UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR INFORMATIONSSYSTEME

Context is the Key:

Context-aware Corpus Annotation Using Subjective Content Descriptions

Colloquium

Felix Kuhr

Institute of Information Systems

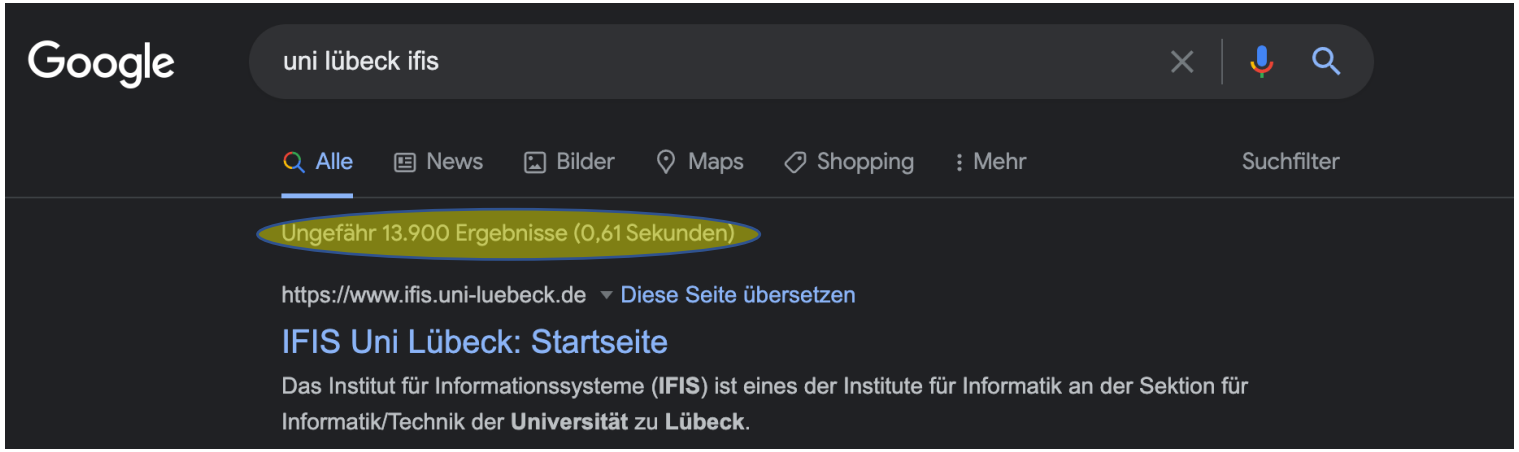
University of Lübeck

March 24, 2022

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:
  - Compare words in a query with content of documents

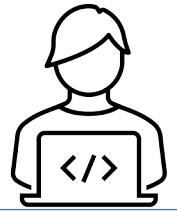
An information retrieval system can be characterized by:

- **Corpus** of documents
- **Queries** posed in a query language
- **Result** set of relevant documents
-



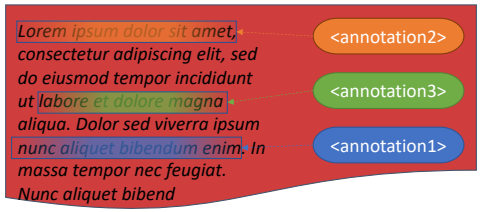
Are *documents* in the result set relevant to the information need of a user?

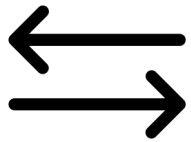
Annotation Systems

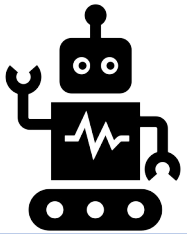


Manual Annotation Systems

- ⊕ Quality of annotations
- ⊕ User-centric annotations
- ⊖ Human annotation experts
- ⊖ High costs / time-consuming

