# Pragmatics and Evaluation

## Einführung in die Informationsverarbeitung

Nils Reiter

December 15, 2022

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

# Section 1

## Language and Linguistics

# Subsection 1

## Pragmatics

# Pragmatics

- Pragmatics: Language and the rest of the world
  - ›pragmatic wastebasket‹ Bar-Hillel (1971)
  - What semantics can’t explain belongs to pragmatics ☺

# Pragmatics

- Pragmatics: Language and the rest of the world
  - ›pragmatic wastebasket‹ Bar-Hillel (1971)
  - What semantics can't explain belongs to pragmatics ☺
- Pragmatic phenomena Levinson (1983)
  - Deixis

# Pragmatics

- Pragmatics: Language and the rest of the world
  - ›pragmatic wastebasket‹ Bar-Hillel (1971)
  - What semantics can't explain belongs to pragmatics ☺
- Pragmatic phenomena Levinson (1983)
  - Deixis: Person: I/time: now/place: here
  - Conversational implicature
    - Grice: The co-operative principle Grice (1975)

# Pragmatics

- Pragmatics: Language and the rest of the world
  - ›pragmatic wastebasket‹ Bar-Hillel (1971)
  - What semantics can't explain belongs to pragmatics ☺
- Pragmatic phenomena Levinson (1983)
  - Deixis: Person: I/time: now/place: here
  - Conversational implicature
    - Grice: The co-operative principle Grice (1975)
    - E.g., the maxim of Quantity
      (i) make your contribution as informative as is required for the current purposes of the exchange
      (ii) do not make your contribution more informative than is required

# Pragmatics

- Pragmatics: Language and the rest of the world
  - ›pragmatic wastebasket‹ Bar-Hillel (1971)
  - What semantics can't explain belongs to pragmatics ☺
- Pragmatic phenomena Levinson (1983)
  - Deixis: Person: I/time: now/place: here
  - Conversational implicature
    - Grice: The co-operative principle Grice (1975)
    - E.g., the maxim of Quantity
      (i) make your contribution as informative as is required for the current purposes of the exchange
      (ii) do not make your contribution more informative than is required
  - Presupposition
  - Speech acts
    - ›I hereby christen this ship the H.M.S. Flounder.‹ Austin (1962)
    - Change of the state of the world
  - Conversational structure

# Presupposition

## Implicit assumptions about the world

### Example

(1) John managed to stop in time.

(2) John stopped in time.

(3) John tried to stop in time.

# Presupposition

## Implicit assumptions about the world

### Example

(1) John managed to stop in time.
(2) John stopped in time.
(3) John tried to stop in time.

From (1), we can infer (2) and (3).

### Example

(4) John didn't manage to stop in time.

From (4), we cannot infer (2), but (3).

# Presupposition

- Entailments are cancelled under negation
- Presuppositions remain stable

# Presupposition

- Entailments are cancelled under negation
- Presuppositions remain stable
- Where does the presupposition come from?
  - The word ›manage’ – let’s replace it by ‘try‹

## Example

(5) John tried to stop in time.

(6) John didn’t try to stop in time.

(5) is not presupposed by (6).

## Presupposition triggers

- Some words trigger presuppositions
- Trigger words have been collected and categorized

# Presupposition triggers

- Definite descriptions
  - John saw/didn't see the man with two heads

  $\rightarrow$ there exists a man with two heads

# Presupposition triggers

- Definite descriptions
  - John saw/didn't see the man with two heads
  
  $\rightarrow$ there exists a man with two heads
- Implicative verbs
  - John forgot/didn't forget to lock the door
  
  $\rightarrow$ John ought to have locked, or intended to lock, the door

# Presupposition triggers

- Definite descriptions
  - John saw/didn't see the man with two heads
  
  $\rightarrow$ there exists a man with two heads
- Implicative verbs
  - John forgot/didn't forget to lock the door
  
  $\rightarrow$ John ought to have locked, or intended to lock, the door
- Iteratives
  - The flying saucer came/didn't come again
  
  $\rightarrow$ The flying saucer came before

# Presupposition triggers

- Definite descriptions
  - John saw/didn't see the man with two heads
  - $\rightarrow$ there exists a man with two heads
- Implicative verbs
  - John forgot/didn't forget to lock the door
  - $\rightarrow$ John ought to have locked, or intended to lock, the door
- Iteratives
  - The flying saucer came/didn't come again
  - $\rightarrow$ The flying saucer came before
- Temporal clauses
  - Before Strawson was even born, Frege noticed/didn't notice presuppositions
  - $\rightarrow$ Strawson was born

