



EVALUATION

Luísa Coheur

OVERVIEW

- Learning Objectives
- Topics
 - Prelude: comparing strings
 - Automatic evaluation
 - Baseline
 - Some evaluation metrics
 - Human evaluation
 - Automatic vs. Human Evaluation
 - Intrinsic vs. extrinsic evaluation
 - Evaluation forums/shared tasks
- Key takeaways
- Suggested readings

LEARNING OBJECTIVES

LEARNING OBJECTIVES

- After this class, the student should be able to:
 - Explain some evaluation measures widely used in NLP
 - Define several concepts, such as evaluation fora, human evaluation, intrinsic/extrinsic evaluations, etc.
 - Apply some similarity/distance/evaluation metrics

TOPICS

OVERVIEW

- Learning Objectives
- Topics
 - Prelude: comparing strings
 - Automatic evaluation
 - Baseline
 - Some evaluation metrics
 - Human evaluation
 - Automatic vs. Human Evaluation
 - Intrinsic vs. extrinsic evaluation
 - Evaluation forums/shared tasks
- Key takeaways
- Suggested readings

COMPARING STRINGS

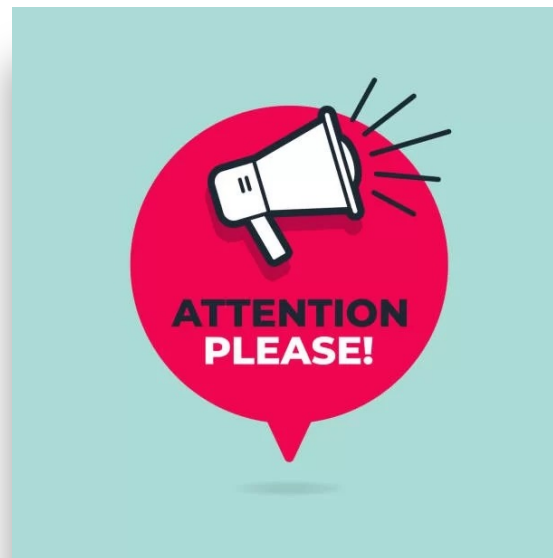
- Comparing strings is at the basis of an NLP evaluation
 - We compare words, sentences, paragraphs and even documents
 - These comparisons might be lexical or semantic
 - Example:
 - "The discovery of DNA structure was a revolutionary achievement in science."
 - vs.
 - "The revelation of the DNA structure was a revolutionary milestone in science."

COMPARING STRINGS

- Comparing strings can be useful in other tasks:
 - Spell Checking:
 - Example: "The brrown fox jumpped over the lazy dog."
 - Data Cleaning: Matching and merging duplicate records
 - Ex: "John Fitzgerald Kennedy", "Jonh F. Kennedy", "Jonh Kennedy " => "John Kennedy"
 - Plagiarism Detection:
 - Ex: "The discovery of DNA structure was a revolutionary achievement in science." vs "The revelation of the DNA structure was a revolutionary milestone in science."
 - Code Similarity Detection
 - ... (and many, many more)

BEFORE WE MOVE ON: DISTANCE VS. SIMILARITY

- Pay attention to the metric you use: some are similarity metrics, some are distance/difference metrics
- You may need to normalize them



COMPARING STRINGS

- Edit-based metrics allow to quantify how dissimilar two strings are to one another, by counting the minimum number of operations required to transform one string into the other
 - The Levenshtein distance is an Edit-based metric that calculates the minimum number of insertions, deletions or substitutions required to change one sequence into the other

```
 $C_1, C_2, C_3 \leftarrow 1$   
if  $n = 0$  then  
    return m  
end if  
if  $m = 0$  then  
    return n  
else  
    Build matrix M, with  $m+1$  lines and  $n+1$  columns  
     $j \leftarrow 1$   
    while  $j \neq n + 1$  do  
         $i \leftarrow 1$   
        while  $i \neq m + 1$  do  
            if  $s[i] = t[j]$  then  
                 $M[i, j] = M[i - 1, j - 1]$  // Take the diagonal value if characters  
                match  
            else  
                 $M[i, j] = \min(M[i - 1, j] + C_1, M[i, j - 1] + C_2, M[i - 1, j - 1] + C_3)$   
            end if  
             $i \leftarrow i + 1$   
        end while  
         $j \leftarrow j + 1$   
    end while  
    return  $M[m, n]$   
end if
```

