Universidad de Buenos Aires
Departamento de Computación

Métricas de Evaluación NLP

Temas de Procesamiento del Lenguaje Natural - NLP Aplicado

Dra. Viviana Cotik

2do bimestre 2025

Sistemas *Sequence to Sequence*

Sistemas, cuya salida es texto en lugar de una clasificación del texto de entrada.

Algunas tareas

- Machine translation
- Resúmenes automáticos
- Chatbots
- Question Answering (QA)

Evaluación:

- Evaluación humana era el estándar
- No es trivial

Tipos de Question Answering

Los sistemas de QA pueden diferir en función de:

- cómo se crean las respuestas
    - **Extractive QA**: El modelo extrae la respuesta de un contexto y se la proporciona directamente al usuario. El contexto podría ser un texto, una tabla, HTML, etc. Suele resolverse con modelos tipo BERT.
    - **Generative QA**: El modelo genera texto libre directamente en función del contexto. Aprovecha los modelos de generación de texto.
    - **Closed QA:** No se provee contexto y la respuesta es completamente generada por el modelo.

Métricas de Evaluación

- Precision, Recall, F1
- BLEU, ROUGE, METEOR, and F1
  - Usadas normalmente para traducción.
  - Computadas usando similaridad de n-gramas. Tienen inconvenientes conocidos.
  - Dependiendo el tipo de Question Answering puede diferir el tipo de métrica a utilizar

N-grams (n-gramas)

Subsecuencia contínua de n elementos de una secuencia dada.

Los elementos pueden ser palabras, sílabas, caracteres, etc.

n=1: Unigrama, n=2: bigrama, n=3: trigrama

Ej: “ser o no ser”

- bigramas de palabras: “ser o”, “o no”, “no ser”
- bigramas de caracteres: “se, er, r_, _o, o_, _n, no, o_, _s, se, er ”

n-grams similarity

- Jaccard Similarity
  - $Jaccard(A,B) = |A \cap B| / |A \cup B|$, A, B conjuntos de n-gramas extraídos de distintos textos.
- Coeficiente de Dice
  - $Dice(A,B) = 2 \cdot |A \cap B| \ / \ (|A| + |B|)$
- Overlap de n-gramas - Precision -Recall
- BLEU, ROUGE

Precision

Cantidad de palabras en el texto que también están en el texto de referencia.

$$Precision = \frac{(\text{Number of Words in the Output Text that Occurred in the Reference Text})}{(\text{Total Number of Words in the Output Text})}$$

Ej: Salida: Se va de paseo

Referencia: Se fue de paseo Precision = 3/4

Pero: Salida: Paseo paseo paseo Precisión = 3/3 = 1

- Se usa **clipped precision**: pone un **límite superior** en la cantidad de palabras de acuerdo a máx. cantidad de apariciones de la palabra en los textos de referencia.

BLEU (Bilingual Evaluation Understudy)

- Suele usarse para **Machine Translation** o A**utomatic Summarization**, pero puede usarse para cualquier tarea que involucre pares de texto entrada-destino.
- Mide la **calidad del texto predicho**, en comparación con un conjunto de referencias. Se mueve entre 0 y 1, cuanto más alto, mejor es.
- Analiza la **superposición de n-gramas** entre la salida y las traducciones de referencia con una penalización por salidas más cortas
- Normalmente se basa en un promedio entre la precisión de unigram, bigram, trigram and 4-gram precision,

Desventajas:

- No considera significado
- No considera estructura de las frases
- No trabaja bien con lenguajes ricos morfológicamente (como el alemán)

Referencias: Explicación de qué es y de sus fallas: [2] Evaluating Text Output in NLP: BLEU at your own risk | by Rachael Tatman | Towards Data Science

METEOR (Metric for Evaluation of Translation with Explicit Ordering)

Basada en BLEU

Incluye pasos adicionales, como considerar sinónimos y comparar las raíces de las palabras (para que "correr" y "corriendo" se cuenten como coincidencias).

A diferencia de BLEU, está diseñado explícitamente para comparar oraciones en lugar de corpora.

Referencias: [2,5]

Perplexity

Método de evaluación intrínseca de modelos de NLP.

Evalúa cuán bien un modelo de lenguaje predice una secuencia de texto.

- La **perplejidad mide la "sorpresa" del modelo** al ver el texto real.
- **Menor perplexity = mejor modelo**, porque significa que el modelo asignó alta probabilidad a las palabras verdaderas.
- Se interpreta como: "en promedio, el modelo está tan sorprendido como si tuviera que elegir entre *k* palabras en cada paso". U**na perplexity de 100** implica que el modelo se comporta como si eligiera al azar entre 100 opciones.

