PROYECTO INTEGRADOR (GPO 10)

Avance 3

PROFESORES TITULARES

DRA. GRETTEL BARCELÓ ALONSO

DR. LUIS EDUARDO FALCÓN MORALES.

EQUIPO 31

JUAN CARLOS VILLAMIL ROJAS A01794003

MATEO CRUZ LANCHERO A01793882

ANDREA MARGARITA OSORIO GONZÁLEZ A01104776

Fecha:19 de mayo de 2024

El funcionamiento del PLN implica varias etapas complejas que imitan el proceso de comprensión del lenguaje humano por parte de los seres humanos:

1. **Tokenización**: El texto se divide en unidades más pequeñas, como palabras o frases (llamadas tokens).
2. **Análisis morfológico:** Se identifican las estructuras gramaticales y las formas de las palabras.
3. Análisis sintáctico: Se determina la estructura gramatical y la relación entre las palabras en una oración.
4. **Análisis semántico:** Se comprende el significado de las palabras y las oraciones en un contexto determinado.
5. **Análisis pragmático:** Se tiene en cuenta el contexto más amplio y las intenciones del hablante.
6. **Generación de texto:** Se produce una respuesta coherente o acción basada en la comprensión del texto de entrada.

Clasificación de los sistemas GLN (Generación de Lenguaje Natural)

Clasificación	Tipos	Definición
Entrada al sistema	*T2T*	Reciben como entrada texto u oraciones
	D2T	Reciben como entrada datos generalmente numéricos, bases de datos, bases de conocimiento o corpus etiquetados
Objetivo del sistema: tipo de texto producido	Textos informativos	Generan informes a partir de un conjunto de datos
	Resúmenes	Generan texto que incluye las ideas fundamentales de uno o varios documentos
	Borradores	Generan versiones preliminares de textos técnicos
	Textos simplificados	Simplifican el pasaje de entrada para facilitar su comprensión
	Textos persuasivos	Generan relatos con el objetivo de convencer o motivar a los usuarios sobre un tema
	Sistemas de diálogos	Generan producciones interactivas manteniendo una comunicación entre el usuario y el sistema
	Explicaciones de razonamiento	Exponen en un texto los pasos seguidos en la resolución de un problema
	Recomendaciones	Generan sugerencias y valoraciones relativas a lugares, productos, servicios, etc

Según la entrada del sistema

Según el tipo de entrada que se introduce en el sistema se consideran dos posibles enfoques en la *GLN:* datos-a-texto *(D2T: data-to-text)* y texto-a-texto *(T2T: text-to-text).* Mientras que en la perspectiva *D2T la* entrada al sistema es un conjunto de datos que no conforman un texto (p. ej., datos numéricos representando información meteorológica), en el enfoque *T2T* el sistema sí parte de un texto del que se extrae la información relevante para construir la salida.

Generación de resúmenes.

Este tipo de generación tiene como objetivo producir una versión abreviada de una o más fuentes de información

Arquitectura:

- ***Macro planificación.*** El primer módulo del sistema debe determinar qué decir y organizarlo en una estructura coherente, dando lugar a un plan del documento. Lo hace mediante dos tareas:
    - Selección de contenido
    - Estructuración del documento
- ***Micro planificación.*** Partiendo del plan del documento que llega como entrada desde el módulo anterior, se generará una planificación del discurso. Se seleccionan las palabras y las referencias adecuadas, se dota a los mensajes de una estructura lingüística y se agrupa la información en oraciones. Las tareas que intervienen son las siguientes:
    - Agregación de sentencias
    - Lexicalización de estructuras sintácticas
    - Generación de expresiones referenciales
- ***Realización.*** A estas alturas del proceso, se dispone de una representación de las oraciones que van a conformar la salida del sistema. El módulo de realización genera la salida final, sea ésta texto o habla, las oraciones concretas que la conforman así como la estructura que hayan de presentar. Dos tareas se suceden finalmente:
    - Realización lingüística
    - Realización de la estructura

Selección de contenido

La selección de contenido es la tarea que permite al sistema elegir y obtener la información que debería ser comunicada en el texto final: la más relevante para el usuario acorde con el objetivo comunicativo y la situación, que incluye aspectos tan diversos como el tamaño que corresponde a la salida del sistema, el nivel de conocimiento del usuario o la historia del discurso hasta el momento.

