Foundation Models

Jacopo Staiano
jacopo.staiano@unitn.it

Marco Bronzini
marco.bronzini-1@unitn.it

University of Trento

October 10, 2025

Overview

Friday, October 10

Lab 1

- Metrics for Natural Language Generation (NLG)
- Real-world use cases (QA with RAG)

Wednesday, October 15

Lab 2

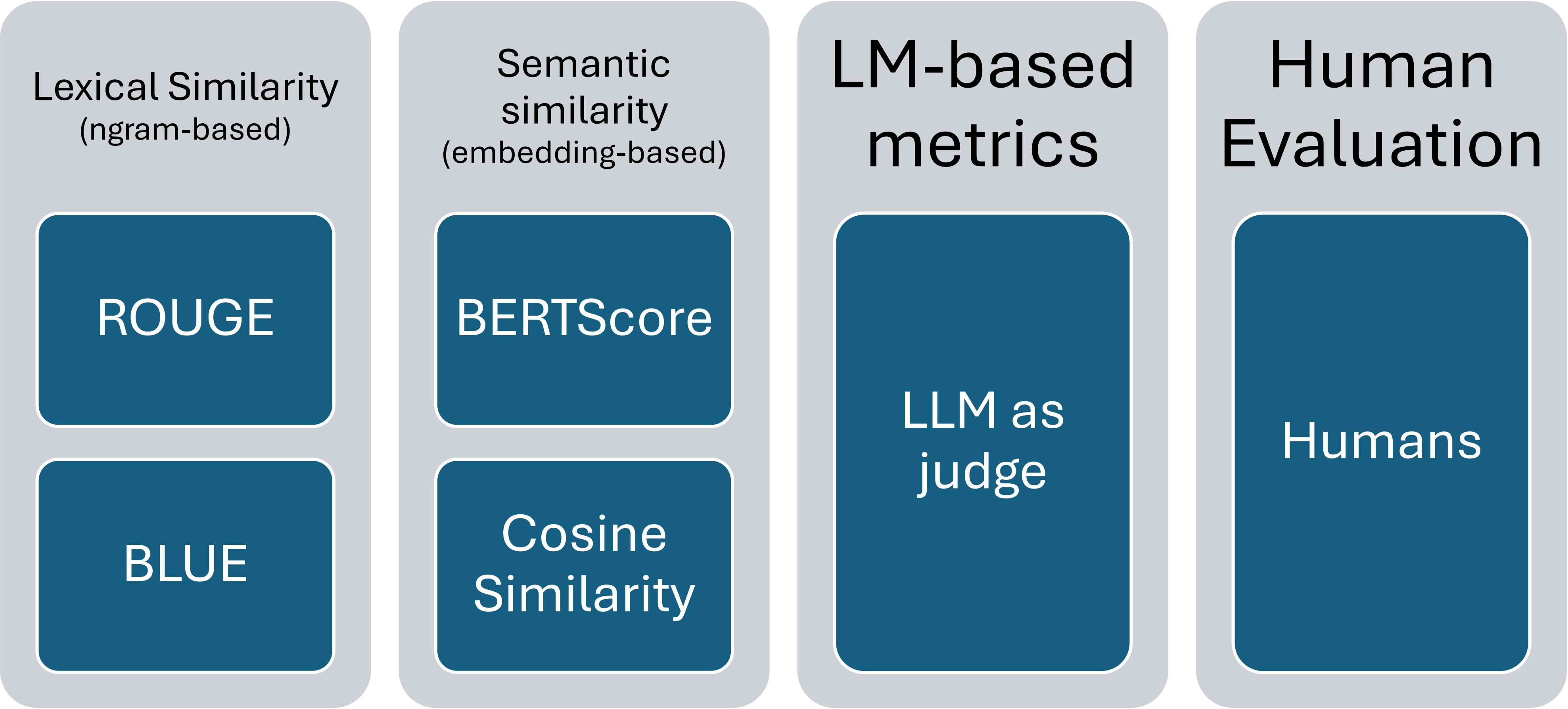
- Extract latent features from LLM embeddings by training a classification model

Lab 3

November

Lab 4

Metrics for Text Generation



Lexical Similarity
(ngram-based)

ROUGE

BLUE

Semantic similarity
(embedding-based)

BERTScore

Cosine Similarity

LM-based metrics

LLM as judge

Human Evaluation

Humans

[A.1] ROUGE

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

- is a *set of metrics* used to evaluate the *quality of automatically generated text* by comparing it with one or more human-written reference texts.