| | | | | |
|---|----------|---|----------|----------|
| | | 0 | 1 | 2 |
| | • | | M | E |
| 0 | • | 0 | 1 | 2 |
| 1 | M | 1 | 0 | 1 |
| 2 | Y | 2 | 1 | 1 |

- An empty string



DELETION OPERATION



INSERTION OPERATION



DO NOTHING (both letters are equal)



SUBSTITUTION OPERATION

ACTIVE LEARNING MOMENT



EXERCISE

- What is the MED between MORDOR and LORD?

| | | M | O | R | D | O | R |
|------------------|---|---|---|---|---|---|---|
| L O R D | 0 | | | | | | |
| | 1 | | | | | | |
| | 2 | | | | | | |
| | 3 | | | | | | |
| | 4 | | | | | | |

EXERCISE

- What is the MED between MORDOR and LORD?

| | | M | O | R | D | O | R |
|------------------|---|---|---|---|---|---|---|
| L O R D | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| | 1 | 1 | 2 | | | | |
| | 2 | 2 | 1 | | | | |
| | 3 | 3 | | | | | |
| | 4 | 4 | | | | | |

EXERCISE

- What is the MED between MORDOR and LORD?

| | | M | O | R | D | O | R |
|------------------|---|---|---|---|---|---|---|
| L O R D | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| | 1 | 1 | 2 | 3 | 4 | 5 | 6 |
| | 2 | 2 | 1 | 2 | 3 | 4 | 5 |
| | 3 | 3 | 2 | 1 | 2 | 3 | 4 |
| | 4 | 4 | 3 | 2 | 1 | 2 | 3 |

MORE EDIT DISTANCES

- Other Edit-based metrics:
 - The **Longest Common Subsequence (LCS) distance** allows as operations only insertion and deletion, not substitution
 - The **Hamming distance** allows only substitution (it only applies to strings of the same length)
 - The **Damerau–Levenshtein distance** allows insertion, deletion, substitution, and the transposition (swapping) of two adjacent characters
 - The **Jaro distance** allows only transposition
- Just to name a few...

COMPARING STRINGS

- Jaccard and Dice are other metrics used to compare strings (but they operate on sets)
 - While the MED says how distant two strings are (the highest the value, the less similar they are), Jaccard and Dice are similarity metrics: the highest the value, the more similar they are
- Let us see Jaccard and Dice as sets and not bags
 - that is, no repetitions; a set does not have repeated elements
- Both are examples of token-based similarity-metrics
- There are many, many more token-based metrics

COMPARING STRINGS

- Jaccard(s, t) = $|s \cap t| / |s \cup t|$
- Dice(s, t) = $2 \times |s \cap t| / (|s| + |t|)$
- ...
- Overlap(s, t) = $|s \cap t| / \min(|s|, |t|)$ ← just another example
- ...

Remember: consider s and t as sets!
(each token appears only once)

COMPARING STRINGS

- Notice that you can pre-process strings and compare them afterwards
- For instance, you can “translate” your string to a form that represents how it sounds

COMPARING STRINGS (NOW AT THE SOUND LEVEL)

- Soundex (but also others, such as Metaphone, ...)

1. Retain the first letter of the name;
2. Drop all occurrences of a, e, i, o, u, y, h, w (unless they appear in the first position).

