Asignatura	Datos del alumno	Fecha
Procesamiento del Lenguaje Natural	Apellidos:	
	Nombre:	

Trabajo: Etiquetado morfosintáctico

Objetivos

Aplicar un método basado en modelos ocultos de Markov (HMM) para realizar el etiquetado morfosintáctico de una oración.

Descripción

En esta actividad debes aplicar el método basado en modelos ocultos de Markov (HMM) para realizar el etiquetado morfosintáctico de una oración.

Parte 1: construir el etiquetador morfosintáctico

En esta primera parte de la actividad tienes que construir el etiquetador morfosintáctico basado en un HMM bigrama a partir de un corpus etiquetado. Para ello debes utilizar el corpus que te indicará el profesor en clase y en el foro. Además, el profesor también te indicara la parte de la cadena de Markov que tienes que crear.

Para construir el etiquetador con los datos de entrenamiento, calcula las probabilidades que rigen el HMM bigrama, es decir, las probabilidades de emisión y las probabilidades de transición del HMM. Además, dibuja el HMM indicando claramente las probabilidades en cada estado de la cadena de Markov.

Asignatura	Datos del alumno	Fecha
Procesamiento del Lenguaje Natural	Apellidos:	
	Nombre:	

Parte 2: Etiquetar morfosintácticamente una oración

En esta segunda parte de la actividad tienes que calcular la mejor secuencia de etiquetas para una oración, dicho de otro modo, realizar el etiquetado mofosintáctico la oración. Para ello debes utilizar el etiquetador que has construido en la parte 1 de esta actividad y aplicar el algoritmo de Viterbi. El profesor te indicará en clase y en el foro la oración que debes etiquetar.

Para aplicar el algoritmo de Viterbi, dibuja la matriz de probabilidades donde se representen claramente las observaciones y los estados de la máquina de estados finitos. Calcula el valor de Viterbi para cada celda de la matriz e indica claramente los valores obtenidos en la representación gráfica. Una vez se haya llegado al final del algoritmo, traza la ruta inversa para obtener la mejor secuencia de etiquetas. Indica claramente el resultado obtenido del etiquetado morfosintáctico de la oración estudiada.

Criterios de evaluación

- Se debe entregar un informe con la descripción detallada de todos los pasos seguidos para realizar la actividad y los resultados intermedios que se van obteniendo en cada paso. Además, se debe detallar toda la información que se solicita en el enunciado.
- Se valorará la explicación clara y argumentada.
- Si solo se presenta el resultado final, independientemente de que este sea correcto o no, la actividad se calificará con cero puntos.
- Si para resolver la actividad se implementa código (algo que no es obligatorio, ya que la actividad también se podría resolver a mano), este deberá entregarse como un anexo al informe. Sin embargo, este hecho no exime de presentar un informe con una descripción detallada de los pasos seguidos para realizar la actividad. Si

solo se entrega el código, independientemente de que este esté comentado o no, la actividad se calificará con cero puntos.

- No se pueden utilizar recursos disponibles para el procesamiento del lenguaje natural, por ejemplo, librerías, API… Si se utiliza alguno de estos recursos para resolver la actividad y no se realizan de forma manual todos los pasos descritos en el enunciado, la actividad se calificará con cero puntos.
- El informe de la actividad debe tener el formato marcado en el apartado **Extensión máxima**.
- La extensión máxima marcada para la actividad en el apartado **Extensión máxima** no se puede sobrepasar en ningún caso. Por lo tanto, si el informe presentado tiene más páginas que el valor establecido, solo se corregirán y evaluarán las *N* primeras páginas. Por ejemplo, si se presenta un documento con 10 páginas y la extensión máxima es de 8 páginas, se evaluará solamente el contenido de las páginas 1 a la 8, por lo que todo el contenido presentado en las páginas 9 y 10 no contará para la nota de la actividad.
- Se pueden añadir anexos al informe que contengan, por ejemplo, código, figuras, etc. Estos anexos no computan para la extensión máxima del trabajo, pero tampoco se valorará su contenido para la calificación final de la actividad. Por lo tanto, el informe debe ser entendible por sí solo y responder a todas las cuestiones planteadas en la actividad sin necesidad de recurrir al contenido presentado en los anexos.
- Si se detecta plagio en el informe o actividades que copien parte de las respuestas de otro alumno, todos los alumnos involucrados obtendrán una calificación para la actividad de cero puntos.