# Presupposition triggers

- Definite descriptions
  - John saw/didn't see the man with two heads
  
  $\rightarrow$ there exists a man with two heads
- Implicative verbs
  - John forgot/didn't forget to lock the door
  
  $\rightarrow$ John ought to have locked, or intended to lock, the door
- Iteratives
  - The flying saucer came/didn't come again
  
  $\rightarrow$ The flying saucer came before
- Temporal clauses
  - Before Strawson was even born, Frege noticed/didn't notice presuppositions
  
  $\rightarrow$ Strawson was born
- Comparisons and contrasts
  - Marianne called Adolph a male chauvinist, and then HE insulted HER
  
  $\rightarrow$ For Marianne to call Adolph a male chauvinist would be to insult him
- ...

# Presupposition properties

- So far: Presuppositions
  - are implicit assumptions about the world
  - survive under negation
- Now:
  - Defeasibility

# Presupposition

## Defeasibility

- Presuppositions can be cancelled/prevented/defeated

# Presupposition

## Defeasibility

- Presuppositions can be cancelled/prevented/defeated
- By background knowledge (that John didn't to a PhD)
  - At least John won't have to regret that he did a PhD.

# Presupposition

## Defeasibility

- Presuppositions can be cancelled/prevented/defeated
- By background knowledge (that John didn't to a PhD)
  - At least John won't have to regret that he did a PhD.
- By the meaning of the sentence

(1) Sue cried before she finished her thesis.

$\rightarrow$ Sue finished her thesis

- ›before‹ triggers a presupposition

# Presupposition

## Defeasibility

- Presuppositions can be cancelled/prevented/defeated
- By background knowledge (that John didn't to a PhD)
  - At least John won't have to regret that he did a PhD.
- By the meaning of the sentence

(1) Sue cried before she finished her thesis.

$\rightarrow$ Sue finished her thesis

- ›before‹ triggers a presupposition

(2) Sue died before she finished her thesis.

$\nrightarrow$ Sue finished her thesis

# Presupposition

## Defeasibility

- By more context

(1) He isn't aware that Serge is on the KGB payroll
$\rightarrow$ Serge is on the KGB payroll

# Presupposition

## Defeasibility

- By more context
  (1) He isn't aware that Serge is on the KGB payroll
  $\rightarrow$ Serge is on the KGB payroll
  - A: Well we've simply got to find out if Serge is a KGB infiltrator
    B: Who if anyone would know?
    C: The only person who would know for sure is Alexis; I've talked to him and he isn't aware that Serge is on the KGB payroll. So I think Serge can be trusted
- A specific discourse context can override a presuppositional inference

# Evaluation

# Introduction

- We *always* want to know how well machine learning works
- Straightforward evaluation: Comparison with a gold standard

# Introduction

- We *always* want to know how well machine learning works
- Straightforward evaluation: Comparison with a gold standard
- Most simple metric: Accuracy
  - Percentage of correctly classified instances (the higher the better)
  - Inverse: Error rate (percentage of incorrectly classified instances)

# Introduction

- We *always* want to know how well machine learning works
- Straightforward evaluation: Comparison with a gold standard
- Most simple metric: Accuracy
  - Percentage of correctly classified instances (the higher the better)
  - Inverse: Error rate (percentage of incorrectly classified instances)
- Accuracy is nice, but not enough
  - When improving systems, we want to *compare* our accuracy with the previous accuracy
  - When developing new systems, we want to know how difficult the task is
    - E.g., 60% accuracy when distinguishing 35 parts of speech is better than 60% accuracy when distinguishing nouns and all the rest

# Baseline

The baseline performance is the performance of a simple system, rule or thought experiment

# Baseline

The baseline performance is the performance of a simple system, rule or thought experiment

- Example 1: Gender of students in Stuttgart and Cologne
  - Task: Classify students according to their gender
  - Data
    - Stuttgart: $8585$ of $25\,705$ students are female
    - Cologne: $29\,793$ of $48\,841$ students are female
  - Majority baseline: Everyone is female (Cologne) or male (Stuttgart)
  - Classification accuracies: $61\%$ / $66.6\%$

# Baseline

The baseline performance is the performance of a simple system, rule or thought experiment

- Example 1: Gender of students in Stuttgart and Cologne
- Example 2: Gender of arbitrary Germans
  - Task: Classify a random German according to their gender
  - male: 40.7m vs. female: 41.8m
  - Random baseline: Toss a coin
  - Classification accuracy: about 50%

# Baseline

The baseline performance is the performance of a simple system, rule or thought experiment