It measures the overlap of **n-grams, word sequences,** or **word pairs** between the output and the reference → *more overlap, the better the generated text is assumed to be*

ROUGE-N

- Measures ***n-gram overlap*** (e.g., ROUGE-1 for unigrams, ROUGE-2 for bigrams)

ROUGE-L

- Based on the ***longest common subsequence*** (LCS) between output and reference

ROUGE-S

- Measures ***skip-bigram overlap*** (word pairs in order but not necessary consecutive)

Each metric can be calculated in terms of:

Precision

Recall

F1 score

[A.2] BLEU

Bilingual Evaluation Understudy (BLEU) metric is widely used for evaluating the *quality of machine generated translations* by comparing them to human reference translations.

- The BLEU score is **based on precision of n-grams**
- *Brevity penalty (BP)* for penalize unnaturally short translations

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \ln(p_n)\right)$$

A: The sun set behind the hills
B: The moon set behind the hills

BLEU Score: 53.73

https://cloud.google.com/translate/docs/advanced/automl-evaluate

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

Practical Tutorial

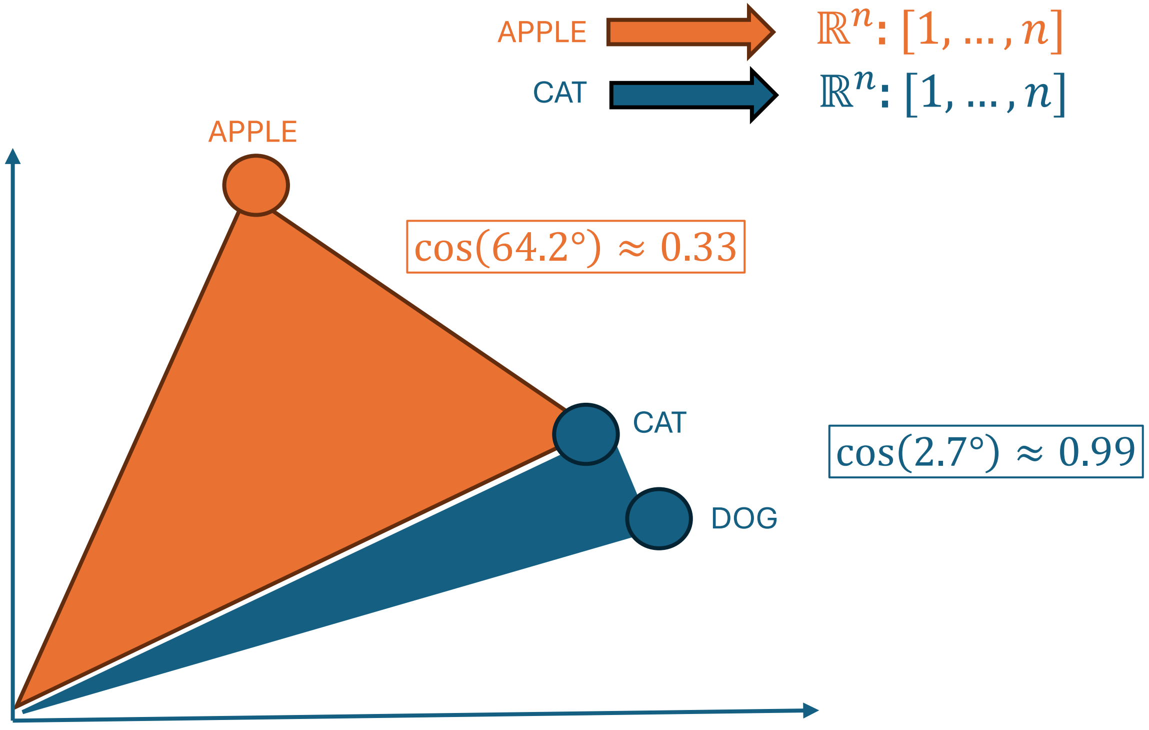
Please, open the following notebook:

https://colab.research.google.com/github/saturnMars/FM_2025/blob/main/Lab1_NLG_metrics.ipynb