Referencias: [9]

MEWR (Machine Translation Evaluation without Reference Texts)

MEWR:

- no requiere traducciones de referencia
- Utiliza una combinación de embeddings de palabras y oraciones y perplexity

BERTScore

Aprovecha las representaciones de palabras contextualizadas, lo que le permite ir más allá de la coincidencia exacta y captura las paráfrasis mejor.

Referencia: [5]

MoverScore

Medida para evaluar la similitud entre un par de oraciones escritas en el mismo idioma.

Aparentemente correlaciona mejor con los juicios humanos que BLEU.

Referencias: [7]

SAS -semantic answer similarity -

Busca similitud semántica y no sintáctica.

Referencia: [9]

Evaluación de QA

Accuracy : medir la exactitud para responder correctamente. Para esto usar las preguntas y sus respuestas correspondientes y medir el porcentaje de preguntas que el sistema responde correctamente.

F1: mide la media armónica entre *precision* y *recall*. Para calcular estas tres métricas también hay que utilizar un conjunto de datos con preguntas y sus respuestas.

Evaluación humana: a diferencia de las métricas comentadas anteriormente, la evaluación humana sirve para saber cuán bien logra el sistema que las respuestas a las preguntas planteadas sean naturales y fáciles de entender. Sería útil calificar la calidad de las respuestas con una escala fija y con criterios de qué corresponde a cada valor de la escala.

Robustez: mide cuán bien funciona el sistema con preguntas que están fuera de su conjunto de entrenamiento. Para esto se puede evaluar con preguntas que el sistema no haya visto y medir cuán buenas son las respuestas.

Velocidad: puede ser importante el tiempo que se demora el sistema en responder preguntas. Se podría medir en alguna unidad de tiempo y compararlo con el tiempo que le lleva dar respuestas a otro sistema de respuesta a preguntas.

Evaluación de performance de chatbot

Algunos indicadores:
1. **Cantidad de interacciones**: Mide cuántas veces los usuarios interactuaron con el chatbot para obtener una respuesta
2. Fallback Rate (tasa de falta de respuesta): Mide la tasa con que los usuarios no reciben información del chatbot. El que sea alta podría deberse a la inhabilidad del chatbot de proveer respuestas satisfactorias a las consultas del usuario o a la no comprensión de las mismas.
3. **Bounce Rate (tasa de rebote)**: Mide la tasa con la que el usuario abandona al chatbot después de una sola interacción.
4. **Confusion Rate:** Mide cuán frecuentemente los usuarios se confunden por las respuestas del chatbot.
5. **Disminución de contacto por manera habitual:** Si se asume que ese contacto se hace mediante el chatbot. Por ej. Llaman menos por teléfono porque resuelven las dudas con el chatbot.

Evaluación de performance de chatbot

6. **Accuracy en respuestas**: Para esto hay que comparar las respuestas del chatbot con un gold standard
7. **Relevancia**: Mide cuán relevantes son las respuestas del chatbot.
8. **Tasa de completitud de la tarea:** mide el porcentaje de conversaciones que puede concluir el chatbot sin tener que acudir a un operador humano.
9. **Satisfacción del usuario:** se podría medir a través de surveys, ratings, etc. Se podría preguntar sobre si es amigable, si ayudó, etc.
10. **Tasa de errores:** medición de errores (respuestas irreleventas, respuestas incorrectas, falta de comprensión de pregunta del usuario)
11. **Flujo conversacional:** Mide cuán bien maneja contexto, recuerda interacciones pasadas, etc.
12. **Tasa de retención:** mide el porcentaje de usuarios que vuelven a usar el chatbot.

Referencias

[1] Evaluation Metrics in Natural Language Processing — BLEU | by Priyanka | Medium
[2] Evaluating Text Output in NLP: BLEU at your own risk | by Rachael Tatman | Towards Data Science
[3] Two minutes NLP — Quick intro to Question Answering | by Fabio Chiusano | NLPlanet | Medium
[4] What is Question Answering? - Hugging Face
[5] Evaluating Question Answering Evaluation (aclanthology.org)
[6] Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary | Transactions of the Association for Computational Linguistics | MIT Press
[7] GitHub - AIPHES/emnlp19-moverscore: MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance
[8] Perplexity in NLP: Definition, Pros, and Cons — Techslang
[9] Semantic Answer Similarity for Evaluating Question Answering Models - ACL Anthology
[10]Question Answering in Natural Language Processing [Part-I] | by Ranjan Satapathy | Lingvo Masino | Medium (open datasets for question answering)