Introducción a Natural Language Processing

Lingüística Básica

Ronald Cárdenas Acosta

Setiembre, 2016

Outline

Outline

Notación básica (1)

- Corpus:
  Dataset de texto anotado con información lingüistica específica para una tarea de NLP (parsing, NER, WSD)
- Token:
  Palabra tal y como se encuentra en el corpus o documento.
- Type:
  Palabra única presente en el corpus.
- Vocabulario:
  Cantidad de types/tipos de palabra.

Notación básica (2)

- N-grama:
  Secuencia de N tokens seguidos. Ejemplo: unigramas (w), bigramas (w_{i-1}, w_i), trigramas (w_{i-2}, w_{i-1}, w_i)
- Lexicon:
  Diccionario de categorías gramaticales compatibles con cada palabra o type del vocabulario.
- Raíz (*stem*):
  Parte de la palabra que no cambia ante los cambios morfológicos (conjugación).
- Lexema (*Lemma*):
  Forma infinitiva o "de diccionario" de un conjunto de palabras/typos.

Morfología

- Las diversas conjugaciones de una palabra son derivadas a partir de procesos morfológicos.
- Inflexión:
  Se agrega sufijos y/o prefijos a la raíz. No cambia la categoría gramatical.
- Derivación:
  Agrega sufijos que cambian la categoría gramatical.
- Composición:
  Unión de dos o más palabras. Pueden estar separadas por un guión.

Outline

Categorías gramaticales (POS)

- Se distinguen 9 categorías o clases de palabras para Español: Sustantivo, Adjetivo, Pronombre, Preposición, Conjunción, Interjección, Adverbio, Articulo, Verbo.
- Se pueden dividir en dos grupos:
  - Categorías abiertas: admiten palabras nuevas (sustantivo, adjetivo, verbo, interjección)
  - Categorías cerradas: no admiten palabras nuevas (pronombres, preposiciones, etc)
- Los corpus codifican cada categoría con información morfológica (conjugación) adicional en etiquetas (TAGSET).
- Cada lenguaje dispone de un Tagset oficial.
- Existe también un Tagset universal (nuevo).

EAGLE TAGSET: POS para Español

Figure: Ejemplo de EAGLE tagset para adjetivos, para otras categorías ver `http://nlp.lsi.upc.edu/freeling-old/doc/tagsets/tagset-es.html`.

ADJETIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
		Ordinal	O
3	Grado	Aumentativo	A
		Diminutivo	D
		Comparativo	C
		Superlativo	S
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Función	-	0
		Participi	P

Penn Treebank POS TAGSET: POS para Inglés

Figure: Ver lista completa en `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective

Universal POS tags

Figure: Ver descripción completa en `http://universaldependencies.org/u/pos/index.html`.

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	