Pos-Tagging

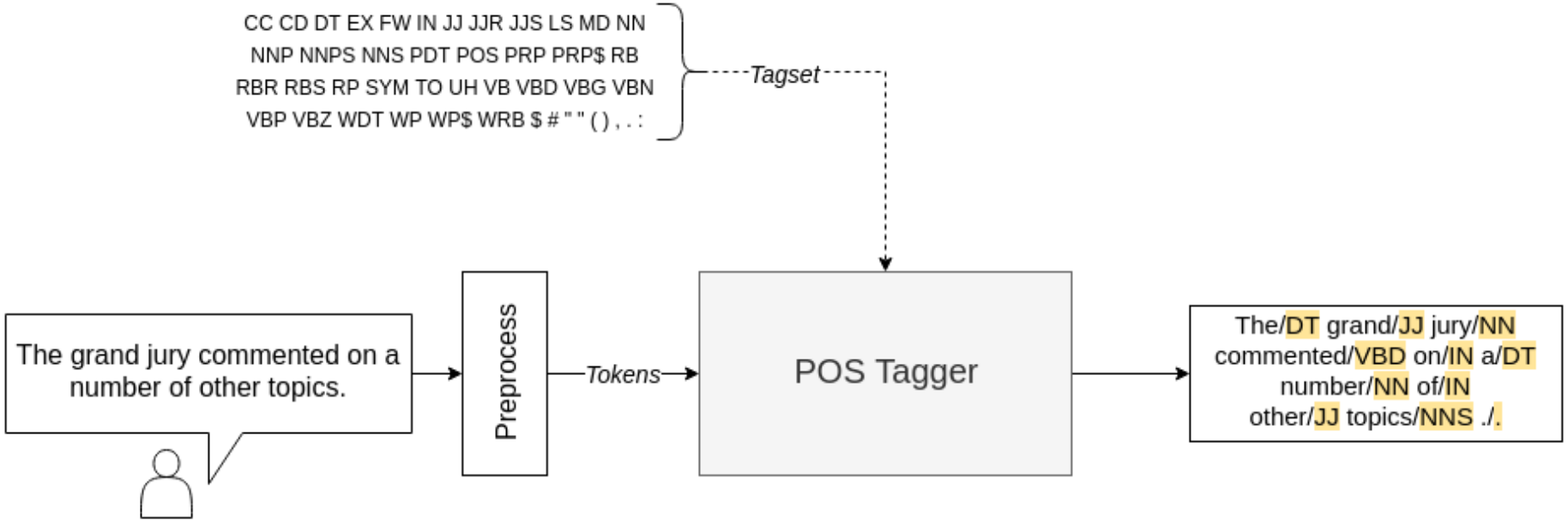
Andrés Rosso

¿Qué es el etiquetado de partes de la oración (POS-Tagging)?

La asociación de cada palabra de una frase con una POS (parte de la oración) adecuada se conoce como etiquetado POS o anotación POS.

Anteriormente, la anotación POS era realizada manualmente por anotadores humanos, pero al ser una tarea tan laboriosa, hoy en día disponemos de herramientas automáticas que son capaces de etiquetar cada palabra con una etiqueta POS adecuada dentro de un contexto.

Arquitectura Básica



Categorías POS

Clase cerrada.

- **Palabras de función**: preposiciones, pronombres determinantes, conjunciones, verbos auxiliares y partículas .

Clase abierta:

- **Sustantivos**: personas, lugares y cosas sustantivos propios, comunes sustantivos comunes, sustantivos contables y sustantivos incontable
- **Verbos**: acciones y procesos. Verbos principales, no auxiliares
- **Adjetivos**: Propiedades
- **Adverbios**: Modifica el verbo, lugar, tiempo, modo, cantidad.

Mejor explicación en https://universaldependencies.org/u/pos/all.html

Beneficios POS-Tagging

Es útil para muchas de las tareas de NLP como son:

- Recuperación de información
- Extracción de información
- Sistemas de conversión de texto en voz
- Reconocimiento de entidades nombradas
- Question-Answering
- **Desambiguación del sentido de las palabras**

Tipos de Tag

Desde los griegos se han distinguido 8 TPV básico:

- Sustantivo
- Verbo
- Pronombre
- Preposición
- Adverbio
- Conjunción
- Adjetivo
- Artículo

Ahora se usan más de 87 (Brown)

TASG en Penn Tree Bank

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular
NNP	Proper noun, singular
NNS	Noun, plural
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Posessive ending
PRP	Personal pronoun
PP	Possessive pronoun