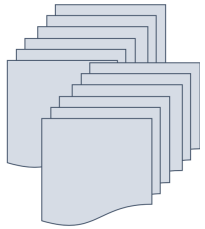






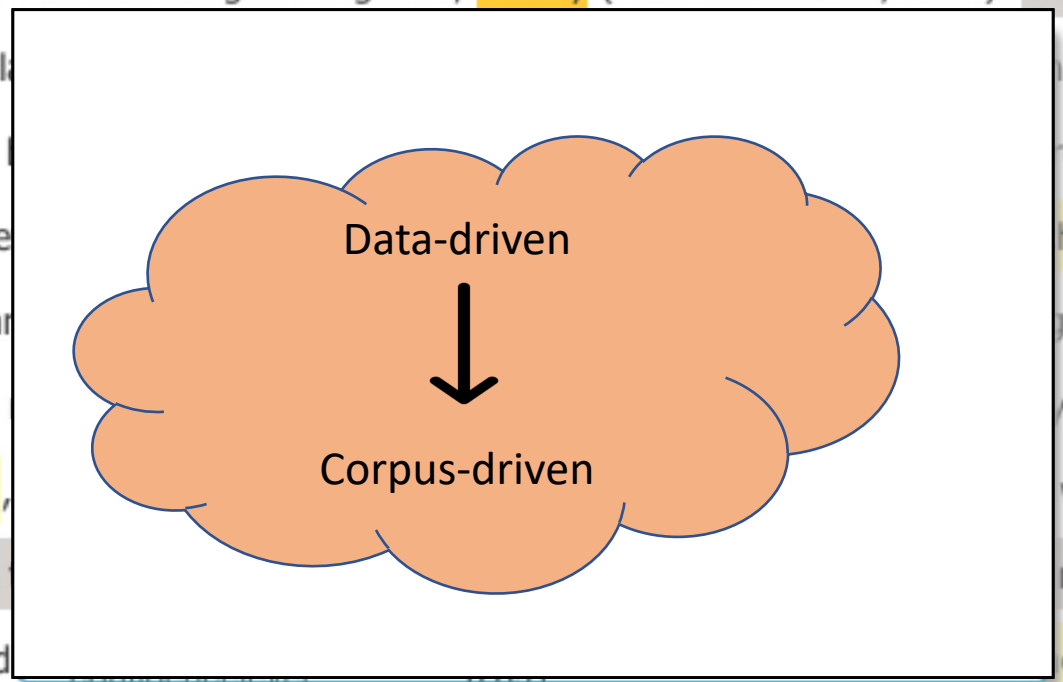
Automatic Annotation Systems

- ⊕ Low costs
- ⊕ Fast annotation process
- ⊖ Quality of annotations
- ⊖ Missing explainability



Context is the Key: Example (One Calais)*

January 6, 1984). Ivana became a naturalized United States citizen in 1988. By early 1990, Trump's troubled marriage to Ivana and affair with actress Marla Maples had been reported in the tabloid press. They were divorced in 1992.. Trump married his second wife, actress Marla Maples in 1993. They had one daughter together, Tiffany (born October 13, 1993). The couple were separated in 1997 and l… Slovene model Melania Knauss, who l… e married on January 22, 2005, at Be… h, Florida. In 2006, Melania becar… gave birth to their son, whom they … his doctor, Harold Bornstein M.D., … were in normal range. Trump says … ncluding marijuana. He also does not d… olism. His BMI, according to his