Extensión máxima: 8 páginas, fuente Calibri 12 e interlineado 1,5.

Asignatura	Datos del alumno	Fecha
Procesamiento del Lenguaje Natural	Apellidos:	
	Nombre:	

Anexo 1: Corpus

El corpus etiquetado que se debe utilizar para crear el etiquetador morfosintáctico se encuentra en el fichero de texto llamado **"mia07_t3_tra_Corpus-tagged-PER715"**.

El corpus se compone de 16 frases en español etiquetadas con conocimiento sobre las partes de la oración (categorías gramaticales o POS tags). Estas frases etiquetadas han sido extraídas de cinco documentos que forman parte de Wikicorpus[1]. Wikicorpus es un corpus trilingüe (español, catalán e inglés) compuesto de más de 750 millones de palabras. Wikicorpus fue creado por investigadores de la Universitat Politècncnica de Catalunya a partir de documentos de la Wikipedia que fueron anotados con la librería opensource FreeLing[2].

La Tabla 1 muestra en formato de texto plano y sin etiquetar las 16 frases que componen el corpus. De hecho, también se muestra el extracto del documento del cual han sido extraídas las frases etiquetadas.

doc id	texto
27315	(…) Tristana es una película del director español nacionalizado mexicano Luis Buñuel. Está basada en la novela del mismo nombre de Benito Pérez Galdós. Fue nominada al Oscar a la mejor película de habla no inglesa en 1970. (…)
216784	(…) En su primer viaje el comportamiento desadaptado y agresivo de Cole lleva a que sea apresado y recluido en un Centro Psiquiátrico, acusado de ser enfermo mental puesto que defiende venir del futuro y habla sobre un virus mortal del que nadie tiene sospechas. En este Centro Psiquiátrico conoce a la psiquiatra Kathryn Railly (Madeleine Stowe) y a un excepcional enfermo mental, Jeffrey Goines (Brad Pitt), con quienes entabla una particular relación que le permite establecer que un grupo radical probablemente ecoterrorista llamado Doce Monos podría ser responsable de la propagación del mortal virus. (…)
394089	(…) Mr. Bean habla muy pocas veces y cuando lo hace es siempre con pocas palabras y voz grave pero divertida. Tiene una novia "normal" (Irma Gobb) que a pesar de su banalidad lo adora y lo odia a la vez. (…)
986299	(…) Mientras la historia se desarrolla conoce a Denny Duquette, un paciente enfermo del corazón con quien entabla amistad. Al pasar los capítulos Denny regresa, pero aún

[1] http://www.cs.upc.edu/~nlp/wikicorpus/
[2] http://nlp.lsi.upc.edu/freeling/

	más enfermo del corazón por lo cual le ponen un bypass que no funcionó. Afortunadamente y a la vez desafortunadamente Denny obtiene un corazón en el programa de trasplantes, el corazón lo obtuvo puesto que 2 hermanos tienen un accidente, pero Burke al ir a recoger el corazón a otro hospital se "muere" el corazón que pertenecía a Denny pero pide el del otro hermano, pero se le niega pues el dueño de ese corazón está un poco más grave que Denny y se inscribió en el programa de trasplantes diecisiete segundos antes que Denny. Al saber la noticia, Izzie actúa sin pensar y corta los cables de Denny para simular un ataque grave y así ganar el corazón, pero para que no muera le tiene que cuidar hasta que su corazón quede débil; en este tiempo Denny le propone matrimonio a Izzie. (…)
1419147	(…) Poco después de que él comenzara nuevamente a trabajar, el capataz de Young, Bob Egle, enfermó repentinamente y murió. Young había hecho el té para sus colegas, envenenándolos con elementos tales como antimonio y talio. Debido a los envenamientos producidos por Young, muchos colegas caen enfermos y se habla de la posibilidad de un virus extraño, por lo que es apodado "Bovingdon Bug". (…) Luego de la muerte de Biggs, era obvio que tantas enfermedades y dos muertes requerían de una investigación médica y policial en el lugar de trabajo. Inexplicablemente, Young habla con el doctor de la compañía y le insinúa si él no creía que se trataba de envenenamiento por talio, debido a los síntomas. (…)

Tabla 1. Definiciones de las diferentes categorías morfosintácticas o categorías gramaticales según el diccionario de la lengua española de la Real Academia Española (RAE).