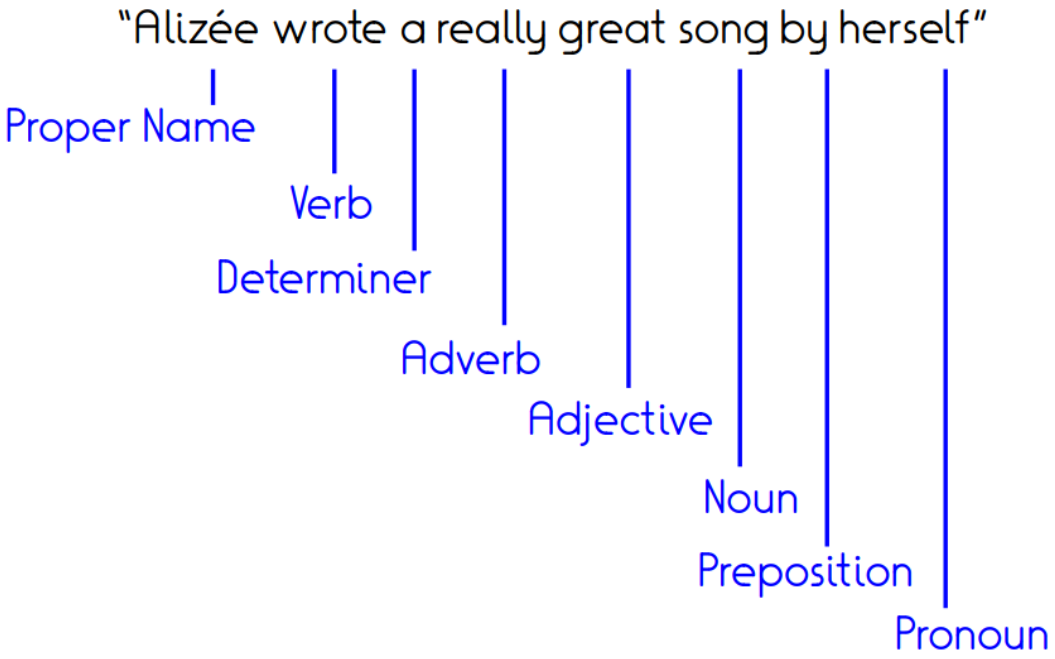
Penn Tree Bank tagset

RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund
VBN	Verb, past participle
VBP	Verb, non-3rd ps. sing. present
VBZ	Verb, 3rd ps. sing. present
WDT	wh-determiner
WP	wh-pronoun
WP	Possessive wh-pronoun
WRB	wh-adverb

TAGS Comunes

POS Category	Identifier	Corresponding POS Tags
Adjectives	ADJ	AQ, AC
Conjunctions	CON	CC, CS
Determiners	DET	DA, DD, DE, DI, DN, DP, DT
Punctuation	PUN	Fa, Fc, Fd, Fe, Fg, Fh, Fi, Fp, Fs, Fx, Fz
Nouns	NOU	NC, NP
Pronouns	PRO	PO, PD, PI, PN, PP, PR, PT, PX
Adverbs	ADV	RG, RN
Prepositions	PRE	SP
Verbs	VER	VA, VM, VS

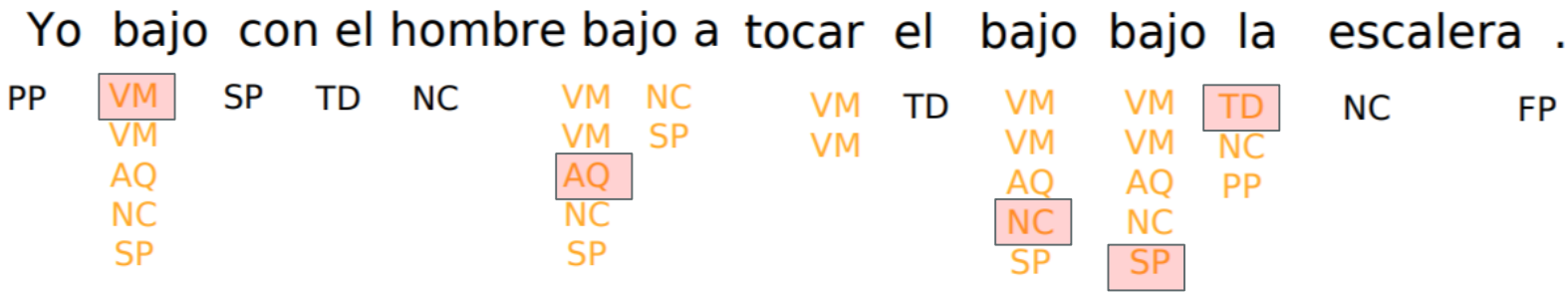
Ejemplo


"Alizée wrote a really great song by herself"
Proper Name
Verb
Determiner
Adverb
Adjective
Noun
Preposition
Pronoun

