Procesamiento del lenguaje natural
Introducción y modelos de lenguaje

INTRODUCCIÓN

Procesamiento del lenguaje natural

- El procesamiento del lenguaje natural (PLN) es un área de intersección entre la Inteligencia Artificial y la Lingüística, que estudia la comunicación entre personas y máquinas por medio del lenguaje natural.
  - Sistemas capaces de interpretar o generar lenguaje natural

- El lenguaje natural es una forma de comunicación *imprecisa* y *ambigua* que se apoya en el conocimiento *compartido* por los que se comunican

Procesamiento del lenguaje natural

- Tareas de análisis de lenguaje natural
  - ❑ Análisis morfo-léxico, sintáctico, semántico y pragmático

- Aplicaciones
  - Recuperación de información
  - Clustering y clasificación de textos
  - Análisis de sentimientos
  - Extracción de entidades y relaciones
  - Traducción automática
  - Generación automática de resúmenes
  - Sistemas de diálogo
  - Generación de lenguaje natural

Análisis del lenguaje natural

- Gran dificultad debido a que el lenguaje es algo vivo: en continua expansión, que va modificándose...
  - El lenguaje se modifica tanto en vocabulario como en sintaxis
- Existencia de jergas locales, profesionales, por franjas de edad...
- Ambigüedad
  - Léxica: *Entró en el banco. Se sentó en el banco.*
    - Polisemia: ambigüedad de las palabras
  - Sintáctica: *Juan vio a María con unos prismáticos.*
    - A veces es imposible de solucionar
  - Semántica: *Los niños compraron el libro de Peter Pan.*
  - Referencial: *El jamón está en el armario. Sácalo. Ciérralo.*
- Requiere mucho conocimiento: objetivos del hablante, hipótesis, contexto... No es mera transmisión de palabras...

Niveles de análisis del lenguaje I

- Tokenization:
  - Frases -> Palabras
    - “kids made good snacks.”=[“kids”, “made”, “good”, “snacks”, “.”]
- Análisis morfologico:
  - Palabra -> raiz + sufijos/prefijos + feature
    - kids=kid+s+<plural>
    - made=make+<past tense>
- Análisis léxico:
  - Palabras -> Etiquetas léxicas (POS tags)
    - [“kids/NN”, “made/V”, “good/ADJ”, “snacks/NN”, “./PUNCT”]

Niveles de análisis del lenguaje II

- Análisis sintáctico (parsing):
  - Palabras + POS tags -> Estructura de la frase
  - 2 tipos:
    - Análisis de constituyentes.
    - Análisis de dependencias.
- Análisis semántico:
  - Palabras + POS tags -> Significado de la palabra (desambiguación)
  - Significado de palabras + estructura de frase: significado de frase (normalmente, independientemente del contexto)
- Resolución de referencias:
  - Expresiones de referencia + contexto: referente semántico
- Análisis Pragmático:
  - Significado de frase + contexto: significado más profundo

Ejemplo de herramienta de PLN

- Demo de herramienta de PLN
  - Freeling:
    - http://nlp.lsi.upc.edu/freeling/demo/demo.php
  - Se pueden configurar distintas opciones de análisis y distintas configuraciones de salida, en distintos idiomas

Etapas históricas

- Podemos distinguir 3 grandes etapas históricas
  - PLN simbólico (1950s – 1990s): Técnicas lingüísticas formales
    - Se basan en el desarrollo de reglas estructurales que se aplican en las distintas fases del análisis.
    - Encaje de patrones, gramáticas, sistemas de reglas, ontologías, etc.
  - PLN estadístico (1990s-2010s): Técnicas empiricistas o probabilísticas
    - Se basan en el estudio de una serie de características de tipo probabilístico asociadas a las distintas fases del análisis del lenguaje.
    - Estas características son extraídas de un conjunto de textos de referencia (corpus)
    - Aprendizaje supervisado, semi-supervisado y no supervisado
  - PLN y redes neuronales (2010s-presente): Deep Learning
    - Arquitecturas específicas para trabajar con lenguaje natural
    - Requieren grandes corpus de documentos y poder computacional
    - Grandes avances en los últimos años en áreas como la traducción automática o los sistemas conversacionales

Procesamiento del lenguaje natural

- El problema de los métodos basados en técnicas lingüísticas formales es la dificultad de codificar manualmente todo el conocimiento lingüístico necesario (diccionarios, gramáticas, etc.)
  - Esto obliga a trabajar con un lenguaje reducido
  - Este tipo de métodos no suelen contemplar la capacidad de aprendizaje
- En cambio, los métodos basados en técnicas probabilísiticas aprenden a partir de datos prácticos (corpus de documentos)
- Los sistemas que utilizan métodos necesitan una fase de entrenamiento en la que se les debe proporcionar un número suficiente de ejemplos
  - Corpus anotado
- El uso de redes neuronales profundas permite encontrar patrones estadísticos complejos en los corpus de documentos

