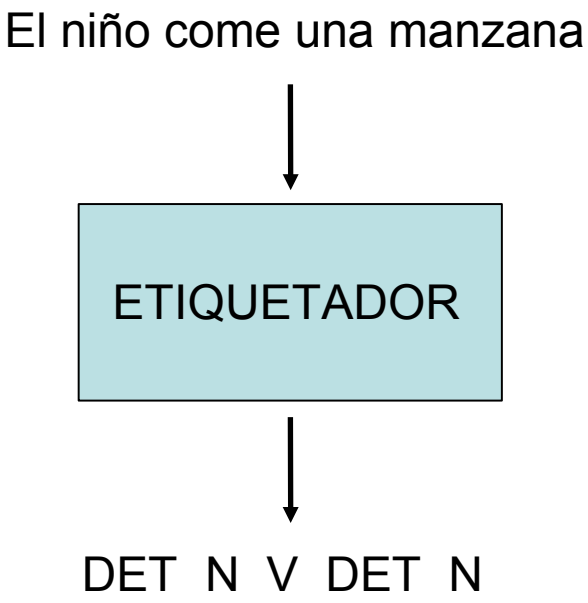
Práctica 1

Definición

- Un etiquetador/anotador recibe una secuencia de palabras y le asigna una secuencia de etiquetas gramaticales (tags)



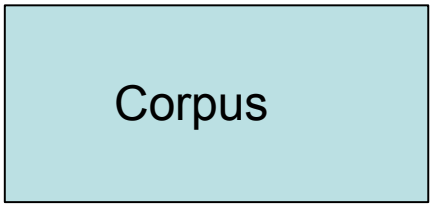
Dificultad de la anotación: Ambigüedad

- Una misma palabra puede tener etiquetas diferentes en diferentes frases.
  - Ha pesado/V tres kilos
  - El profesor es un pesado/N
  - Un discuro pesado/ADJ
- La categoria gramatical de una palabra en una frase depende del contexto de utilización.
  - Ejemplo.
    - La palabra “médico” después del determinante (“un”) generalmente actuará como nombre (y no como adjetivo). “Un medico/N” vs. “un equipo médico/ADJ”

¿Cómo automatizar el proceso?

- Opción 1:
  - Obtener un diccionario
  - Escribir reglas
    - si w_i="médico" y t_{i-1} es un DET entonces
    - entonces t_i = N
- Opción 2: Utilizando Aprendizaje Automático (Machine Learning)

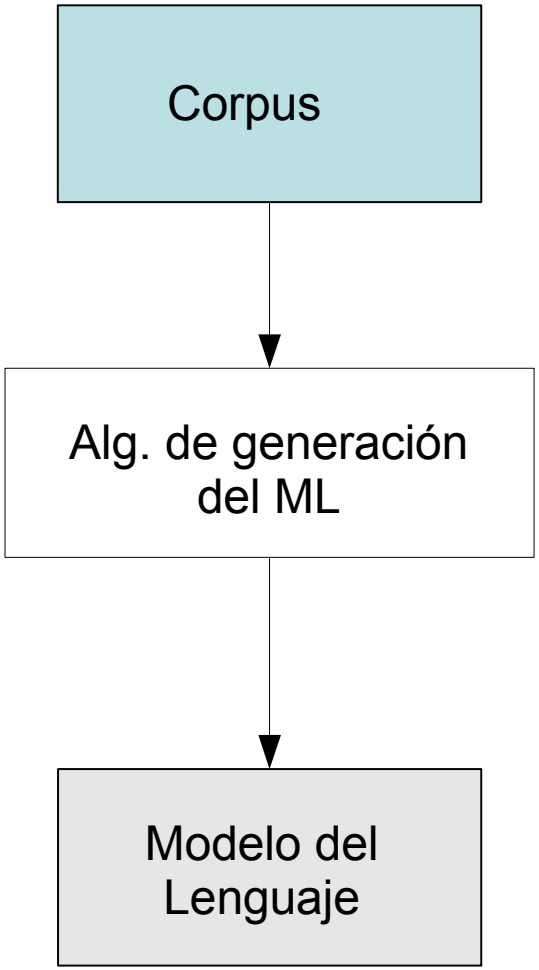
Tagging como *Supervised* Machine Learning



- Datos anotados manualmente: “Training test”. En LC se suele llamar a estos datos un “Corpus anotado”.

