CDLN-2022

EJEMPLOS DE PREGUNTAS DE EVALUACIÓN

Solución

CATEGORÍAS GRAMATICALES, TAGGING

1. Indique la categoría gramatical de cada palabra en el siguiente fragmento de texto. Utilice la clasificación en 8 categorías vista en el curso.

> *Gradualmente, el hermoso universo fue abandonándolo; una terca neblina le borró las líneas de la mano, la noche se despobló de estrellas, la tierra era insegura bajo sus pies. Todo se alejaba y se confundía.*

2. ¿Cuál es la diferencia entre etiquetado morfo-sintáctico y tagging? ¿Qué consecuencias para el tagging tiene una de las versiones de la ley de Zipf?

1. Las 8 categorías son, abreviadas: N (nombre), V (verbo), Adj (adjetivo), Adv (Adverbio), Prep (preposición), Det (Determinante), Pron (pronombre), C (conjunción)

> *Gradualmente, el hermoso universo fue abandonándo-lo; una terca neblina le*
> *Adv Det Adj N V V Pron Det Adj N Pron*
>
> *borró las líneas de la mano, la noche se despobló de estrellas, la tierra era*
> *V Det N Prep Det N Det N Pron V Prep N Det N V*
>
> *insegura bajo sus pies.*
> *Adj Prep Det N*
>
> *Todo se alejaba y se confundía.*
> *Pron Pron V Conj Pron V*

2. ¿Cuál es la diferencia entre etiquetado morfo-sintáctico y tagging?

El etiquetado morfo-sintáctico puede asignar varias categorías morfo-sintácticas a una palabra, en el tagging se desmbigua considerando el contexto y finalmente hay una única categoría morfo-sintáctica por cada palabra.

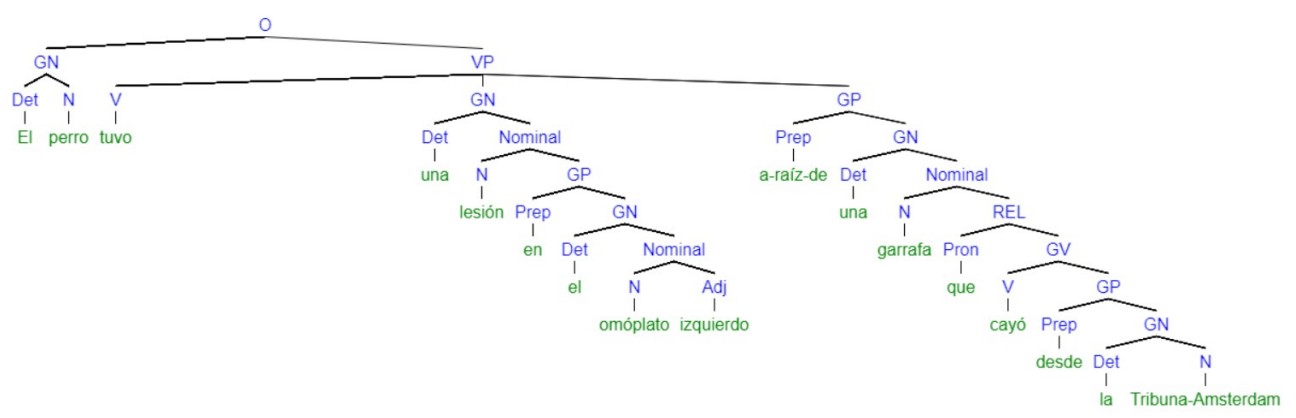
ESTRUCTURAS SINTÁCTICAS, ÁRBOLES DE CONSTITUYENTES Y DE DEPENDENCIAS

a) Realice el análisis en constituyentes y en dependencias de la siguiente oración :

> *El perro tuvo una lesión en el omóplato izquierdo a raíz de una garrafa que cayó desde la Tribuna Ámsterdam.*

Veremos una versión parentizada del árbol de constituyentes y un dibujo del árbol:

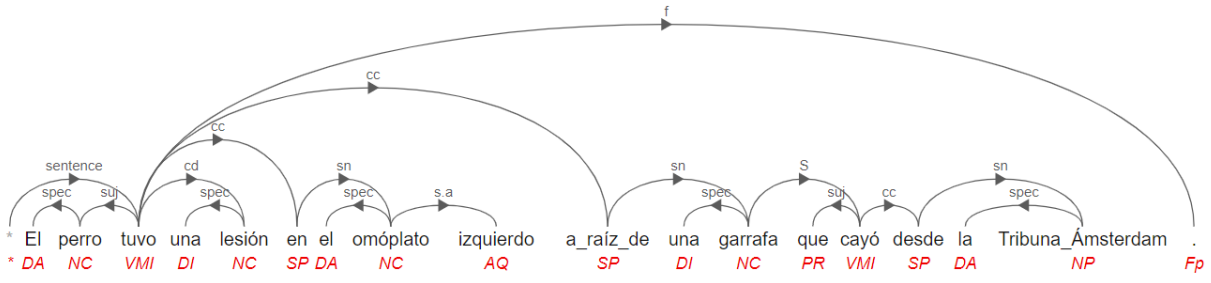
> [O [GN [Det El] [N perro]][VP [V tuvo] [GN [Det una] [Nominal [N lesión] [GP[Prep en] [GN [Det el] [Nominal [N omóplato] [Adj izquierdo]]]]]] [GP [Prep a-raíz-de] [GN [Det una] [Nominal [N garrafa] [REL [Pron que] [GV [V cayó] [GP [Prep desde] [GN [Det la] [N Tribuna-Amsterdam]]]]]



Algunos comentarios respecto a las categorías y los constituyentes que se proponen:

- Las categorías no pre-léxicas usadas son O (oración), GN (grupo nominal), GV (grupo verbal), GP (grupo preposicional), Nominal (nombre con calificadores y REL (subordinada relativa). En las herramientas de parsing se suele ver las abreviaturas en inglés (S (sentence), VP (Verbal Phrase), etc.).
- Hay algunas unidades multipalabra que se juntaron con guiones (a-raíz-de, Tribuna-Amsterdam). Esto no es obligatorio, se podría haber armado un pequeño constituyente con reglas.
- La frase relativa (REL) es un complemento de nombre, por eso entra en la categoría Nominal en el mismo lugar en el que puede haber un adjetivo.

Veremos el árbol de dependencias que generó Freeling:



- Las abreviaciones en rojo debajo de cada palabra no forman parte del árbol, corresponden a los tags morfosintácticos que Freeling asignó a cada palabra o unidad léxica.
- Si bien se trata de un formalismo distinto al del primer árbol, se puede ver que las estructuras son básicamente equivalentes. Hay sin embargo una diferencia importante : ¿reconoce cuál es?

b) ¿El árbol de dependencias de la parte a) es proyectivo ? Fundamente.

Es proyectivo, ya que no hay cruzamiento de aristas.

c) ¿Qué herramientas para análisis sintáctico conoce ? Descríbalas brevemente.

Breve descripción de Freeling y Spacy, cómo se usan, qué funcionalidades tienen. Se extraen de lo sitios web de las herramientas

EXTRACCIÓN DE INFORMACIÓN : ENTIDADES CON NOMBRE

1. Considere el siguiente ejemplo

> *El perro tuvo una lesión en el omóplato izquierdo a raíz de una garrafa que cayó desde la Tribuna Ámsterdam y que también hirió al oficial que lo llevaba. "La lesión (de Duque) fue un traumatismo a nivel de la cruz, eso ocasionó un hematoma y ahora está siendo tratado con antinflamatorios y analgésicos", dijo a El Observador el capitán John Severo, jefe del K-9, la unidad canina de la Guardia Republicana. Duque es uno de los perros que trabaja controlando disturbios y su función es más "psicológica que represiva", dijo Severo.*

a) Indique, en el texto anterior, cuáles son las entidades con nombre y cuál es su tipo semántico. Utilice las clases PERSONA, ORGANIZACIÓN, LUGAR y agregue clases si lo considera necesario.