3. Replace consonants by digits, as follows (after the first letter):

(a) b, f, p, v \rightarrow 1

(b) c, g, j, k, q, s, x, z \rightarrow 2

(c) d, t \rightarrow 3

(d) l \rightarrow 4

(e) m, n \rightarrow 5

(f) r \rightarrow 6

Luisa

L

Ls

L2

L200

4. Two adjacent letters with the same number are coded as a single number (ex: 55 \rightarrow 5)
5. Continue until you have one letter and three numbers. If you run out of numbers, add zeros until there are three numbers (ex: L2 \rightarrow L200); if you have too much numbers drop them after the third one (ex: L2345 \rightarrow L234).

ACTIVE LEARNING MOMENT



SOUNDEX – EXAMPLE

- Can soundex help you to understand the following joke?
 - How does superman likes his milk?
 - Claro quente (vs Clark Kent)

1. Retain the first letter of the name;
2. Drop all occurrences of a, e, i, o, u, y, h, w (unless they appear in the first position).
3. Replace consonants by digits, as follows (after the first letter):

(a) b, f, p, v \rightarrow 1

(b) c, g, j, k, q, s, x, z \rightarrow 2

(c) d, t \rightarrow 3

(d) l \rightarrow 4

(e) m, n \rightarrow 5

(f) r \rightarrow 6

C460 Q530 vs C462 K530

+ Levenshtein (for instance)

And if it was Claroquente vs. Clarkkent?

4. Two adjacent letters with the same number are coded as a single number (ex: 55 \rightarrow 5)
5. Continue until you have one letter and three numbers. If you run out of numbers, add zeros until there are three numbers (ex: L2 \rightarrow L200); if you have too much numbers drop them after the third one (ex: L2345 \rightarrow L234).

COMPARING STRINGS

- We have seen how to apply these metrics to words
 - $\text{MED}(\text{Monserate}, \text{Moncerrate}) = \dots$
 - $\text{Jaccard}(\text{Olá}, \text{Ola}) = \dots$
 - ...
- But we can also apply them to sentences (although they are not very effective):
 - $\text{MED}(\text{"Niagara Falls is viewed by thousands of tourists every year."}, \text{"Each year, thousands of people visit Niagara Falls."})$
 - $\text{Jaccard}(\text{"She was a successful author and speaker."}, \text{"She found success as a public speaker and writer."})$
 - ...

Can be used to detect paraphrases!
(at this moment, still at the lexical level –
no semantics)

COMPARING SENTENCES DRAWBACKS

(example from Read more: <http://www.city-data.com/forum/writing/1115620-two-sentences-have-same-words-but-2.html>)

- Only he told his mistress that he loved her. (Nobody else did)
- He only told his mistress that he loved her. (He didn't show her)
- He told only his mistress that he loved her. (Kept it a secret from everyone else)
- He told his only mistress that he loved her. (Stresses that he had only ONE!)
- He told his mistress only that he loved her. (Didn't tell her anything else)
- He told his mistress that only he loved her. ("I'm all you got, sweetie--nobody else wants you.")
- He told his mistress that he only loved her. (Not that he wanted to marry her.)
- He told his mistress that he loved only her. (Yeah, don't they all...).
- He told his mistress that he loved her only. (Similar to above one).

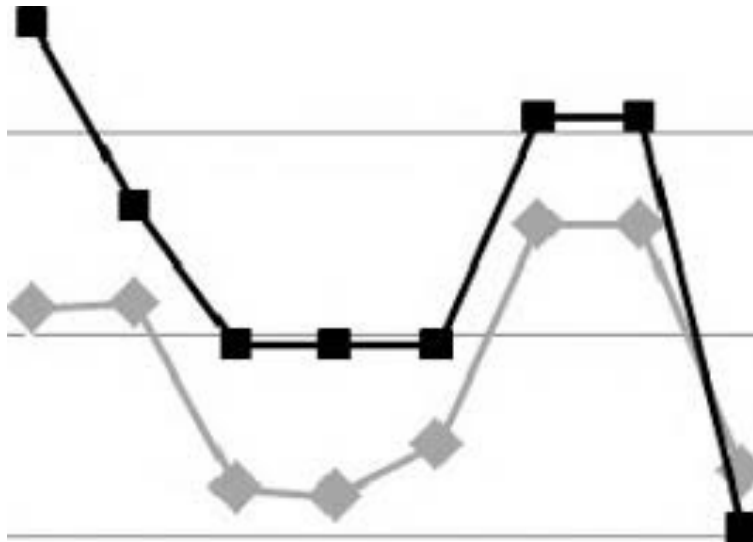
MED is different, but considering the other measures we talked about, they all end in the same sets!