La versión anotada de las frases presentadas en la Tabla 1 conforman el corpus anotado proporcionado para realizar esta actividad. El formato del fichero de texto que contiene el corpus es el mismo que el utilizado en Wikicorpus. Por lo tanto, cada uno de los documentos de Wikipedia se identifica con el tag XML <doc> donde se indica el identificador del documento (id). Además, cada una de las frases en el documento viene separada por una línea en blanco. La información relativa a cada palabra de la frase se representa en una nueva línea del fichero. Para cada palabra, es decir en cada línea del fichero, se proporciona además del token que representa a la propia palabra, su lema, la etiqueta gramatical (POS tag) asociada a la palabra y el sentido de la palabra.

La Figura 1 muestra una captura del corpus anotado donde se observa la frase “Tristana es una película del director español nacionalizado mexicano Luis Buñuel.” perteneciente al documento de Wikicorpus con identificador 27315 y titulado Tristana. Si se analizan las anotaciones para la palabra “es” se observa que su lema es “ser” y que la categoría gramatical a la que pertenece esa palabra es la

identificada por la etiqueta gramatical "VSIP3S0" y que el sentido de la palabra es el identificado por el código "01775973". También se observa que la palabra "del" en la frase se representa en dos líneas y se anota con dos tokens, el primero "de" y el segundo "el". Esto se deba a que la palabra "del" es la contracción de la preposición "de" y el artículo "el". Por el contrario, el nombre propio "Luis Buñuel" que está formado por dos palabras (el nombre "Luis" y el apellido "Buñuel") se anota como un único token "luis_buñuel". Además, se observa que el punto final de la frase también viene anotado como un token ".".

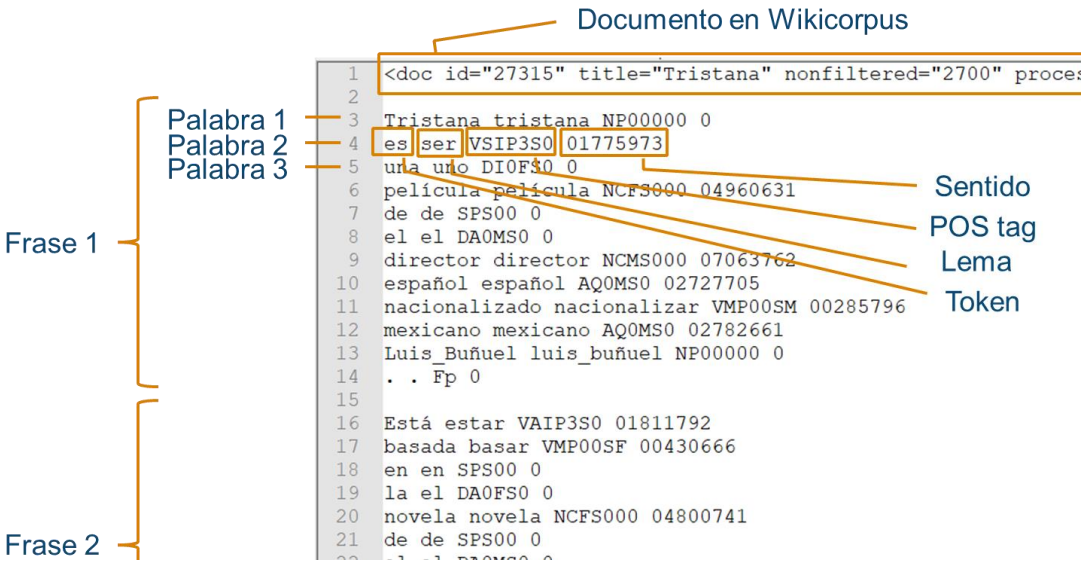

Figura 1. Muestra de parte del corpus etiquetado.