XXXX Persona
XXXX Organización
XXXX Otra
XXXX Lugar

> *El perro tuvo una lesión en el omóplato izquierdo a raíz de una garrafa que cayó desde la Tribuna Ámsterdam y que también hirió al oficial que lo llevaba. "La lesión (de Duque) fue un traumatismo a nivel de la cruz, eso ocasionó un hematoma y ahora está siendo tratado con antinflamatorios y analgésicos", dijo a El Observador el capitán John Severo, jefe del K-9, la unidad canina de la Guardia Republicana. Duque es uno de los perros que trabaja controlando disturbios y su función es más "psicológica que represiva", dijo Severo.*

Podría ser necesario agregar más clases, es dudoso que “Tribuna Amsterdam” sea un lugar. Otra clase podría ser nombres propios de animales (ej., Duque). Dependerá del contexto y la aplicación.

b) ¿En qué consiste la “wikificación”? ¿Por qué puede ser necesaria? Ilustre con el texto anterior, discuta si le parece pertinente en ese caso. ¿Que´procesos serían de interés para ese texto?

Un nombre propio no es necesariamente una expresión que denote un objeto único en el contexto. Hay nombres que pueden aplicarse a personas, a un país, a un río (Uruguay). O aplicarse a varias personas distintas (ej., Cristina Fernández). La “wikificación” refiere a desambiguar un nombre propio utilizando como base páginas de desambiguación de Wikipedia.

En el texto anterior un caso de ambigüedad es “Amsterdam”, ciudad o nombre de una tribuna en un estadio. Otro caso es el nombre del diario “El Observador”, existió un diario con el mismo nombre en Perú. Sabemos que “El País” es un diario de Madrid además de ser un diario uruguayo. De

hecho, en https://es.wikipedia.org/wiki/El_Pa%C3%ADs_(desambiguaci%C3%B3n) , se puede ver un conjunto de periódicos con nombre “El País”, listado en una página de desambiguación de Wikipedia. Obviamente, Wikipedia no proporciona información para todos los casos de nombres propios que se reusan para distintas entidades.

Respecto a los procesos, en un caso de ambigüedad puede ser necesario vincular la mención en el texto con el referente adecuado, por ejemplo para responder adecuadamente en una aplicación de respuestas a preguntas.

c) ¿Qué representan los esquemas BIO y BILOU? Etiquete según el esquema BIO la última oración del texto ejemplo.

Los esquemas BIO (Begin In Out) y BILOU (Begin In Last Out Unique) formas de realizar etiquetado secuencial de un texto, que dada una clase que deseamos reconocer (p.ej., nombre propio de persona) asigna una etiqueta a cada palabra, indicando si inicia un segmento de la clase en cuestión (B), si es interior a un segmento de la clase en cuestión (I), si es la última palabra de un segmento de la clase en cuestión (L), si es la única palabra de un segmento de la clase en cuestión (U) o si no pertenece a ningún segmento de la clase en cuestión (O).

Para la oración se etiqueta simplemente nombre propio, sin indicar tipo de nombre propio. Notar que si trabajamos con las subclases anteriores debemos multiplicar por 2 (BIO) o por 4 (BILOU) la cantidad de etiquetas distintas, y sumar 1 por la etiqueta O.

> *Duque es uno de los perros que trabaja controlando disturbios y su función es más*
> *B O O O O O O O O O O O O O*
>
> *"psicológica que represiva", dijo Severo.*
> *O O O O B*

d) Indique qué atributos utilizaría para la extracción de entidades con nombre siguiendo un método de aprendizaje supervisado.

Los nombres propios tienen algunas características diferenciadas en el texto:
- empiezan con mayúscula
- suelen ser expresiones multipalabra
- en el caso de nombre de persona pueden ir precedidas por palabras como Ing. Sr. Sra. ...
- existen listas de nombres de pila
- existen listas de nombres geográficos (gazeeter)

Todas estas características se pueden expresar como atributos. Es posible que los atributos pertinentes varíen según la clase del nombre propio.

2. Considere el siguiente ejemplo

> *Leonardo Cipriani, presidente de ASSE, redefinió en las últimas horas la situación del Hospital de Bella Unión, en Artigas. En diálogo con el periodista Leonardo Sarro, de Radio Monte Carlo, explicó que no se trataba en sentido estricto de un brote -término que él mismo empelara ayer- debido a las características de los casos presentados.*
>
> *El jerarca informó que cerca de la pasada medianoche se diagnosticaron seis nuevos casos*

que "no están relacionados al hospital sino a la familia" de uno de los diagnosticados anteriormente, y que "parte de esa familia vive en Bella Unión, y otra parte en Barra de Quaraí", ya en territorio brasileño.

a) Indique, en el texto anterior, cuáles son las entidades con nombre y cuál es su tipo semántico. Utilice las clases PERSONA, ORGANIZACIÓN, LUGAR y agregue clases si lo considera necesario.

