Group Assignment 3 Information extraction of a timeline from a biography

Stat date: 30 September 2024

Due date: 21 October 2024

Very clear **group** assignment with different roles for group members

Group Assignment 3 Information extraction of a timeline from a biography

Connected to Lecture on Information Extraction.

In the assignment you are asked **to extract entities**: time expressions

We look at events in the lives of three famous people and we can loosely define an event as "something that happened at a specific point in time and is worthy to be reported"

The underlying assumption is that each date represents an event

ISO standard https://en.wikipedia.org/wiki/ISO_8601

5 December 2020 -> '2020-12-05'
1999 -> '1999'

Group Assignment 3 Information extraction of a timeline from a biography

Learning goals of this assignment:

- Become familiar with the intricacies and challenges of time expression labelling
- Practice with manual annotation using BIO tags
- Learn how to compute inter-annotator agreement scores with Cohen's kappa, and understand how to interpret these kappa scores
- Define a set of patterns for extracting time patterns from text
- Calculate precision for the output of your code on an unseen text
- Gain insight into the importance of pattern generalizability
- Reflect on the gap between labeling time expressions and the actual IE task of extracting and matching events from texts.

Group Assignment 3 IE: two parts

Part 1:

- manual annotation of time expressions in text using BIO labels (individually)
- compute inter-annotator agreement scores with Cohen's kappa

Part 2:

- create a rule-based application to extract time expressions from text
- evaluate your rule-based approach
- reflect on your findings

We do not expect a perfect solution for this assignment: we look forward to your report with your reflections on the challenges in this task.

Data sample creation

Manual labeling for data set creation

Developing Text and multi-media Mining applications

You need a manual gold-standard labeled data sample for training /tuning /testing

For rule-based or ML-based approaches

Manual labeling: language is often vague, ambiguous and interpretable in different ways and within different contexts

‘I like your pants` (do not say this in the UK!)

How to get example data

2. Manually label data
   - Make a selection of documents
   - Define a set of categories
   - Human classification
     - Experts
     - Crowdsourcing (Amazon Mechanical Turk, Prolific)

- How many examples do you need?
  - At least dozens/hundreds per category
  - The more, the better
  - The more difficult the problem, the more examples needed

Inter-rater agreement

- Human labelled reference data = ‘gold standard’ / ‘ground truth’

- But 2 human classifiers do never fully agree
- We therefore always have part of the example data labelled by 2 or 3 raters
- and then compute the inter-rater agreement
- to know the reliability of the example data

- Measure for inter-rater agreement: Cohen’s Kappa

Classic metric, still often used, some critical remarks like kappa is sensitive to sample size.

Cohen's Kappa

Agreement table		A2	
		Yes	No
A1	Yes	20	5
	No	10	15

- Pr(a) = actual (measured) agreement: percentage agreed
- Pr(e) = expected (chance) agreement
- Pr(a) = (20+15)/(20+5+10+15) = 35/50 = 0.70
- Pr(e):
  - A1 says 'yes' to 25 and 'no' to 25 → 50% of the time
  - A2 says 'yes' to 30 and 'no' to 20 → 60% of the time
  - Pr(e,yes) = 0.50*0.60 = 0.30
  - Pr(e,no) = 0.50*0.40 = 0.20
  - Pr(e) = Pr(e,yes) + Pr(e,no) = 0.50
- K = (0.70 – 0.50)/(1 – 0.50) = 0.20/0.50 = 0.40

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$