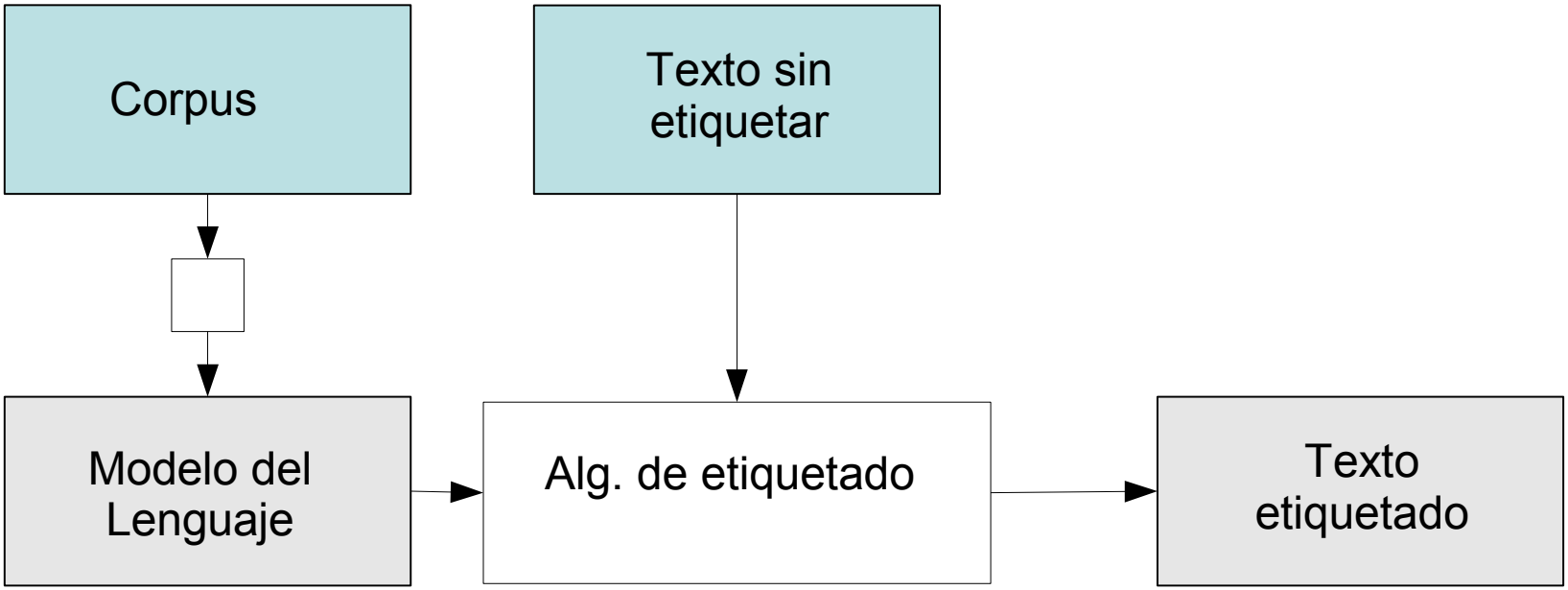
- Objetivo: extraer de estos datos “**conocimento**” para poder repetir la operación de etiquetado con otros textos