- Example 1: Gender of students in Stuttgart and Cologne
- Example 2: Gender of arbitrary Germans
- Example 3: Detecting nouns
  - Task: Classify words into noun and non-noun
  - Most words are not nouns
  - Majority baseline: Every word is a non-noun
  - Accuracy (in a German text): 81.8%

# Looking Closer

- Not all errors are the same
  - A verb can be wrongly classified as noun
  - A noun can be classified wrongly as something else
- Errors can be different for different classes
  - Detection of nouns might be better than verbs

⇒ Precision and recall Manning and Schütze (MS99, pp. 267 sqq.)
  - German: ›Genauigkeit‹ and ›Sensitivität‹
  - Other metrics in other disciplines (e.g., ›Spezifizität‹ in virology)

# Precision and Recall

- Both are calculated *per class*

# Precision and Recall

- Both are calculated *per class*

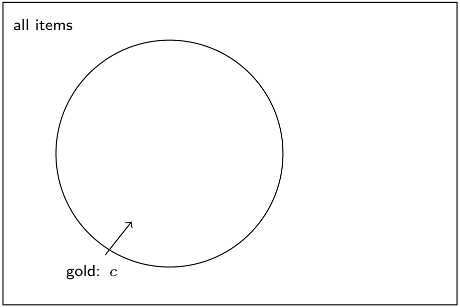



Figure: Identifying true/false positives/negatives for class $c$

# Precision and Recall

- Both are calculated *per class*

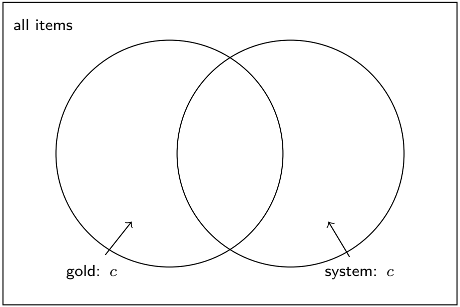



Figure: Identifying true/false positives/negatives for class $c$

# Precision and Recall

- Both are calculated *per class*

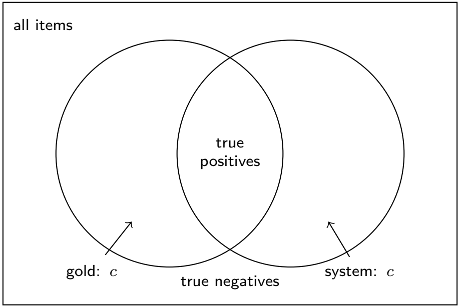



Figure: Identifying true/false positives/negatives for class $c$

# Precision and Recall

- Both are calculated *per class*

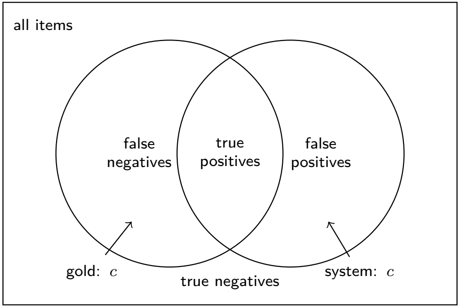
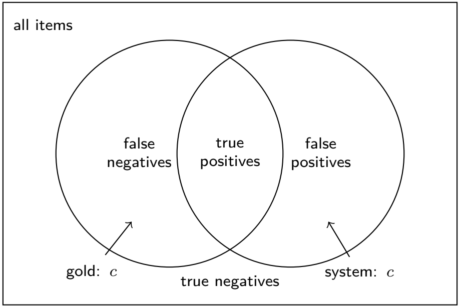



Figure: Identifying true/false positives/negatives for class $c$

# Precision and Recall



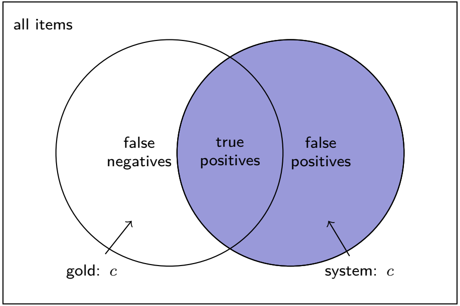

true positives Correctly identified items of class $c$

true negatives Correctly identified items of other classes

false positives System predicts $c$, but it's another class

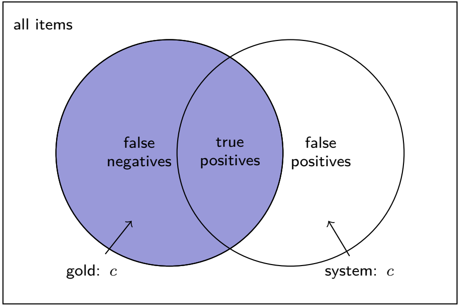
false negatives System predicts something else, but it's $c$