OVERVIEW

- Learning Objectives
- Topics
 - Prelude: comparing strings
 - Automatic evaluation
 - Baseline
 - Some evaluation metrics
 - Human evaluation
 - Automatic vs. Human Evaluation
 - Intrinsic vs. extrinsic evaluation
 - Evaluation forums/shared tasks
- Key takeaways
- Suggested readings

AUTOMATIC EVALUATION

- Baseline:
 - a known starting configuration(s) against which results are compared



You should always compare your system with others (at least with a (random) baseline)!!!

Hopefully:
Baseline (grey) vs Your System

AUTOMATIC EVALUATION

- Metrics:
 - Precision (and macro- and micro-precision)
 - Recall (and macro- and micro-recall)
 - Accuracy
 - F-measure and F1-measure – use Precision and Recall

AUTOMATIC EVALUATION

- Consider:
 - True Positives (TP)
 - True Negatives (TN)
 - False Positives (FP)
 - False Negatives (FN)
- Precision (P) = $TP / (TP + FP)$
- Recall (R) = $TP / (TP + FN)$
- F1 = $2PR / (P + R)$

ACTIVE LEARNING MOMENT



The capital of Portugal, Lisbon (Portuguese: Lisboa) has experienced a renaissance in recent years, with a contemporary culture that is alive and thriving and making its mark in today's Europe. Lisbon lacks a defined “~~downtown~~”, but the the vast Praça do Comércio, facing the river at the base of the pedestrianized grid of Baixa (lower town), occupies a central position. Further northwest from Baixa stretches ~~Lisbon's “Main Street”~~, Avenida da Liberdade, a broad boulevard resplendent in leafy trees, chic hotels and up-scale shops, terminating at the circular Praça de Marques de Pombal. To the east are old neighborhoods of Mouraria and Alfama, both relatively spared during the Great Earthquake (as they are on a firmer rock) and therefore both retaining the charm of the winding alleys and azulejo-covered crumbling walls (further north lie relatively boring residential quarters).

Named Entity Recognition is an NLP TASK!

$$\text{Precision (P)} = \text{TP}/(\text{TP}+\text{FP})$$

$$\text{Recall (R)} = \text{TP}/(\text{TP} + \text{FN})$$

$$F_1 = 2\text{PR}/(\text{P}+\text{R})$$

- Reference:

Portugal, Lisbon, Lisboa, Europe, Praça do Comércio, Baixa, Avenida da Liberdade, Praça de Marquês de Pombal, Mouraria, Alfama

- System A: Portugal, Lisboa, Avenida da Liberdade, Alfama

- System B: Portugal, Lisbon, Portuguese, Lisboa, Praça do Comércio, Europe, Main Street, Alfama, Great Earthquake

Which is the best system?

- System A

- $TP = 4, FP = 0, FN = 6$
- $P = 4 / (4+0) = 1$
- $R = 4 / (4+6) = 0.4$
- $F1 = 2PR / (P+R) = 2*0.4/1.4 = 0.57$

- System B

- $TP = 6, FP = 3$ (Portuguese, Main Street, Great Earthquake), $FN = 4$
- $P = 6 / (6+3) = 0.667$
- $R = 6 / (6+4) = 0.6$
- $F1 = 0.63$



AUTOMATIC EVALUATION

- More metrics:
 - BLEU – Machine Translation
 - METEOR – Machine Translation
 - ROUGE – initially summarization
 - Perplexity – to evaluate Language Models
 - ...
 - COMET – trained measure (deep learning model)

Recently: LLMs are used to evaluate systems

- Most of these metrics are now used in generative tasks such as dialogue systems
 - And they are awful!