[B.1] Cosine similarity

It is a mathematical concept that measures how similar two vectors are *based on the angle between them*.

$$\cos(\theta) = \frac{A \cdot B}{||A||\ ||B||}$$


APPLE
$\mathbb{R}^n$: $[1, \ldots, n]$
CAT
$\mathbb{R}^n$: $[1, \ldots, n]$
APPLE
$\cos(64.2°) \approx 0.33$
CAT
$\cos(2.7°) \approx 0.99$
DOG

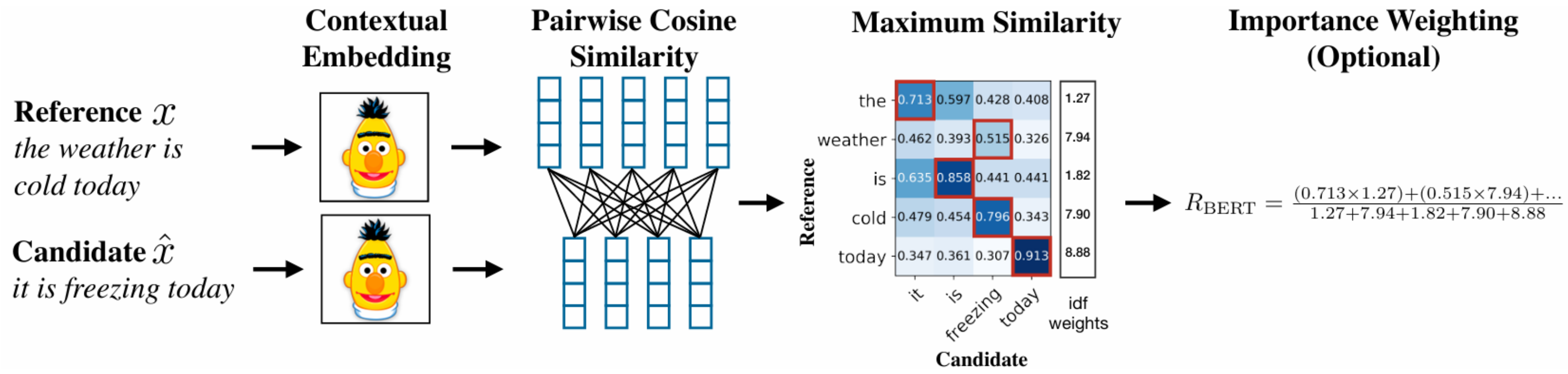
STEPS:

1. Texts (word, sentence, ...)
2. Vectorization (embeddings)
3. Compute cosine similarity

[B.2] BERTScore

BERTScore computes a **token-wise similarity score** based on *contextual embeddings*

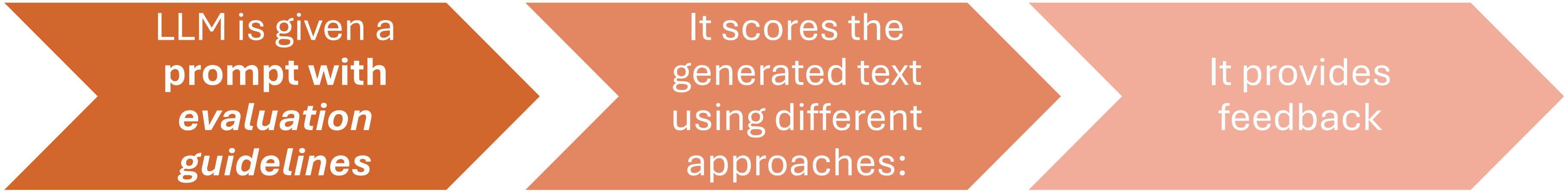
- each token in the candidate sentence with each token in the reference sentence.
- computes token similarity using BERT-based contextual embeddings. .


