PYTHON PARA LINGÜISTAS

PART OF SPEECH TAGGING

ALEJANDRO ARIZA

CENTRE DE LLENGUATGE I COMPUTACIÓ

UNIVERSITAT DE BARCELONA

¿QUÉ SABEMOS POR AHORA?

- Fundamentos de la programación:
  - Tipos de datos y variables
  - Funciones y métodos
  - Estructuras de decisión
  - Bucles
- Preprocesado básico de corpus:
  - Limpieza
  - Tokenización y segmentación de frases
  - Extracción de estadísticas básicas y n-gramas

ÚLTIMAS CLASES DEL CURSO: TAREAS LINGÜÍSTICAS

- En las próximas clases, exploraremos diferentes tareas relacionadas con la lingüística
- Part of Speech Tagging
- Análisis sintáctico
- Clasificación de texto general

PARTS OF SPEECH

- El “part of speech” es una categoría que se le asigna a cada palabra de acuerdo a sus funciones
- Existen diferentes conjuntos de etiquetas:
  - Los conjuntos simplificados únicamente marcan la categoría principal: “nombre”, “verbo”, “adjetivo”...
  - Otros conjuntos más complejos contienen una variedad de características gramaticales: número, persona, forma verbal
  - Originariamente, los conjuntos de etiquetas eran específicos para cada lenguaje
  - Con el desarrollo de CL y NLP, se introdujeron los conjuntos de etiquetado universales

PART OF SPEECH TAGGING

- Part of speech tagging (o POS-tagging) es un proceso que asigna una etiqueta POS a cada palabra de un texto:
  - jugando: verbo, estatua: nombre, dirección: ?

- Un problema con el etiquetado automático de POS es la ambigüedad de algunas palabras
  - Muchas palabras pueden tener más de una etiqueta potencial
  - (to) address (VB) someone vs someone's address (NN)

DESAMBIGUACIÓN POS

- Para desambiguar entre múltiples etiquetas necesitamos mirar el contexto:
  - The reason why I address you today is that...
  - The courier delivered the gift to the address.
- ¿Cuál es el contexto?
  - Lineal – un número limitado de palabras antes y después (recordad los n-gramas y Markov)
  - ¿Por qué no sintáctico? – el análisis sintáctico es más complicado que el etiquetado POS; a menudo necesitamos las etiquetas POS antes de poder realizar un análisis sintáctico

ETIQUETADO POS BASADO EN N-GRAMAS

- La idea principal: predecir la etiqueta POS de una palabra basándonos en la palabra actual y un contexto predefinido de palabras adyacentes
  - Recordad la idea detrás de un modelo de lenguaje – predecir la siguiente palabra dadas las anteriores

- ¿Cuántas palabras incluimos en ese contexto?
- ¿Cómo hacemos las predicciones exáctamente?

ELECCIÓN DE N-GRAMA

- Un modelo de unigramas predice la etiqueta usando únicamente la palabra actual (i.e.: “book” siempre será un nombre)
- Un modelo de bigramas predice la etiqueta basándose tanto en la palabra actual como en la anterior (“the book” asingará “noun” a “book”; “I book” asignará “verb”)
- Un modelo trigrama utilizará 3 palabras consecutivas para predecir la etiqueta (“mouse or book” vs “cancel or book”)
- N-gramas de mayor orden son más potentes pero cuesta más calcularlos y nos encontramos con un problema denominado “data sparsity” (lo veremos más adelante)

N-GRAM TAGGERS: APRENDER A ETIQUETAR

- Los etiquetadores basados en N-gramas seleccionan la etiqueta más probable para cierta palabra (o n-grama) basándose en las ocurrencias de un corpus.
- Un etiquetador por unigramas contará en un corpus el número de veces que aparece “book” en el corpus como nombre y como verbo. Entonces, predecirá la etiqueta que aparece con mayor frecuencia.
- Un etiquetador por bigramas hará lo mismo que el anterior pero tomará la frecuencia de “book” siendo nombre para cada bigrama en el que aparece la palabra “book” en segundo lugar. A la hora de predecir, si el bigrama es “the book” devolverá la etiqueta de “book” más frecuente para ese bigrama.