Utilizamos las clases y código de colores de la parte b)

Leonardo Cipriani, presidente de ASSE, redefinió en las últimas horas la situación del Hospital de Bella Unión, en Artigas. En diálogo con el periodista Leonardo Sarro, de Radio Monte Carlo, explicó que no se trataba en sentido estricto de un brote -término que él mismo empelara ayer- debido a las características de los casos presentados.

El jerarca informó que cerca de la pasada medianoche se diagnosticaron seis nuevos casos que "no están relacionados al hospital sino a la familia" de uno de los diagnosticados anteriormente, y que "parte de esa familia vive en Bella Unión, y otra parte en Barra de Quaraí", ya en territorio brasileño.

b) La salida de un reconocedor de entidades con nombre para el texto anterior genera la siguiente salida. Las entidades con nombre reconocidas están marcadas con fondo de color :

XXXX Persona
XXXX Organización
XXXX Otra
XXXX Lugar

Leonardo Cipriani, presidente de ASSE, redefinió en las últimas horas la situación del Hospital de Bella Unión, en Artigas. En diálogo con el periodista Leonardo Sarro, de Radio Monte Carlo, explicó que no se trataba en sentido estricto de un brote -término que él mismo empelara ayer- debido a las características de los casos presentados.

El jerarca informó que cerca de la pasada medianoche se diagnosticaron seis nuevos casos que "no están relacionados al hospital sino a la familia" de uno de los diagnosticados anteriormente, y que "parte de esa familia vive en Bella Unión, y otra parte en Barra de Quaraí", ya en territorio brasileño.

Calcule Precision y Recall globales y por Clase.

Para calcular Precision y Recall consideraremos que una instancia es correcta si coincide exactamente con la anotación humana (parte a del ejercicio) e incorrecta en caso contrario. Así, Bella Unión (en Hospital de Bella Unión) es una instancia incorrecta, aunque haya un match parcial. Este error se va a reflejar tanto en Precision (se marca una entidad incorrecta) como en Recall (no se marca una entidad correcta).

Precision y Recall se aplican a problemas de clasificación binarios. Cuando se habla de Precision y Recall globales, nos referimos a las clase más amplia de Entidad con Nombre, sin considerar si es

Persona, Organización etc.

Llamamos Verdaderas Positivas (VP) a las entidades correctas detectadas.
Llamamos Falsas Positivas (FP) a las entidades erróneamente detectadas.
Llamamos Falsas Negativas (FN) a las entidades correctas que no fueron detectadas.

Las definiciones de Precision y Recall son:

Precision = VP / (VP + FP)
Recall = VP / (VP + FN)

(*Notar que para Precision y Recall no nos interesan las Verdaderas Negativas, que sí influyen en Accuracy*)

Valores Globales
Precision = 5 / (5+4) = 5/9
Recall = 5/ (5+3) = 5/8

Valores por Clase
Precision-Persona = 2/(2+0) = 1
Recall-Persona =2/(2+0) = 1

etc.

c) ¿En qué consiste la "wikificación"? ¿Por qué puede ser necesaria? Ilustre con el texto anterior, discuta si le parece pertinente en ese caso. ¿Que´procesos serían de interés para ese texto?

La parte general de esta pregunta ya fue contestada en la parte 1. Respecto al ejemplo concreto, observamos que Monte Carlo es el nombre de una radio y el nombre de un barrio de un país-ciudad y esto podría provocar problemas.