Contextual Embedding
Pairwise Cosine Similarity
Maximum Similarity
Importance Weighting (Optional)
Reference x
the weather is cold today
Candidate $\hat{x}$
it is freezing today
Reference
the 0.713 0.597 0.428 0.408 1.27
weather 0.462 0.393 0.515 0.326 7.94
is 0.635 0.858 0.441 0.441 1.82
cold 0.479 0.454 0.796 0.343 7.90
today 0.347 0.361 0.307 0.913 8.88
it
is
freezing
today
idf weights
Candidate
$R_{\mathrm{BERT}} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$

BERTScore: Evaluating Text Generation with BERT

LLM as judge

Evaluation strategy where a **large language model acts as an automated evaluator**, assessing the quality of AI-generated text based on predefined criteria.



LLM is given a **prompt with *evaluation guidelines***

- it is instructed to judge responses based on relevant aspects such as:
  - *relevance*
  - *coherence*,
  - *fluency*
  - *factual accuracy*

It scores the generated text using different approaches:

- **Pointwise**: Evaluates a *single response* independently.
- **Pairwise**: *Compares two responses* and determines which is better.
- **Listwise**: *Ranks multiple responses*

It provides feedback

- a numerical score
- an explanation of its reasoning.

Hugging Face tutorial: Using LLM-as-a-judge

Human Evaluation

IMPORTANCE

Provides **deeper insights** into abstract features such as coherence, relevance, and creativity.

Helps **compare results** more accurately, ensuring real-world applicability.

LIMITATIONS

Time-consuming and **costly**, requiring significant resources

Requires expertise: finding or training evaluators can be difficult

Subjectivity & **inconsistency**: different evaluators may disagree, impacting reliability

Limitation

The evaluation of Natural Language Generation (NLG) can be tricky as **each metrics provides only a partial understanding of the output quality**

Low scores do not always indicate poor quality

The sun set behind the hills

The moon set behind the hills

METRICS	SCORE	
ROUGE-1 [F1]	0.83	Lexical
ROUGE-2 [F1]	0.60	
BLEU Score	53.73	
BERTScore (F1)	0.92	Semantic
Cosine Similarity	0.71	

TASK: Machine Translation (ML)

Machine Translation (MT) is the process of automatically converting text or speech **from one language to another using computational models**.

Life is full of surprises [**ENGLISH**] → *La vita è piena di sorprese* [**ITALIAN**]

EVALUATION: Mainly using lexical-based metrics (token-based accuracy)

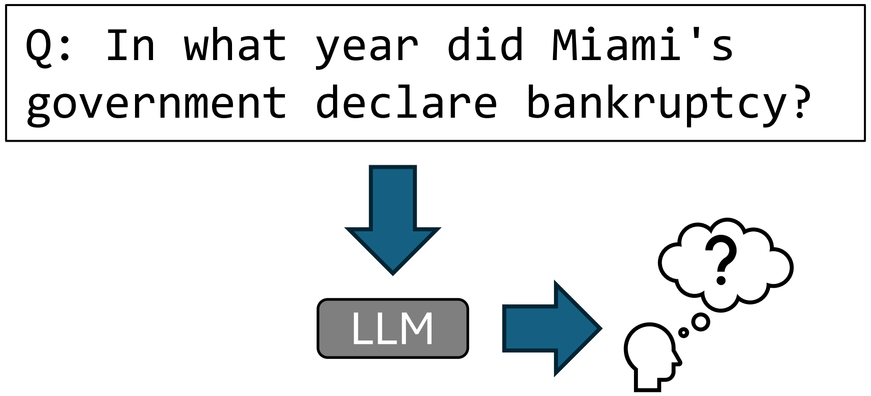
1. TARGET: «La vita è piena di sorprese»
2. OUTPUT: «La vita è piena di sorprendenti»