Estructuración del documento

Para conseguir un texto coherente es preciso que los elementos que lo configuran estén debida-mente estructurados. Cohesión y coherencia son los principios que permiten que un conjunto de oraciones constituyan un discurso. Hacen referencia al modo en que las unidades textuales se relacionan entre sí y permiten que se puedan realizar inferencias a partir de la información proporcionada o que se puedan identificar de forma no ambigua los elementos correferentes. Esa coherencia compete tanto a las oraciones como a los mensajes que las componen.

Por todo ello, se hace necesaria una tarea que, ya sea durante el proceso de seleccionar los mensajes o después de haberlo hecho, determine la estructura que vaya a tener en el texto final, la relación que guardan unos elementos con otros, dado que tal ordenación supone el primer paso hacia el discurso correcto.

Lexicalización

La lexicalización es la etapa de la *GLN* que se encarga de seleccionar las palabras específicas o estructuras sintácticas concretas con las que referirse al contenido seleccionado en fases anteriores.

Cuando se dispone de varias opciones deben considerarse aspectos como el conocimiento y preferencias de los usuarios, la consistencia tanto con el léxico ya empleado como con la historia del discurso o la relación con las tareas de agregación y Generación de Expresiones Referenciales *(GER)*

Desafíos en el Procesamiento del Lenguaje Natural (PLN). Algunos de los más comunes son:

1. **Ambigüedad:** El lenguaje humano es inherentemente ambiguo. Las palabras pueden tener múltiples significados según el contexto, lo que dificulta la interpretación precisa.

2. **Variabilidad lingüística:** Las personas hablan y escriben de manera diferente según su región, cultura o nivel educativo. El PLN debe lidiar con esta variabilidad.

3. **Resolución de correferencias**: Identificar a qué se refiere un pronombre o una palabra como “esto” o “aquello” puede ser complicado.

4. **Análisis sintáctico y gramatical:** Determinar la estructura gramatical correcta de una oración es un desafío, especialmente en idiomas con reglas complejas.

5. **Recursos limitados**: El PLN depende de grandes cantidades de datos etiquetados para entrenar modelos. La falta de datos puede afectar la calidad del procesamiento.

6. **Traducción automática:** Lograr traducciones precisas entre idiomas es difícil debido a las diferencias gramaticales y culturales.

7. **Entendimiento del contexto:** Comprender el contexto es crucial para interpretar correctamente el significado de una oración.

8. **Generación de texto coherente:** Crear respuestas naturales y coherentes es un desafío, especialmente en sistemas de chat o asistentes virtuales.

Las siguientes son algunas de las preguntas que deberán abordar durante esta fase:

1. ¿Qué algoritmo se puede utilizar como baseline para predecir las variables objetivo?

- Se utilizaron varios algoritmos de clasificación, como Regresión Logística, Naive Bayes, SVM y Árboles de Decisión, para predecir si una reunión se clasificaría como "Alta" en función de la frecuencia de las palabras clave. Estos algoritmos pueden servir como baseline para el problema de generación de resúmenes.Sin embargo, para generar resúmenes, lo más adecuado sería utilizar algoritmos basados en redes neuronales, como modelos de lenguaje (por ejemplo, transformers como BERT o GPT) o técnicas de extracción de resúmenes como

TextRank o LSA (Latent Semantic Analysis). Estos algoritmos están diseñados específicamente para tareas de procesamiento de lenguaje natural y generación de texto.

2. ¿Se puede determinar la importancia de las características para el modelo generado?