ENTRENAR UN ETIQUETADOR POS POR N-GRAMAS

- Lo primero que necesitamos es un corpus anotado con POS
- Seleccionamos el tamaño de los n-gramas
- Entrenamos el etiquetador POS
  - El etiquetador POS obtiene las frecuencias de todos los n-gramas en el corpus junto a sus POS.
  - Entonces, calcula las probabilidades para las etiquetas POS de todos los n-gramas.
- Finalmente, usamos el etiquetador POS en otro corpus no anotado
  - También podéis comprobar la efectividad del etiquetador en una porción del corpus anotado que no hayáis utilizado para entrenarlo

PROBLEMAS (POTENCIALES) CON EL ENTRENAMIENTO

- El etiquetador es tan bueno como el corpus utilizado para entrenarlo
  - Si “book” siempre es un verbo en el corpus de entrenamiento, el etiquetador siempre lo anotará como verbo
  - Si “book” nunca aparece en el corpus de entrenamiento, el etiquetador no será capaz de asignarle ninguna etiqueta. Este problema se denomina “data sparsity” (escasez de datos).
- Data sparsity es un grave problema cuando se trabaja con n-gramas de mayor orden (3,4,5):
  - Es muy difícil que el corpus de entrenamiento contenga “mouse or book”
  - Para resolver este problema, utilizaremos una técnica llamada “backoff”

ETIQUETADO POS CON NLTK

- NLTK tiene corpus anotados y funciones para el etiquetado POS
- Para entrenar un etiquetador con NLTK, es necesario:
  - Importar un corpus anotado con POS
  - Elegir un etiquetador
  - Entrenar el etiquetador
- Entonces, podremos evaluar el etiquetador o etiquetar un nuevo corpus

APRENDER DE LOS DATOS: ENTRENAMIENTO Y EVALUACIÓN

- El etiquetado POS es un ejemplo de “aprender de los datos”, o en particular “aprendizaje supervisado”.
  - Creamos ejemplos con el resultado deseado (en este caso, un corpus anotado con POS)
  - Entrenamos el modelo para que aprenda de los ejemplos (la relación palabra/n-grama y la etiqueta correspondiente)
  - Típicamente, evaluamos el sistema usando datos que no ha visto en su entrenamiento y monitorizamos el grado de aciertos y fallos que tiene
  - El conjunto de datos de evaluación y entrenamiento deben salir de la misma distribución de datos pero es importante que el modelo no haya visto los ejemplos del test para obtener un resultado no sesgado del rendimiento del sistema
  - El “aprendizaje supervisado” es similar a como aprendemos los humanos (clases + examen)

BACKOFF

- Backoff es una técnica utilizada para lidiar con la escasez de datos
- El “backoff” combina múltiples etiquetadores en uno solo:
  - Sabemos que utilizar n-gramas de mayor orden es más preciso pero nos enfrentamos a menos casos de los que aprender
  - Primero, miramos si el etiquetador que utiliza n-gramas de mayor orden es capaz de darnos una etiqueta
  - Si el etiquetador no conoce ese n-grama, reducimos la n en 1 y preguntamos al etiquetador correspondiente.
  - Continuamos el proceso hasta que un etiquetador nos devuelve una etiqueta. En última instancia, encontraremos el etiquetador por unigramas que es más probable que haya visto la palabra actual en el corpus de entrenamiento

LA PRÁCTICA DE HOY

- Exploraremos diferentes etiquetadores POS
- Veremos su comportamiento y precisión
- Comprobaremos la importancia de usar backoff
- Veremos la importancia de seleccionar un buen corpus de entrenamiento

¡GRACIAS!
¿PREGUNTAS?