BLEU Score
64.3%

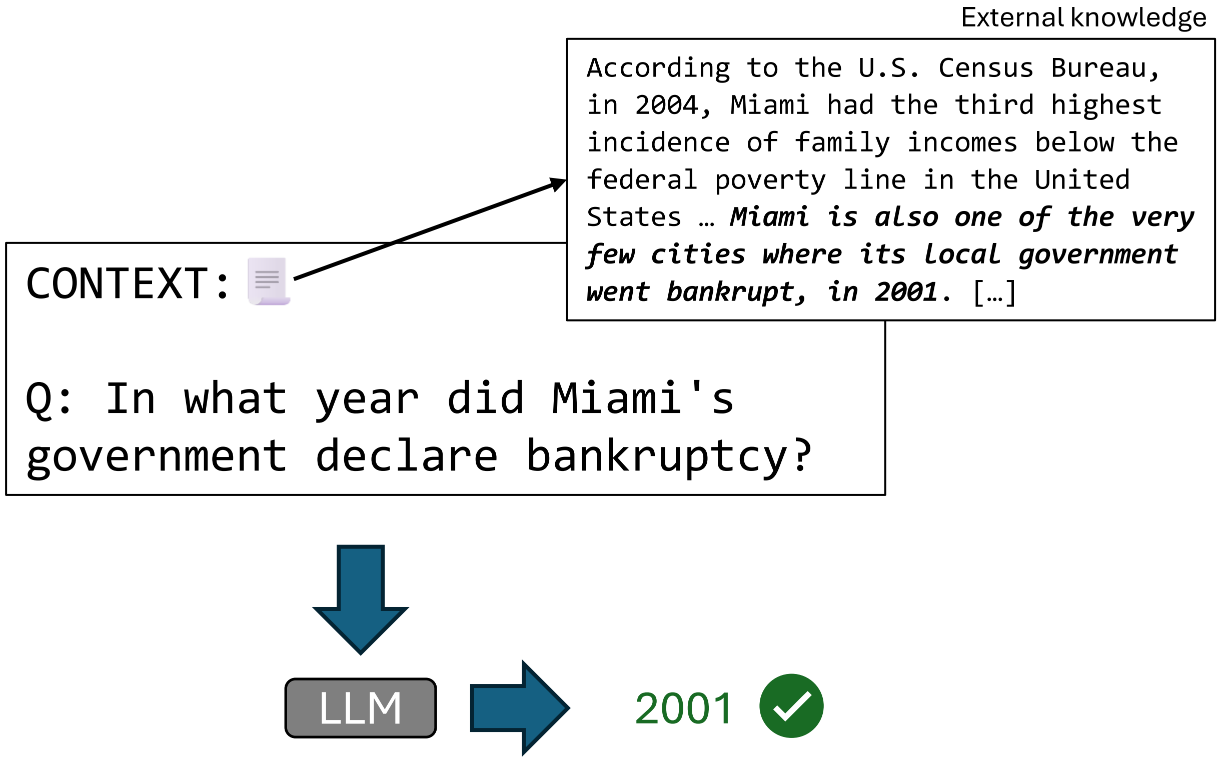
TASK: Retrieval-Augmented Generation (RAG)

It is an approach combines **information retrieval** with generative LLMs to generate more accurate, fact-based responses. It combines:

- the *internal knowledge of a language model*
- with **contextually relevant documents**.



Q: In what year did Miami's government declare bankruptcy?



External knowledge

According to the U.S. Census Bureau, in 2004, Miami had the third highest incidence of family incomes below the federal poverty line in the United States … ***Miami is also one of the very few cities where its local government went bankrupt, in 2001.*** […]

CONTEXT:

Q: In what year did Miami's government declare bankruptcy?

QUIZ

MENTI **(64286241)**

https://www.menti.com/aljoa3zgejuz