Tagging como *Supervised* Machine Learning
1 - Entrenamiento



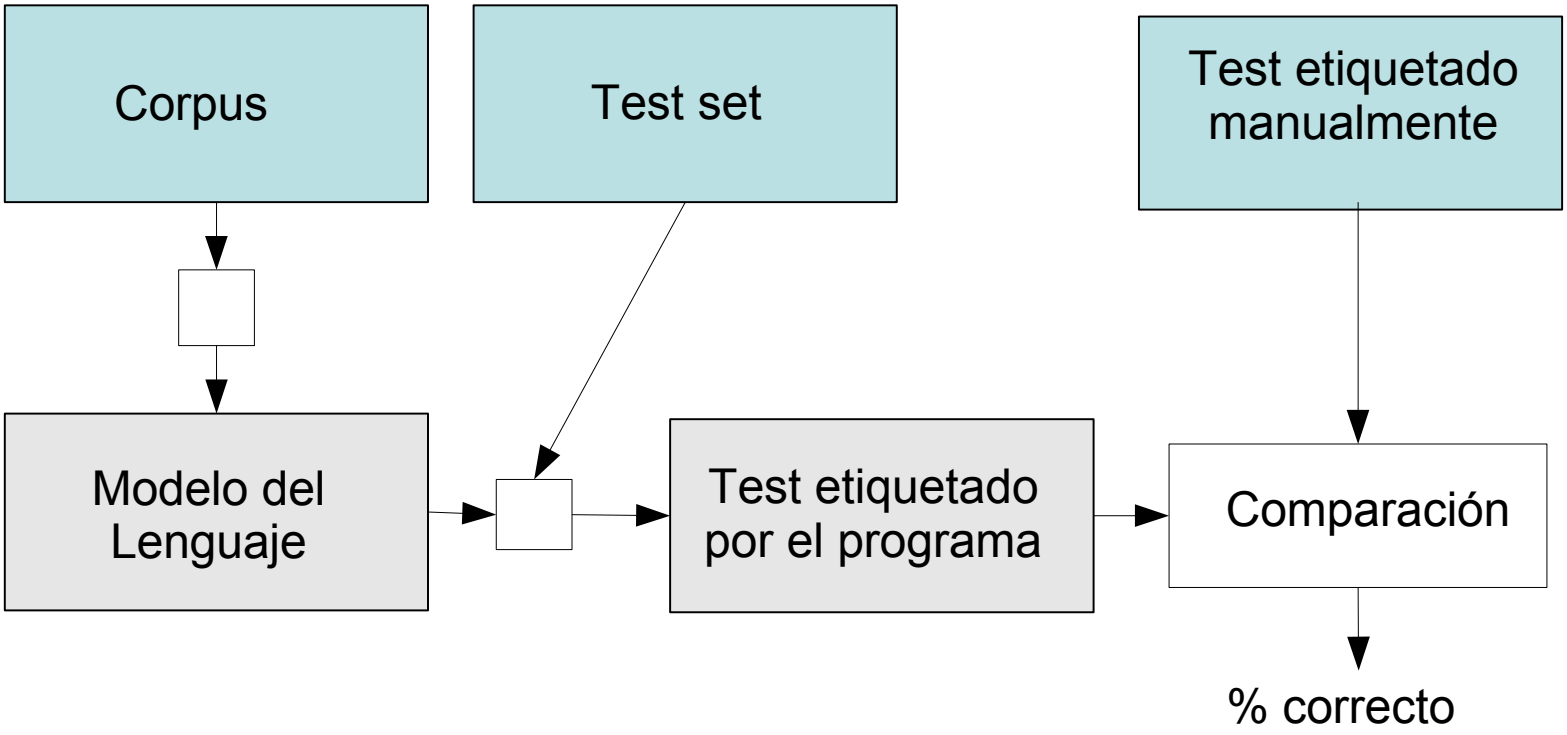
- El Modelo del Lenguage (ML) puede consistir en reglas, conjunto de probabilidades, pesos de un red neuronal, etc.

Tagging como *Supervised* Machine Learning
2 – Etiquetado



- Utilizando el ML se puede ahora etiquetar cualquier texto nuevo

Tagging como *Supervised* Machine Learning
3 – Evaluación



- Para conocer la precisión del algoritmo de etiquetado se compara con un texto anotado manualmente que no se ha usado para el entrenamiento

Anotación Estadística

- Dada una secuencia de palabras $\mathbf{W=(w_1, \ldots, w_n)}$ y una secuencia de tags $\mathbf{T=(t_1, \ldots, t_n)}$ definimos la probabilidad de que a **W** le corresponda **T** como:

$$P(T|W)$$

- Consideramos que la mejor secuencia será la maximize la eq. anterior

$$\hat{T} = argmax_T P(T|W)$$

Anotación Estadística

- La realidad (aplicando la regla de la cadena)

$$P(T|W) = \prod_{i=1}^{N} P(t_i|w_N, \ldots, w_{1,} t_{i-1}, \ldots, t_1)$$

- Lo que haremos en la práctica 1

$$P(T|W) \approx \prod_{i=1}^{N} P(t_i|w_i)$$

¿Cómo podemos utilizar el corpus?

- Podemos extraer $P(t_i|w_i)$ aplicando la regla general

$$P(A|B)=\frac{P(A \wedge B)}{P(B)}$$

P(A & B) o P(B) las podemos estimar a partir de las ocurrencias en el corpus.

  - Ejemplo:

$$P(t_i|w_i)=\frac{count(t_i, w_i)}{count(w_i)}$$

    - numerador: veces que la palabra w_i aparece etiquetada como t_i
    - denominador: veces que aparece la palabra w_i