MODELOS DE LENGUAJE

Modelos de lenguaje probabilísticos

- Un modelo probabilístico del lenguaje define una distribución de probabilidad sobre el conjunto de elementos a partir de los valores observados en un corpus de documentos
  - Según cual sea el objeto de análisis los elementos pueden ser fonemas, letras, sílabas, o palabras
  - Las frecuencias de aparición de cada uno de los elementos son las que se observen en el corpus

- Son realmente útiles hoy día en multitud de tareas de PLN
  - Texto predictivo: P(Qué tal **estás**) > P(Qué tal **has comido**)
  - Traducción automática: P(Voy de visita a su **casa**) > P(Voy de visita a su **hogar**)
  - Corrección ortográfica: P(Tenemos **calor**) > P(Tenemos **color**)
  - Reconocimiento del habla: P(Se hizo daño **a sí** mismo) > (Se hizo daño **así** mismo)

Modelos de lenguaje probabilísticos

- Vamos a considerar que nuestros elementos son palabras, pero lo que veamos aplica igual para fonemas, sílabas, etc.

- Un modelo probabilístico del lenguaje permite entre otras cosas
  - Calcular la probabilidad de encontrar una frase o secuencia de palabras determinada
    - $P(\text{Yo}, \text{quiero}, \text{comer}, \text{macarrones}, \text{con}, \text{tomate})$
  - Calcular la probabilidad de la siguiente palabra
    - $P(\text{tomate} \mid \text{Yo}, \text{quiero}, \text{comer}, \text{macarrones}, \text{con})$

- Como el texto es secuencial, podemos pensar que la probabilidad de una palabra depende de todas las anteriores
  - Para ello necesitamos refrescar ciertas nociones de probabilidad

Teoría de probabilidad aplicada a PLN

- Probabilidad condicionada: La probabilidad de B habiendo observado A
  - $P(B|A) = P(A,B) \,/\, P(A)$
- La probabilidad de que una palabra sea (por ejemplo) 'perro', sabiendo que la primera palabra es 'el', es la fracción de veces que 'el' aparece seguido de 'perro' en nuestro corpus
  - $P(\text{perro} \mid \text{el}) = numVeces(\text{el}, \text{perro}) \,/\, numVeces(\text{el})$
- La probabilidad condicionada se reescribe como $P(A,B) = P(A)\,P(B \mid A)$
  - Si añadimos elementos: $P(A,B,C,D) = P(A)P(B|A)\,P(C|A,B)P(D|A,B,C)$
  - La probabilidad de cada palabra depende de todas las anteriores
- Esto se generaliza mediante la regla de la cadena

$$P(w_1,\ldots,w_n) = P(w_1)\;\; P(w_2 \mid w_1)\;\; P(w_3 \mid w_1,w_2) \ldots\; P(w_n \mid w_1,\ldots,w_{n-1})$$

Relajando la hipótesis de la regla de la cadena

- Según la regla de la cadena, la probabilidad de una palabra depende de todas las anteriores
  - Sin embargo, esta hipótesis es impracticable porque no hay un corpus tan grande para asignar probabilidad a las posibles combinaciones de palabras
    - Siempre habrá alguna combinación que no esté presente en el corpus
  - Se puede relajar la hipótesis de la regla de la cadena, haciendo así factibles los cálculos y obteniendo excelentes resultados

- En lugar de considerar que la probabilidad de un elemento depende de todos los anteriores, supone que solamente los n-1 elementos anteriores tienen efecto sobre las probabilidades del siguiente elemento i-ésimo.
  - Hipótesis de Markov: $P(w_i|w_1, \dots, w_{i-1}) \approx P(w_i|w_{i-(n-1)}, \dots, w_{i-1})$

- Esta hipótesis no tiene en cuenta que en lenguaje puede haber dependencias de “larga distancia”, sin embargo, funciona muy bien para algunas tareas

Modelos probabilístico de los n-gramas

- El modelo n-grama es uno de los modelos estadísticos del lenguaje más simples pero más útiles.
  - Se puede aplicar a fonemas, letras, sílabas, palabras… aquí nos centraremos en palabras

- El modelo n-grama utiliza la hipótesis de Markov para indicar que la dependencia es con los n-1 anteriores. Por ejemplo:
  - Modelo bigrama (n = 2): $P(w_n|w_{n-1})$
  - Modelo trigrama (n = 3): $P(w_n|w_{n-1}, w_{n-2})$

