3S0* 0.999889	iniciar *VMP00SM* 0.992958	“ *Fra* 1	el *DA0FS0* 0.98926	senda *NCFS000* 1	de *SP* 0.999961	el *DA0FS0* 0.98926	recuperación *NCFS000* 1
ha *I* 5.55463e-05	iniciado *NCMS000* 0.00704225		lo *PP3FSA0* 0.010734		de *NCFS000* 3.85246e-05	lo *PP3FSA0* 0.010734	
haber *VMIP3S0* 5.55463e-05			la *NCMS000* 6.20105e-06			la *NCMS000* 6.20105e-06	

¿Cuántas categorías posibles?

Las palabras más comunes suelen tener varias categorías (leyes de Zipf)

Types:		WSJ	Brown
Unambiguous	(1 tag)	44,432 (**86%**)	45,799 (**85%**)
Ambiguous	(2+ tags)	7,025 (**14%**)	8,050 (**15%**)
Tokens:			
Unambiguous	(1 tag)	577,421 (**45%**)	384,349 (**33%**)
Ambiguous	(2+ tags)	711,780 (**55%**)	786,646 (**67%**)

Figure 8.4 Tag ambiguity in the Brown and WSJ corpora (Treebank-3 45-tag tagset).

La mayoría de las palabras no es ambigua, pero las más frecuentes si lo son.

¿Qué conjunto de categorías se usa en tagging?

- Diversos conjuntos según implementaciones.
- Primer corpus anotado Penn Treebank, Universidad de Pensilvania, 1992,
- Unión Europea, proyecto EAGLES, 1993, tags con estructura interna, Freeling
- spacy actualmente, una versión muy simplificada

REFERENCIAS

PennTagset

Freeling Tagset

Métodos en tagging

- Todos los métodos usan aprendizaje supervisado.
- Han partido históricamente de corpus anotados a mano.
- En el caso del español el corpus es Ancora, desarrollado en la UPC, Universidad Politécnica de Cataluña.
  http://clic.ub.edu/corpus/es

Métodos en tagging

- Modelos de Markov Ocultos, (*HMM, Hidden Markov Models*), modelo probabilístico generativo.
- CRF, *Conditional Random Fields*, modelo probabilístico discriminativo.
- Redes Neuronales
- *Transformers*

Resultados de métodos en *tagging*

- La *accuracy* (el porcentaje de tags correctos, según tags anotados manualmente) de los métodos de *tagging* es muy alta, del orden del 97 %.
- Eso iguala además el acuerdo entre anotadores (para el inglés).
- Es independiente del algoritmo utilizado.

Reconocimiento de entidades con nombre *NER, Named Entity Recognition*

- Una subcategoría de nombre presente en los taggers es la de Nombre Propio
- Suele ir acompañada de información sobre el tipo de nombre propio.
- Las clases usuales son *Persona*, *Lugar*, *Organización*
- Tienen complejidad adicional ya que suelen ser multi-palabra

Ejemplo con Freeling

La Policía llegó cerca de las dos de la madrugada a Rambla Wilson y Sarmiento.

La	Policía	llegó	cerca	de	las	dos_de_la_madrugada	a	Rambla_Wilson	y	Sarmiento	.
el	policía	llegar	cerca	de	el	[??:??/??/??:2.00:am]	a	rambla_wilson	y	sarmiento	.
DA0FS0	*NP00O00*	*VMIS3S0*	*RG*	*SP*	*DA0FP0*	*W*	*SP*	*NP00G00*	*CC*	*NP00G00*	*Fp*

- El 5to carácter codifica el tipo de nombre propio.
- En el ejemplo vemos O por Organización y G por lugar geográfico.
- Se ve también una codificación de fecha hora por la expresión temporal ´dos de la madrugada´

Ejemplo con spacy

In [9]:
```
doc = nlp("La Policía llegó cerca de las dos de la madrugada a Rambla Wilson y Sarmiento.")
for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

```
Policía 3 10 ORG
Rambla Wilson 52 65 LOC
Sarmiento. 68 78 LOC
```

- el proceso ´nlp´ genera todos los análisis de un documento (en este caso, una oración).
- Aparecen las mismas dos clases que en el análisis de Freeling.
- No hay codificación de fecha hora por la expresión temporal ´dos de la madrugada´