ACTIVE LEARNING MOMENT



EXERCISE

- Consider the evaluation of a
 - Sentiment analysis task
 - Humanity is great – positive/negative
 - Translation task
 - Humanity is great
 - A humanidade é maravilhosa
 - A humanidade é porreira
 - A humanidade é fantástica
 - Dialogue task
 - [You] Hi!
 - [Bot] Hi, how are you?
 - [You] Fine, and you?
 - [Bot] Miserable.

Remember the previous class?

- How should we build gold collections for these tasks?

Discuss the difficulty of evaluating a model that performs each task

AUTOMATIC EVALUATION

- To conclude, automatic evaluation:
 - Allows for more agile development cycles in NLP
 - Evaluate a vast number of language samples quickly and consistently, saving time and resources;
 - thus, reduces the need for expensive and time-consuming manual evaluation
 - Experiments are easily reproduced
 - Provide a uniform standard, reducing the subjectivity and potential bias that might come with (more) human evaluations;
 - Although the comparison is almost always against some human reference

AUTOMATIC EVALUATION

- But...
 - Some metrics are just lexical
 - Most of the metrics do not provide insights into subjective aspects like fluency, coherency, and readability
 - Sometimes do not correlate well with human evaluations

OVERVIEW

- Learning Objectives
- Topics
 - Prelude: comparing strings
 - Automatic evaluation
 - Baseline
 - Some evaluation metrics
 - Human evaluation
 - Automatic vs. Human Evaluation
 - Intrinsic vs. extrinsic evaluation
 - Evaluation forums/shared tasks
- Key takeaways
- Suggested readings

HUMAN EVALUATION

- Allows for:
 - **Nuanced understanding**: Humans can grasp context, irony, humor, and cultural references in ways that automated metrics cannot
 - **Quality assessment**: human judges can assess subjective qualities like readability, coherency, and engagement of the text
 - **Error identification**: while automatic metrics can indicate that an error has occurred, human evaluators can provide detailed insights into the nature of the error
 - **Ground truth benchmarking**: human judgment often serves as the gold standard

HUMAN EVALUATION

- But...
 - It is necessary to ensure evaluator expertise and consistency
 - Human evaluation is expensive
 - Human evaluation is time-consuming (and can be boring)



HUMAN EVALUATION

- Some types of human evaluation:
 - Direct Assessment:
 - Human evaluators rate the output on a fine-grained scale
 - Example:
 - 1-10, 1-100
 - Rank-Based Evaluation:
 - Multiple outputs are ranked in order of preference or quality
 - Enables evaluators to assess relative performance
 - Paired Comparison (a subcase of rank-based evaluation):
 - Human evaluators are presented with two outputs and asked to choose the better one

ACTIVE LEARNING MOMENT



EXERCISE

- Consider the following synopses of Harry Potter books:
 - S1: Harry go to magic school and do magic stuff, then he fight bad guy and win at the end.
 - S2: Harry goes to school, learns magic, and beats the villain.
 - S3: Harry embarks on a thrilling journey at Hogwarts, uncovering magical secrets and battling the dark forces of Voldemort in an epic struggle for the fate of the wizarding world.
- (1) Consider that you want to evaluate the **English quality** of each synopsis. Use a **Direct Assessment** (0-10).
- (2) Consider that you want to evaluate how engaging is each synopsis. Use a Rank-Based Evaluation.

BY THE WAY...