Ejemplo de estimación de bigramas

- Supongamos este corpus
  - <s> Yo quiero patatas </s>
  - <s> Yo quiero patatas con carne </s>
  - <s> No quiero carne </s>
  - <s> Quiero carne con patatas </s>
  - <s> Hoy quiero carne </s>
  - <s> Quiero dormir </s>

Estimación optima por máxima verosimilitud

$$P(w_n|w_{n-1}) = \frac{frec(w_{n-1}, w_n)}{frec(w_{n-1})}$$

<s> indica "inicio de oración"
</s> indica "final de oración"

- Estos son las probabilidades de algunos bigramas del corpus
  - P(yo|<s>)= frec(<s>, yo) / frec(<s>) = 2/6 = 0,333
  - P(no|<s>)= frec(<s>, no) / frec(<s>) = 1/6 = 0,167
  - P(quiero|yo)= frec(yo, quiero) / frec(yo) = 2/2 = 1
  - P(patatas|quiero)= frec(quiero, patatas) / frec(quiero) = 2/6 = 0,333
  - P(carne|quiero)= frec(quiero, carne) / frec(quiero) = 3/6 = 0,5
  - P(dormir|quiero)= frec(quiero, carne) / frec(quiero) = 1/6 = 0,167

- Según este modelo, la continuación más segura de <s>Yo quiero…
  - **P(carne|quiero)** > P(patatas|quiero) > P(dormir|quiero)
  - La respuesta cambia si consideramos trigramas → P(patatas | yo, quiero) =1

Estimación de la probabilidad de una frase

- Probabilidades de bigramas obtenidos de un corpus supuesto
  - P(yo | <s>) = 0,25 P(quiero | <s>) = 0,75
  - P(quiero | yo)= 0,5 P(tengo| yo)=0,2 P(soy | yo)=0,3
  - P(ser | quiero) = 0,8 P(tomar | quiero)=0,2
  - P(café | tomar) = 0,6 P(leche | tomar) = 0,3 P(distancia | tomar) = 0,1
  - P(artista| ser) = 0,9 P(informático| ser) = 0,1

- P(*Yo quiero tomar café*) =
  - P(yo | <s>) P(quiero | yo) P(tomar | quiero) P(café | tomar) = 0,25 *0,5*0,2*0,6 =0,015
  - La frase entera puede no estar en el corpus. Lo normal es que no esté.
  - ¡Si un bigrama no está la probabilidad de la frase es cero!

- La frase que se genera siguiendo la opción más probable es: "*Quiero ser artista*"
  - P(quiero | <s>) P(ser | quiero) P(quiero | artista) = 0,75*0,8*0,9 = 0,54

Estimación de probabilidad en casos raros

- Como hemos visto cualquier n-grama que no esté en el corpus recibe probabilidad 0, el alisado de Laplace alivia los problemas de estimación de probabilidades en casos raros
- El alisado de Laplace calcula cualquier probabilidad condicional considerando que ha habido unas observaciones adicionales virtuales de todos y cada uno de los n-gramas posibles
- Siendo AB un bigrama (observado o no), el valor de la probabilidad condicional P(B |A) alisado según Laplace

$$\mathrm{P}(\mathrm{B}|\mathrm{A}) = \frac{\mathrm{frec}(\mathrm{A},\mathrm{B}) \ + t}{\mathrm{frec}(\mathrm{A}) + t * m}$$

Donde

- t es el número de observaciones virtuales adicionales
- m es el número de monogramas (palabras) existentes en el corpus

- Las probabilidades totales siguen sumando 1

Alisado por interpolación lineal

- El alisado por interpolación lineal es ligeramente más sofisticado
- La interpolación lineal usa la probabilidad incondicional $\mathrm{P}(\mathrm{w}_2)$ calculada a partir de los datos para hacer que la probabilidad condicional $\mathrm{P}(\mathrm{w}_2|\mathrm{w}_1)$ se parezca a ella, de la siguiente forma

$$\mathrm{P}_{\mathrm{Int}}(\mathrm{w}_2|\mathrm{w}_1) = \alpha \mathrm{P}(\mathrm{w}_2|\mathrm{w}_1) + (1-\alpha)\mathrm{P}(\mathrm{w}_2)$$

donde $\alpha \epsilon [0,1]$ regula el peso que se le da a la probabilidad condicional y la probabilidad no condicionada

- El valor de α se puede fijar empíricamente con el fin de ajustar el rendimiento
  - También se puede hacer dependiente del "contexto"
    - Si existen muchos bigramas con la palabra w_1 entonces es mejor un valor alto
    - Si no existen muchos bigramas con la palabra w_1 es mejor un valor bajo

Enlaces

- Enlaces interesantes:
  - **Modelos de lenguaje y n-gramas**
    - https://heuristic-bhabha-ae33da.netlify.app/modelos-de-lenguaje-y-n-gramas.html