EXTRACCIÓN DE INFORMACIÓN : RELACIONES

1. Anotación en textos

> *La célula es la unidad funcional de los tejidos. Las células forman los tejidos, los tejidos con características parecidas forman los órganos y los órganos con funciones similares forman los sistemas o aparatos.*
> *Las células presentan un citoplasma muy compartimentado, con orgánulos separados o interconectados, limitados por membranas biológicas que son de la misma naturaleza esencial que la membrana plasmática. El núcleo es solamente el más notable y característico de los compartimentos en que se divide el protoplasma, es decir, la parte activa de la célula. En el protoplasma distinguimos tres componentes principales, a saber, la membrana plasmática, el núcleo y el citoplasma, constituido por todo lo demás. Las células están dotadas de un citoesqueleto complejo, muy estructurado y dinámico, formado por microtúbulos y diversos filamentos proteicos. Además puede haber pared celular, que es lo típico de plantas, hongos y protistas pluricelulares, o algún otro tipo de recubrimiento externo al protoplasma.*

a) Encuentre al menos 3 instancias de relación parte-todo en el texto anterior. Identifique los

argumentos y un patrón sintáctico para la relación.

Patrón A es-parte-de B

1-Las células forman los tejidos

2-En el protoplasma distinguimos tres componentes principales, a saber, la membrana plasmática, el núcleo y el citoplasma,

3-Las células están dotadas de un citoesqueleto complejo

Los patrones se suelen especificar con expresiones regulares, que luego son ejecutables sobre el texto. Hay variantes en los formalismos para especificar las expresiones regulares, dependiendo del contexto de programación en el que estemos.
El ejemplo más sencillo sería en el caso 1, un pattern simplificado es : A forman B –> A es parte de B

b) Explique en qué consiste la supervisión distante aplicada a un problema de extracción de relaciones. Desarrolle un ejemplo para el texto anterior y pares según uno de los patrones de la parte a).

Partiendo de la hipótesis de que la co-ocurrencia de un mismo par de entidades en dos oraciones del mismo o distintos documentos representan casos de la misma relación, se hipotetiza nuevas instancias de la relación a partir de algunas instancias iniciales. Las instancias hipotetizadas proveen nuevos patterns.

2. Semi-supervisión, extracción abierta.

En general un algoritmo de aprendizaje es **semisupervisado** si opera inicialmente con un conjunto reducido de entrenamiento y lo va ampliando incrementalmente.

En general en extracción de información hay una relación previamente definida (p.ej., parte de en el ejemplo anterior). En **extracción abierta** la propuesta es identificar entidades y luego encontrar la relación expresada en el texto. O sea, no hay relaciones predefinidas. Se utiliza para estructurar la información que se encuentra en texto.

LEY DE ZIPF

Conteste las siguientes preguntas fundamentando brevemente.

1 - ¿Qué relación involucra la ley de Zipf?

La ley de Zipf vincula la frecuencia f_p de una palabra en un texto con el rango r_p de dicha palabra, mediante la relación, en textos grandes.

$f_p \approx 1/ r_p^{\ a}$

siendo a una cte cercana a 1

2 - ¿Cuál es (aproximadamente) el porcentaje de ocurrencia de términos que ocurren una sola vez

en un texto?

Se les llama hapax legomena, el porcentaje de ocurrencia es aproximadamente el 50%, es nuevamente una ley empírica.

3- ¿Qué consecuencias tiene, para el manejo de semántica léxica en PLN, la propiedad del punto anterior?

Estas palabras son problemáticas para enfoques estadísticos que se apoyen en conteos de palabras. Por ejemplo, cuando se estima la probabilidad a partir de frecuencias en un corpus.

MEDIDAS DE TEORÍA DE LA INFORMACIÓN

1. Sea X la variable aleatoria que representa el experimento de tirar una moneda 2 veces, el conjunto de valores de X es {CC,NN,CN, NC}, donde C representa Cara y N número

a) Plantee la entropía de X si cara y número son equiprobables.

La entropía de la v.a. X tiene la siguiente formula, siendo p_i la probabilidad de cada uno de los valores de la variable.

$$-\sum_{i=1}^{k} p_i \log_2 (p_i)$$

Como la probabilidad de cada evento es ¼, la entropía en este caso es - 4. ¼ . (-2) = 2

b) Suponga que la probabilidad de Cara es ¾. ¿Cómo se compara la entropía de X con la de la parte anterior ?

