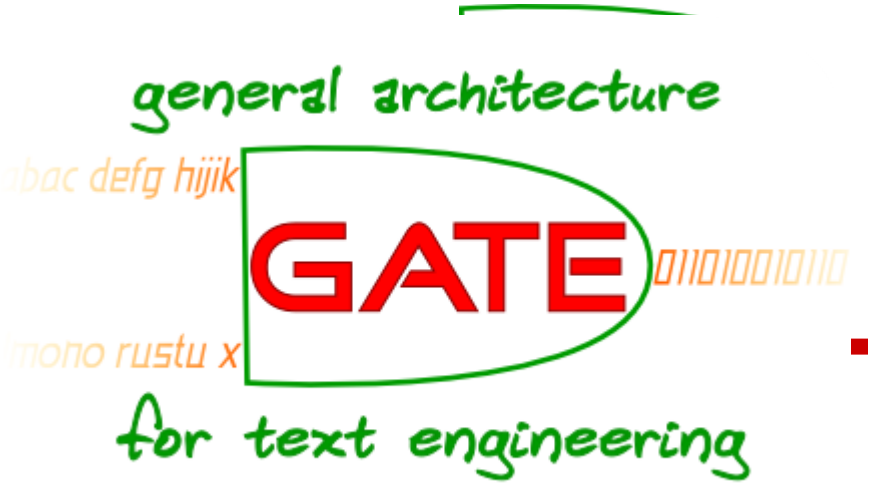


Performance measures (NLP & Clinical)

GATE Team
University of Sheffield

Instances

- Data to be compared:
  - “Key”, “manual”, “gold standard”: set of annotations (& relevant features), made or corrected by human annotators, taken as correct
  - “Response”, “output”, “automatic”: set of annotations (& relevant features) produced by the application

Scoring

- TP (true positive), correct annotation
- FP (false positive), “spurious” (in IE/IR terms): annotation is present in the Response set but not the Key
- FN (false negative), “missing” (IE/IR): annotation is present in the Key but missing in the Response
- TN (true negative): annotation is absent in both sets — not used in IE/IR measures

Strict vs lenient

- Strict: only correct annotations covering exactly the correct span of text count as TPs
- Lenient: “partially correct” or “overlap” annotations *also* count as correct: correct annotation covering part of the correct span of text (shorter, longer, overlapping at either end)
- Precision & recall can use either of these approaches

Why strict vs lenient?

- Diagnosis example
  - "Patient has a diagnosis of schizophrenia with Capgras delusion and is being treated with…"
  - Application output is underlined above; key is in red
  - Key annotation covers the full diagnosis, but the application only annotated "schizophrenia" (3 words missing)
  - Scoring needs to be strict; the output above is wrong

Why strict vs lenient?

- Social care example:
  - "Patient will be transferred to a care home on discharge next week."
  - The human annotated the whole sentence; the application annotated only the underlined part; but the annotation & features are correct:
    - CareHome
    - status = future
    - subject = patient
  - Lenient scoring is appropriate

Sensitivity = recall

- Sensitivity, detection rate, true positive rate; proportion of those with the condition, who have a positive test result
- High sensitivity: the test catches as many people with the condition as possible
- “If a person has the disease, what is the probability that the test will be positive?”
- sens = TPs / (TPs + FNs)
- Same as recall (IE/IR): “What % of the annotations in the key set did my application find?”

Specificity

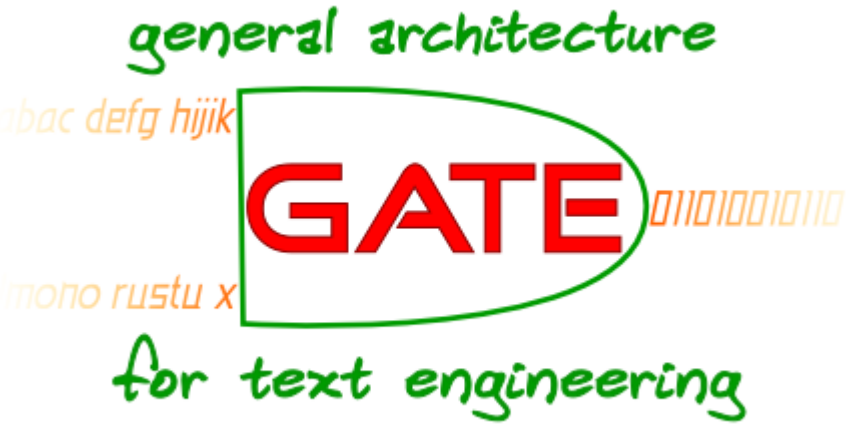
- Specificity; true negative rate; proportion of those without the condition, who have a negative test result
- High specificity: the test has as few false positives as possible
- “If a person does not have the disease, what is the probability that the test will be negative?”
- spec = TNs / (TNs + FPs)
- not used in IE/IR

PV+ = precision

- Positive predictive value (PV+)
- “If the test result is positive, what is the probability that the patient actually has the disease?”
- PV+ = TPs / (TPs + FPs)
- Same as precision (IE/IR term): “What % of the annotations in my application's output are correct?”

PV-

- Negative predictive value (PV-)
- “If the test result is negative, what is the probability that the patient does not have the disease?”
- PV- = TNs / (TNs + FNs)
- not used in IE/IR



Thank you.
Any questions?