Add meta data to text by linking extractable entities to external data

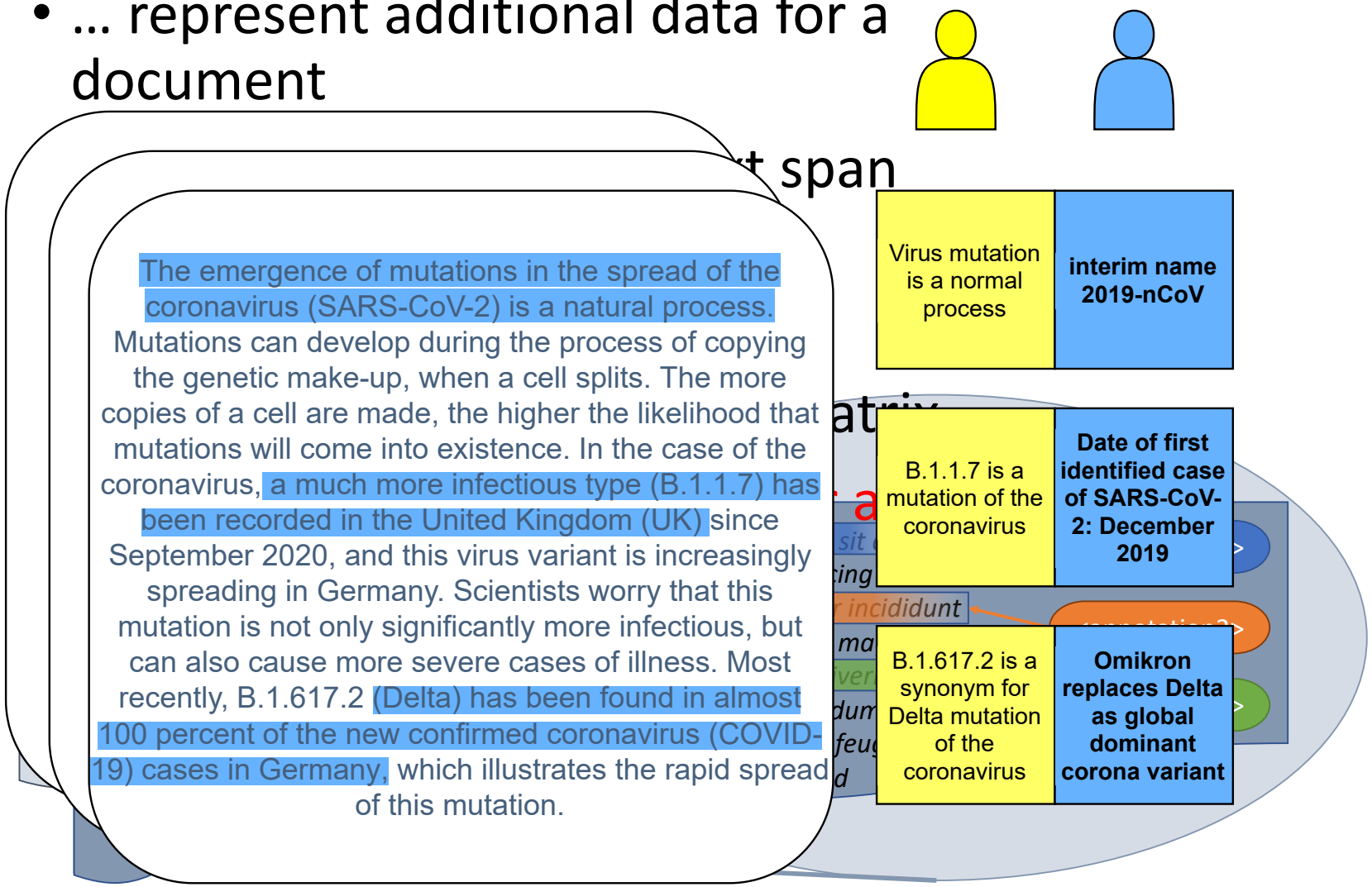
Text related with Company Tiffany & Co.

Persons with names also used for companies

* https://permid.org/onecalaisViewer

Subjective Content Descriptions (SCDs)