Connor McCartan

@MILKCARTAN_

 **Follow**

[#ExplainAFilmPlotBadly](#); Noseless guy has an unhealthy obsession with a teenage boy



From: 64 Times People Explained Movies So Badly It Was Good

OVERVIEW

- Learning Objectives
- Topics
 - Prelude: comparing strings
 - Automatic evaluation
 - Baseline
 - Some evaluation metrics
 - Human evaluation
 - Automatic vs. Human Evaluation
 - Intrinsic vs. extrinsic evaluation
 - Evaluation forums/shared tasks
- Key takeaways
- Suggested readings

AUTOMATIC vs. HUMAN EVALUATION

- The best of two worlds:
 - COMET (Rei, 2019): Automatic metric used in Machine Translation that takes advantage of human evaluation (COMET is trained on human annotations – COMET is a trained metric – deep learning)
 - We will talk about this in one of the next classes

AUTOMATIC vs. HUMAN EVALUATION

- What is the best metric?
 - We also need to evaluate evaluation metrics
 - To do so, we check [how they relate](#) with human evaluation

AUTOMATIC vs. HUMAN EVALUATION

- Pearson, Spearman, and Kendall (remember?) are three statistical methods used to measure the strength and direction of association between two variables
 - They are used to find the correlation between human scores and metric scores. The higher the correlation with human scores, the better the metric is

AUTOMATIC vs. HUMAN EVALUATION

1. Pearson correlation coefficient (r):

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where x_i and y_i are the values of the two variables, \bar{x} and \bar{y} are the means of those variables, respectively.

2. Spearman's rank correlation coefficient (ρ):

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of corresponding variables and n is the number of observations.

3. Kendall's tau (τ):

$$\tau = \frac{2}{n(n-1)} \sum_{i < j} \text{sign}(x_i - x_j) \text{sign}(y_i - y_j)$$

where n is the number of observations, x_i and x_j are the ranks of the x variables, y_i and y_j are the ranks of the y variables, and sign is the sign function.

An example of calculating Spearman's correlation

To calculate a Spearman rank-order correlation on data without any ties we will use the following data:

| | Marks | | | | | | | | | |
|---------|-------|----|----|----|----|----|----|----|----|----|
| English | 56 | 75 | 45 | 71 | 62 | 64 | 58 | 80 | 76 | 61 |
| Maths | 66 | 70 | 40 | 60 | 65 | 56 | 59 | 77 | 67 | 63 |

We then complete the following table:

| English (mark) | Maths (mark) | Rank (English) | Rank (maths) | d | d ² |
|----------------|--------------|----------------|--------------|---|----------------|
| 56 | 66 | 9 | 4 | 5 | 25 |
| 75 | 70 | 3 | 2 | 1 | 1 |
| 45 | 40 | 10 | 10 | 0 | 0 |
| 71 | 60 | 4 | 7 | 3 | 9 |
| 62 | 65 | 6 | 5 | 1 | 1 |
| 64 | 56 | 5 | 9 | 4 | 16 |
| 58 | 59 | 8 | 8 | 0 | 0 |
| 80 | 77 | 1 | 1 | 0 | 0 |
| 76 | 67 | 2 | 3 | 1 | 1 |
| 61 | 63 | 7 | 6 | 1 | 1 |

EVALUATION OF EVALUATION METRICS

- We can also evaluate two metrics considering how they rank two systems for each sample:
 - considering the scores given by two metrics to two systems, check if these metrics share the same sign that humans do

$$\frac{|\text{sign}(\text{metric}\Delta) = \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|}$$

| Sample | Human | | Sign | M1 | | Sign | M2 | | Sign |
|--------|-------|------|------|------|------|------|------|------|------|
| | Sys1 | Sys2 | | Sys1 | Sys2 | | Sys1 | Sys2 | |
| | 7 | 6.5 | + | 5 | 4 | + | 7 | 7.3 | - |

M1 will gain 1 point because it ranked the same way as Humans this sample; M2 will not

ACTIVE LEARNING MOMENT



EXERCISE

- Consider that system X returns 10 translations that are scored by humans as SysXHuman and by metric Y as SysXMetricY. The values are:

- Sys1Human = [6, 8, 9, 7, 9, 9, 7, 7, 6, 6] = S1 S1 S1 = S2 S2 S2 S2 =
- Sys2Human = [6, 7, 5, 5, 9, 10, 8, 10, 7, 6]
- Sys1Metric1 = [5, 7, 8, 6, 8, 9, 7, 7, 6, 6]
- Sys2Metric1 = [6, 8, 5, 7, 7, 6, 7, 7, 7, 9] S2 S2 S1 S2 S1 S1 = = S2 S2
- Sys1Metric2 = [6, 8, 9, 7, 9, 10, 8, 8, 7, 8] S2 S2 S1 S1 = S1 S2 S2 S1 S2
- Sys2Metric2 = [7, 9, 4, 6, 9, 9, 9, 10, 4, 9]