Ejemplo en Español (ambiguo)

Yo	bajo	con	el	hombre	bajo	a	tocar	el	bajo	bajo	la	escalera	.
PP	VM VM AQ NC SP	SP	TD	NC	VM VM AQ NC SP	NC SP	VM VM	TD	VM VM AQ NC SP	VM VM AQ NC SP	TD NC PP	NC	FP

Ejemplo en Español (des-ambiguado)



Formas de Hacer el Pos-Tagging

- Etiquetado POS basado en reglas
- Etiquetado basado en la transformaciones
- Modelos de aprendizaje profundo
- Etiquetado estocástico (probabilístico)

$W = w_1\ w_2\ \ldots\ w_n$ sequence of words
$T = t_1\ t_2\ \ldots t_n$ sequence of POS tags

$$f:\ W \rightarrow T = f(W)$$

Pos-Tagging Basado en Reglas (1)

Los etiquetadores basados en reglas utilizan un diccionario o léxico para obtener las posibles etiquetas de palabra.

- Reglas de patrones contextuales
- Expresiones regulares compiladas

Generalmente estos modelos constan de 2 etapas:

- Primera etapa - se utiliza un diccionario para asignar a cada palabra una lista de posibles partes del discurso.
- Segunda etapa - En la segunda etapa, utiliza grandes listas de reglas de desambiguación escritas a mano para clasificar la lista en una sola parte del discurso para cada palabra.

Pos-Tagging Basado en Reglas (2)

- El prefijo "in-" sugiere un **adjetivo**, como "**in**sondable".
- El sufijo "-ar" sugiere un adverbio, como "camin**ar**".
- Las mayúsculas pueden sugerir un nombre propio, como "Colombia".
- Las formas de las palabras también son útiles, como "35 años", que es un adjetivo.

Ejemplo Brill

- La palabra anterior (siguiente) está etiquetada con z.
- La segunda palabra anterior (después) está etiquetada con z.
- Una de las dos palabras anteriores (siguientes) está etiquetada con z.
- Una de las tres palabras anteriores (siguientes) está se marca con z.
- La palabra anterior está etiquetada como z y la siguiente palabra está etiquetada como w.
- La palabra anterior (siguiente) está etiquetada como z y segunda palabra anterior (después) está etiquetada como w

Pos-Tagging Estocástico

- Conocer las frecuencias de los TAGS de las palabras en un corpus etiquetado permite generalizar y encontrar un modelo probabilístico.
- La secuencia de etiquetas más probable dada la observación está dada por:

$$\hat{P}(t_1^n|w_1^n)$$

$$P(x|y) = P(x)\,P(y|x) / P(y)$$

- Un modelo de Markov oculto (HMM) modelar la secuencia de palabras/TAGS y estimar la etiqueta más probable .

Primera Suposición

La probabilidad de una etiqueta depende de la anterior (modelo de bigramas) o de las dos anteriores (modelo de trigramas) o de las n etiquetas anteriores (modelo de n-gramas) .

- PROB $(C_1,..., C_T) = \Pi_{i=1..T}$ PROB $(Ci|C_{i-n+1}...C_{i-1})$ (modelo de n-gramas)
- PROB $(C_1,..., C_T) = \Pi_{i=1..T}$ PROB $(C_i|C_{i-1})$ (modelo de bigramas)

El comienzo de una frase puede calcularse asumiendo una probabilidad inicial para cada etiqueta.

PROB $(C_1|C_0)$ = PROB inicial (C_1)

Segunda Suposición

Se puede suponer además que una palabra aparece en una categoría independiente de las palabras de las categorías anteriores o posteriores:

$$\text{PROB}(W_1,\ldots, W_T \mid C_1,\ldots, C_T) = \Pi_{i=1..T}\ \text{PROB}(W_i|C_i)$$

Ahora, sobre la base de los dos supuestos anteriores, nuestro objetivo se reduce a encontrar una secuencia C que maximice

$$\Pi_{i=1\ldots T}\ \text{PROB}(C_i|C_{i-1}) * \text{PROB}(W_i|C_i)$$