- … represent additional data for a document



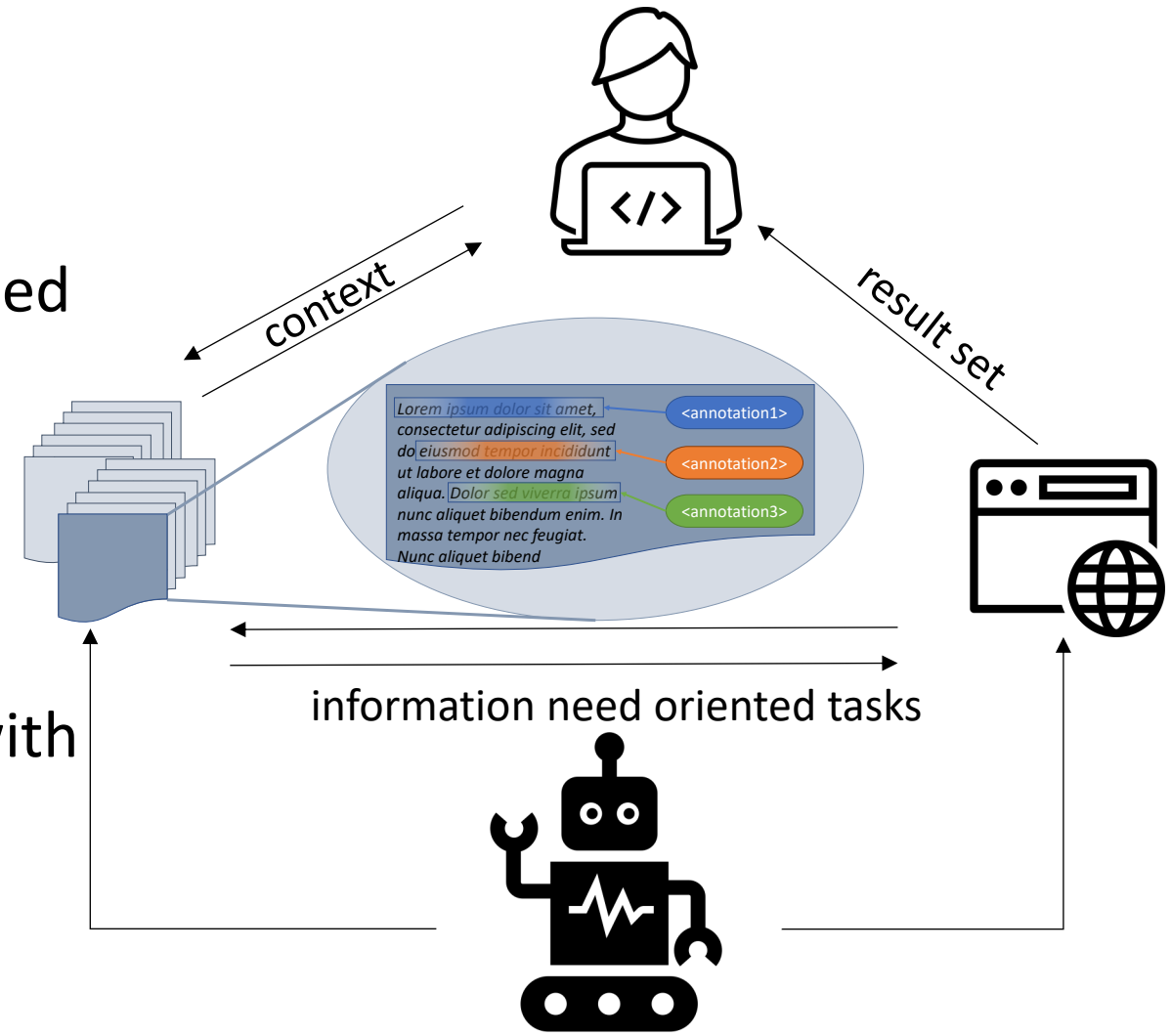
Different Typs of SCDs:

- Additional definitions
- Links to external sources
- Relational data to clarify dependencies between entities

Lead Scenario: Information Retrieval (IR) Agent

- Agent‘s goal: Meet information need of a person
- Agent is working on an IR-corpus that represents a model for the information need of a person
- Agent optimizes the model to meet the information need
- Documents in the corpus are associated with annotations

→ Subjective Content Descriptions (SCDs)

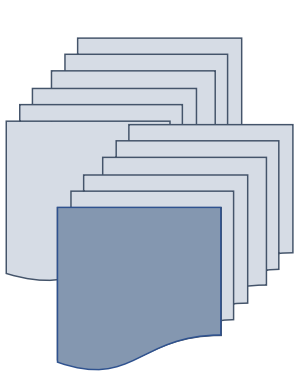


Foundations: SCD-Word Distribution[2]