- Se utilizó la selección de características basada en chi-cuadrado para identificar las palabras clave más relevantes para distinguir las reuniones de alta importancia. Esto nos da una indicación de la importancia de ciertas palabras en la clasificación de las reuniones. Para determinar la importancia de las características en un modelo de generación de resúmenes, se pueden utilizar técnicas como la atención (attention mechanism) en modelos de redes neuronales, que permiten visualizar qué partes del texto de entrada son más relevantes para generar el resumen. También se pueden analizar los pesos o coeficientes de los modelos para identificar las características más influyentes. Como se comentó en la respuesta al punto anterior, para la próxima entrega se abordará un tipo de algoritmo más apropiado.

3. ¿El modelo está sub/sobreajustando los datos de entrenamiento?

- En el ejercicio anterior, se utilizó la validación cruzada estratificada para evaluar el rendimiento de los modelos y se compararon las métricas de accuracy, precisión, recall y F1-score entre los conjuntos de entrenamiento y prueba.

Basándonos en los resultados obtenidos:

Rendimiento en el conjunto de entrenamiento:

- Accuracy: 0.8000
- Precisión: 0.8000
- Recall: 1.0000
- F1 Score: 0.8889

Rendimiento en el conjunto de prueba:

- Accuracy: 0.7500
- Precisión: 0.7500
- Recall: 1.0000
- F1 Score: 0.8571

Podemos observar que el rendimiento del modelo en el conjunto de entrenamiento es ligeramente mejor que en el conjunto de prueba. La accuracy y la precisión son un poco más altas en el conjunto de entrenamiento (0.8000) en comparación con el conjunto de prueba (0.7500). Sin embargo, la diferencia no es muy significativa.

El recall es del 1.0000 tanto en el conjunto de entrenamiento como en el conjunto de prueba, lo que indica que el modelo está identificando correctamente todas las instancias positivas en ambos conjuntos.

El F1 Score, que es una media armónica de la precisión y el recall, es ligeramente superior en el conjunto de entrenamiento (0.8889) en comparación con el conjunto de prueba (0.8571), pero la diferencia no es muy grande.

Dado que las métricas de rendimiento son similares entre el conjunto de entrenamiento y el conjunto de prueba, y no hay una diferencia significativa, podemos concluir que no hay evidencia clara de un sobreajuste o subajuste en el modelo.

Sin embargo, es importante tener en cuenta que esta evaluación se basa en un único conjunto de pruebas y que el tamaño de los datos puede ser limitado. Para obtener una evaluación más robusta, se podrían considerar técnicas adicionales, como la validación cruzada, donde se realizan múltiples divisiones de entrenamiento y prueba para obtener una estimación más confiable del rendimiento del modelo.

- Para evaluar el ajuste de un modelo de generación de resúmenes, se deberán utilizar métricas como ROUGE (Recall-Oriented Understudy for Gisting Evaluation) o BLEU (Bilingual Evaluation Understudy), que comparan los resúmenes generados con resúmenes de referencia creados por humanos. Si las puntuaciones son significativamente diferentes entre los conjuntos de entrenamiento y prueba, podría indicar un sobreajuste o subajuste.

4. ¿Cuál es la métrica adecuada para este problema de negocio?

- Para evaluar la calidad de los resúmenes generados, las métricas más comunes son ROUGE y BLEU, que miden la similitud entre los resúmenes generados y los resúmenes de referencia en términos de precisión y recall de n-gramas.

- Además de las métricas automáticas, es importante considerar la evaluación humana de los resúmenes generados, ya que pueden capturar aspectos como la coherencia, la relevancia y la legibilidad, que son fundamentales para la utilidad de los resúmenes en un contexto empresarial.

5. ¿Cuál debería ser el desempeño mínimo a obtener?

- El desempeño mínimo a obtener dependerá de los requisitos y expectativas específicas del negocio. Es importante establecer un punto de referencia basado en el rendimiento de modelos existentes o en la calidad de los resúmenes generados manualmente.

- Se pueden establecer umbrales mínimos para las métricas ROUGE o BLEU, así como criterios de evaluación humana, como la coherencia, la relevancia y la legibilidad de los resúmenes generados. El desempeño mínimo debe ser lo suficientemente bueno para que los resúmenes sean útiles y valiosos para los stakeholders de la empresa.