# Precision and Recall




precision How many of the items predicted as $c$ are actually correct? $P = \frac{tp}{tp+fp}$

# Precision and Recall




**precision** How many of the items predicted as $c$ are actually correct? $P = \frac{tp}{tp+fp}$

**recall** How many of the items that are $c$ are actually identified? $R = \frac{tp}{tp+fn}$

# Evaluation

## Precision and Recall

**precision** How many of the items *predicted as* $c$ are actually correct?

**recall** How many of the items that *are in class* $c$ are actually found by the system?

- Precision and recall measure different kinds of errors the systems make
  - Precision errors are often easier to spot for humans
  - Recall errors are hurtful, if only instances of one class are looked at or analyzed – missing instances will never be found
- Average P/R values over all classes are often given
- Sometimes combined into an $f_1$-score
  - $f_1 = 2\frac{precision*recall}{precision+recall}$
  - 'harmonic mean' between the two

# Example: Adjective Detection

## Gold Standard

*Die arme Leonore! Und doch war ich unschuldig.*

Goethe, *Die Leiden des jungen Werther*

# Example: Adjective Detection

## Gold Standard

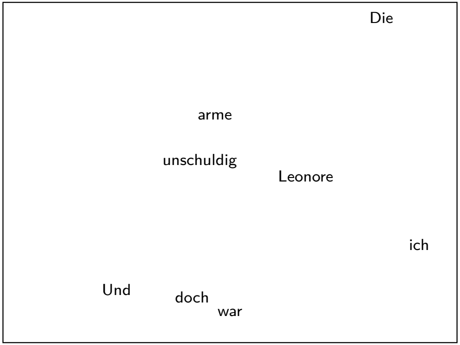
*Die arme Leonore! Und doch war ich unschuldig.*

Goethe, *Die Leiden des jungen Werther*

| Adj | Program 1 | Program 2 |
|---|---|---|
| + | arme | arme, unschuldig, Leonore |
| – | Die, Leonore, Und, doch, war, ich, unschuldig | Die, Und, doch, war, ich |

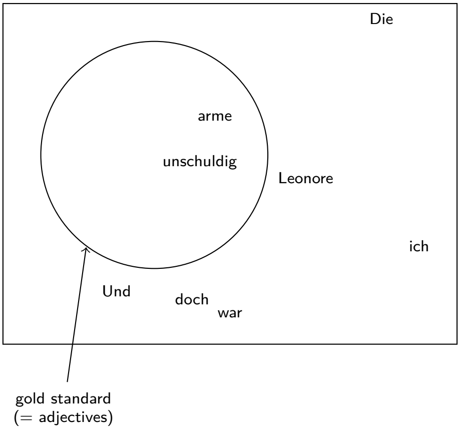
# Example: Adjective Detection

## Different kinds of errors, visually for program 2



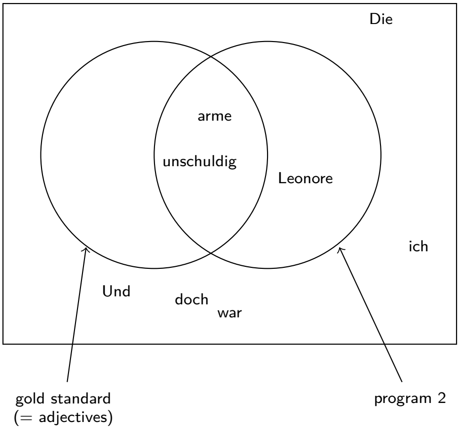
# Example: Adjective Detection

## Different kinds of errors, visually for program 2



# Example: Adjective Detection

## Different kinds of errors, visually for program 2



# Example: Adjective Detection

## Different kinds of errors, visually for program 2

$$
\begin{aligned}
precision &= \frac{tp}{tp + fp} \\
&= \frac{2}{2+1} = 0.66 \\
recall &= \frac{tp}{tp + fn} \\
&= \frac{2}{2} = 1 \\
f_1 &= 2\frac{precision * recall}{precision + recall} \\
&= 2\frac{0.66 * 1}{0.66 + 1} = 0.8
\end{aligned}
$$

# Summary

- Pragmatics: Language and the world
  - Some linguistic expressions have impact on the world
  - Some choices that we make are influenced by non-linguistic factors
- Evaluation
  - Classification: Sort things into previously known categories
  - Precision: Percentage of retrieved items that are correct
  - Recall: Percentage of target items that were retrieved
  - F-Measure: Harmonic mean between P and R