- SCD-word distribution results from SCDs associated with *windows* in documents
- For each SCD estimate relative weighted frequency of words
- Binomial distribution to represent weights

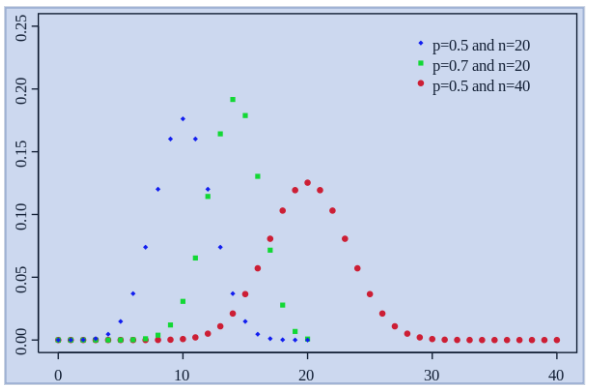
Algorithm 1 Forming SCD-word probability distribution matrix $\delta(\mathcal{D})$

```
1:  function BUILDMATRIX(Corpus D)
2:      Input: Corpus D
3:      Output: SCD-word probability distribution matrix δ(D)
4:      Initialize an m × V matrix δ(D) with zeros
5:      for each d ∈ D do
6:          for each t ∈ T(d) do
7:              for ρ of t do
8:                  for each w ∈ win_{d,ρ} do
9:                      δ(D)[t][w] += I(w, win_{d,ρ})
10:     Normalize δ(D)[t]
11:     return δ(D)
```



Corpus $\mathcal{D}$

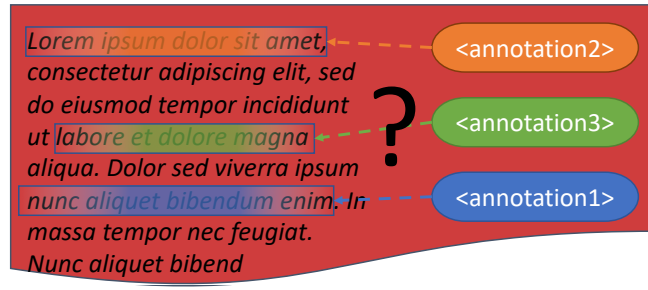
$$\delta(\mathcal{D}) = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{cccccc} w_1 & w_2 & w_3 & \cdots & w_n \end{array} \\ \begin{bmatrix} v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{bmatrix} \end{array}$$

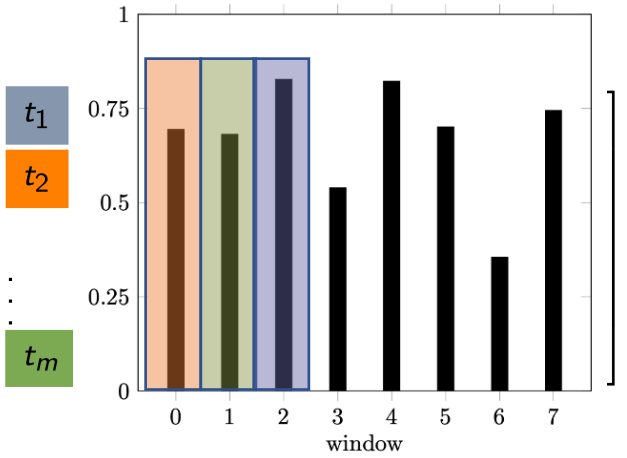


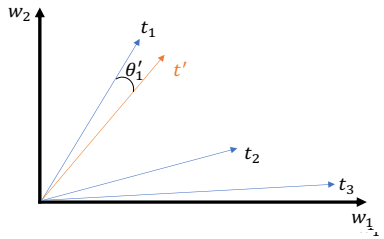
[2] Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of AI 2019: Advances in Artificial Intelligence, 2019