Se demuestra que la entropía es máxima cuando todos los valores de la v.a. son equiprobables. La entropía en este caso va a ser menor que en el caso a, ya que las probabilidades de los 4 eventos mencionados no son iguales.

2. ¿En qué consiste y para qué se usa la medida PMI (*Pointwise Mutual Information*)?

$$PMI(x;y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Se usa PMI para discriminar si la coocurrencia de dos valores x y y ocurre por azar o si hay dependencia entre ambos valores. Un ejemplo en corpus es el de la búsqueda de colocaciones. Las colocaciones son combinaciones de palabras que muestran cierta idiosincrasia en su distribución lingüística. Esta idiosincrasia puede consistir en una menor composicionalidad semántica, una menor modificabilidad sintáctica o, simplemente, la sensación de que la combinación es habitual o incluso fija. Algunos ejemplos son “lluvia torrencial”, “fuertes vientos” y “estrepitoso fracaso”.

3. ¿En qué consiste la entropía cruzada (*Cross Entropy*)? Dé un ejemplo de su uso.

La entropía cruzada H(p,q) para una variable aleatoria ***x*** con distribución **p** (habitualmente la ideal desconocida) y una distribución **q** (experimental) surge al promediar según **p** la distribución **q**.

$$H(p, q) = -\sum_{i=1}^{n} p(x_i) \log_2 q(x_i)$$

Se suele usar como función de error en algoritmos de Aprendizaje Automático.

GRAFOS DE CONOCIMIENTO

1. ¿Cómo es el esquema de organización de Wikidata ?

La unidad básica de información es una tripleta que expresa una relación binaria, con 2 argumento (sujeto y objeto o atributo) y un predicado o propiedad. Estas relaciones se expresan mediante aristas de un grafo, los vértices son sujeto y objeto y la arista está rotulada con el predicado. Estas aristas se organizan en un grafo, ya que cada nodo puede participar en múltiples relaciones.

2. Mencione y dé ejemplos para 3 tipos de relaciones presentes en WordNet.

- hipónimo : perro en un hipónimo (en uno o más pasos) de animal
- merónimo : oveja es un merónimo de rebaño
- antónimo : murmurar es un antónimo de gritar

SEMÁNTICA VECTORIAL

1- ¿Cómo se representa la semántica léxica mediante vectores ?

Existen distintos modos o modelos para representar del contenido léxico de una palabra mediante vectores de números reales. Algunos se desarrollan en el punto 4.

2- ¿Qué es un modelo disperso? Indique un ejemplo.

En un modelo disperso el vector que representa a una palabra tiene muchos ceros. El modelo one hot (4 a) es un ejemplo.

3- ¿Qué es un modelo denso? , mencione al menos un tipo de modelo denso.

Un modelo denso no es disperso. Un vector en un modelo denso suele no tener ceros. Un ejemplo es el modelo de vectores asociado al método skip gram.

4- Explique en qué consisten las siguientes representaciones:
a- one hot

Sea V el vocabulario que vamos a utilizar. Nos referimos a los elementos de V por un índice natural $V = v_1\ v_2\ \ldots\ v_n$, siendo n=|V|

La representación one hot de la palabra k-ésima de V es un vector de largo n, con todos los

elementos iguales a 0 excepto el elemento k-ésimo que vale 1.

b- tf-idf

tf-idf es un escalar asociado a una palabra **t** en un documento **d**, dado un conjunto de documentos **D**. El objetivo es cuantificar el peso de **t** como descriptor de **d** entre todos los documentos de **D**. Se construye de modo tal que disminuye cuando la palabra se encuentra en mucho documentos de la colección y aumenta cuando se encuentra muchas veces en **d.**
Se puede construir a partir de las frecuencias crudas, o normalizando, o tomando logaritmos. El caso más simple sería:

tf (t,d) = frec(t,d)
idf(t,D) = 1/|t ϵ d ^ d ϵ D|
tf(t,d,T) = tf . idf

c- vectores generados por skip-gram

Se generan los vectores , cuya dimensión es habitualmente de algunos cientos de reales, mediante una red neuronal que se entrena para predecir los contextos próximos de las palabras (contextos en una ventana de tamaño pequeño alrededor de la palabra). Son vectores densos, y tienen propiedades interesantes vinculadas a su significado.