- Which metric better correlates with the human evaluation according with this metric?


$$\frac{|\text{sign}(\text{metric}\Delta) = \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|}$$

Metric1 Accuracy: 0,2

Metric2 Accuracy: 0,5

EXERCISE

- In detail:
 - Comparing M1 and M2 with H



| | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 | T9 | T10 |
|----|----|----|----|----|----|----|----|----|----|-----|
| H | = | S1 | S1 | S1 | = | S2 | S2 | S2 | S2 | = |
| M1 | S2 | S2 | S1 | S2 | S1 | S1 | = | = | S2 | S2 |
| M2 | S2 | S2 | S1 | S1 | = | S1 | S2 | S2 | S1 | S2 |

M1 Accuracy: 2 in 10 = 0,2

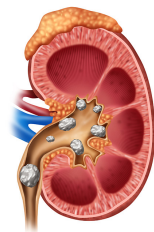
M2 Accuracy: 5 in 10 = 0,5

OVERVIEW

- Learning Objectives
- Topics
 - Prelude: comparing strings
 - Automatic evaluation
 - Baseline
 - Some evaluation metrics
 - Human evaluation
 - Automatic vs. Human Evaluation
 - Intrinsic vs. extrinsic evaluation
 - Evaluation forums/shared tasks
- Key takeaways
- Suggested readings

EXTRINSIC VS. INTRINSIC EVALUATION

- Intrinsic: evaluate your system alone
- Extrinsic: evaluate your system as a component of a more complex system
- Example:
 - Question Classification (QC) vs. Question/Answer (QA)
 - A QC system can be evaluated “per se” and have a score X (intrinsic evaluation)
 - The same system might replace an existing QC system in a QA system, and improve (or not) the QA system (extrinsic evaluation of the QC system)



Kidney Stones



OVERVIEW

- Learning Objectives
- Topics
 - Prelude: comparing strings
 - Automatic evaluation
 - Baseline
 - Some evaluation metrics
 - Human evaluation
 - Automatic vs. Human Evaluation
 - Intrinsic vs. extrinsic evaluation
 - [Evaluation forums/shared tasks](#)
- Key takeaways
- Suggested readings

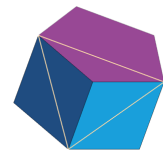
EVALUATION FORA (PL. OF FORUM)

- CLEF (QA, ...)
- IWSLT (Translation)
- SEM-EVAL (Semantics)
- SENSEVAL (Semantics)
- ...
- There are even for a (or shared tasks) to evaluate evaluation metrics (translation)



EVALUATION FORA (PL. OF FORUM)

- As we have seen:
 - Recent evaluation campaigns:
 - [...] new benchmark [...] requiring a single system to perform ten disparate natural language tasks [...]



decaNLP

EVALUATION FORA (PL. OF FORUM)

- As we have seen:
 - Recent evaluation campaigns:
 - \$500,000 prize will be awarded to the team that creates the best socialbot. The second- and third-place [...] \$100,000 and \$50,000, respectively.



KEY TAKEAWAYS

KEY TAKEAWAYS

- There are many different metrics to apply to different NLP tasks. Some tasks are properly evaluated; some do not
- Both human and automatic evaluations have they pros and cons
- Concepts:
 - Baseline, similarity measures, distance measures, evaluation measures, evaluation fora, automatic evaluation, human evaluation, intrinsic/extrinsic evaluations, evaluation fora, etc.

SUGGESTED READINGS

READINGS

- Sebenta:
 - Methodology, corpora and evaluation
 - notice that it does not covers all these slides
 - <https://aclanthology.org/2021.wmt-1.57.pdf>