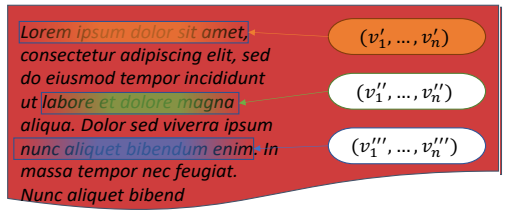
Most-Probably Suited SCDs (MPSCDs)[2]

- Given: Sequence of text windows in a (new) document
- Words in window define word distribution (t')
- Goal: Estimate MPSCDs for windows in document
- Each vector v_i defined in SCD-word matrix $\delta(D)$ defines an angle with window word vector t'
- Define function $MPSCD(M, t')$
  - Provides SCD t_i that is associated with SCD-word matrix vector v_i with smallest angle to window word vector t'
- MPSCD with similarity measure is applied to each window of a document





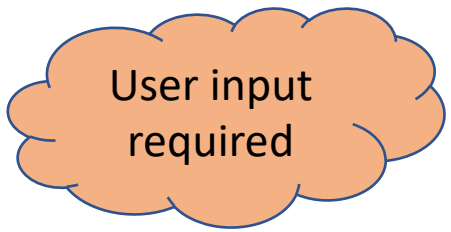


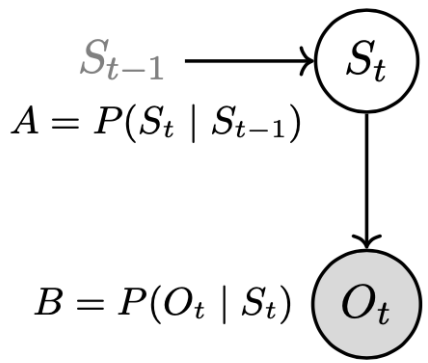


[2] Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of AI 2019: Advances in Artificial Intelligence, 2019

Context-specific Corpus Enrichment [3]

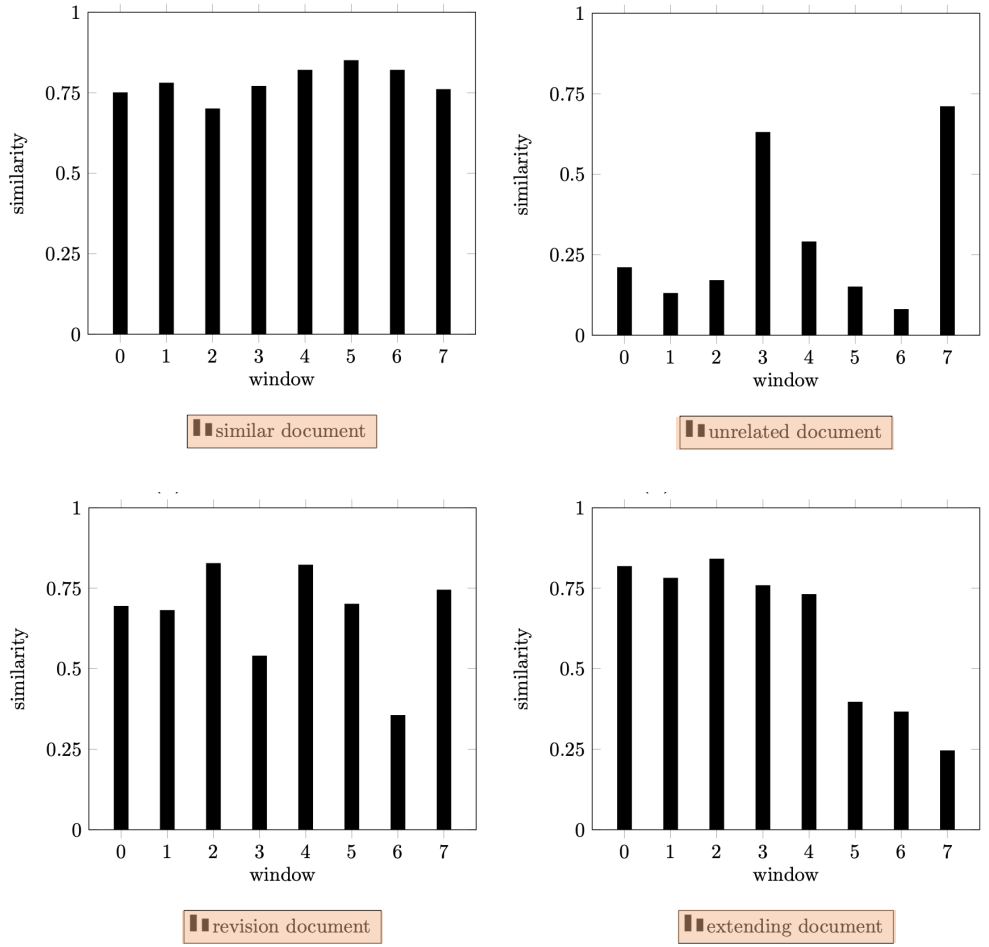
- Goal: Add new documents to IR corpus with an initial set of SCDs already associated with documents in the corpus
- Different Categories: sim, unrel, rev, ext
- Given: 4 category HMMs, each associated with a category label
- HMM Learning by using Baum-Welch Algorithm