Aunque el corpus anotado proporciona más información (ver Figura 1), es importante darse cuenta de que para realizar esta actividad sólo será necesario el conocimiento sobre el token y la etiqueta gramatical (POS tag) de cada palabra. Es decir, la información contenida en el primer y el tercer string de cada línea que representa una palabra en el corpus anotado.

Las etiquetas gramaticales (POS tags) utilizadas para anotar la información morfosintáctica del corpus son las definidas en FreeLing y se basan en EAGLES, una

recomendación para la anotación de la mayoría de las lenguas europeas. La definición del conjunto de etiquetas gramaticales (POS tags) utilizadas por FreeLing en el etiquetado de un corpus en español se puede consultar en la siguiente página web:

https://talp-upc.gitbook.io/freeling-4-1-user-manual/tagsets/tagset-es

Las etiquetas gramaticales de EAGLES utilizadas por FreeLing son etiquetas de longitud variable donde cada carácter corresponde a una característica morfosintáctica. El primer carácter en la etiqueta es siempre la categoría gramatical o parte de la oración. Esa categoría gramatical determina la longitud de la etiqueta y la interpretación de cada uno del resto de caracteres en la misma.

La definición de la etiqueta para la categoría gramatical "verbo" se muestra en la Tabla 2. Entonces, la etiqueta **"VSIP3S0"**, con la que ha sido etiquetada la palabra "es" en la frase que se presentó anteriormente, se interpreta de la siguiente forma: se refiere a un verbo (V) de tipo semiauxiliar (S) en modo indicativo (I) y en tiempo presente (P) para la tercera persona (3) de (número) singular (S), y el carácter **"0"** al final de la etiqueta indica que esta forma verbal no tiene género.

Posición	Atributo	Valores
0	categoría	**V**:*verb*
1	tipo	**M**:*principal*; **A**:*auxiliar*; **S**:*semiauxiliar*
2	modo	**I**:*indicativo*; **S**:*subjunctivo*; **M**:*imperativo*; **P**:*participio*; **G**:*gerundio*; **N**:*infinitivo*
3	tiempo	**P**:*presente*; **I**:*imperfecto*; **F**:*futuro*; **S**:*pasado*; **C**:*conditional*
4	persona	**1**:*1*; **2**:*2*; **3**:*3*
5	número	**S**:*singular*; **P**:*plural*
6	género	**F**:*femenino*; **M**:*masculino*; **C**:*común*

Tabla 2. Definición de la etiqueta para la categoría gramatical "verbo".

Asignatura	Datos del alumno	Fecha
Procesamiento del Lenguaje Natural	Apellidos:	
	Nombre:	

Es importante destacar que para realizar la actividad se deben utilizar las etiquetas con las que se anota el corpus en formato EAGLES, por ejemplo, **"**`VSIP3S0`**"**. Si se utilizan otras etiquetas, la actividad será considerada incorrecta y puntuada con **cero puntos**.

Anexo 2: Oración a etiquetar

La oración para la que se debe realizar el etiquetado mofosintáctico, es:

"Habla de trasplantes con el enfermo grave."

Tienes que construir el etiquetador morfosintáctico basado en un HMM bigrama a partir del corpus etiquetado que se presenta en el *Anexo 1: Corpus* y disponible en el archivo de texto **"mia07_t3_tra_Corpus-tagged-PER715"**. No es necesario que crees toda la cadena de Márkov, sólo debes crear la parte necesaria para realizar el etiquetado mofosintáctico de la oración propuesta. Para crear sólo la parte necesaria del HMM debes seguir las instrucciones que se explican a continuación.

Para cada palabra de la oración extrae sus posibles etiquetas gramaticales según las diferentes apariciones de dicha palabra en el corpus etiquetado. Ten en cuenta que cada una de las palabras puede aparecer en el corpus realizando diferentes funciones gramaticales, es decir, ocupando varias partes de la oración, y por lo tanto generar posibles ambigüedades en el etiquetado morfosintáctico.

Una vez hayas determinado todas las categorías gramaticales posibles para las palabras de la oración, crea una cadena de Márkov donde los estados sean estas categorías gramaticales. Por lo tanto, las categorías gramaticales que no aparecen en

el análisis anterior no deben presentarse en cadena de Márkov. Recuerda que también debes representar los estados inicial y final.

En la cadena de Márkov debes representar las transiciones entre estados.