Related content
or
Unrelated content

MPSCD similarity value of the window (discretized)



[3] Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020

Context-specific Corpus Enrichment - Decision Making Process

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Given: new document

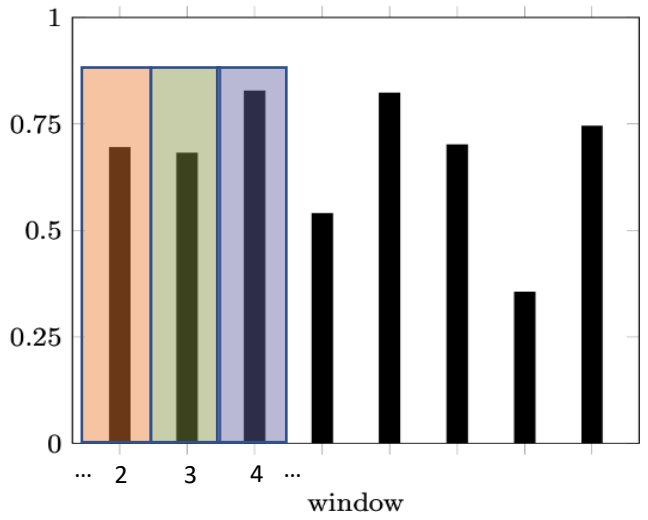
	w_1	w_2	w_3	$\cdots$	w_n
t_1	$v_{1,1}$	$v_{1,2}$	$v_{1,3}$	$\cdots$	$v_{1,n}$
t_2	$v_{2,1}$	$v_{2,2}$	$v_{2,3}$	$\cdots$	$v_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_m	$v_{m,1}$	$v_{m,2}$	$v_{m,3}$	$\cdots$	$v_{m,n}$

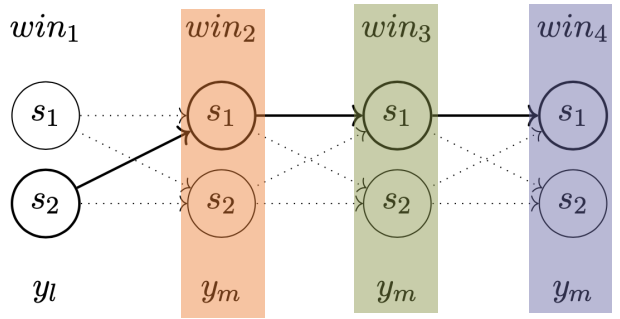
Given: SCD-word distribution of IR corpus

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

<annotation2>
<annotation3>
<annotation1>

Determine the MPSCD sequence for the window sequence of the new document based on available SCD-word distribution





Focus on SCD similarity values

Discretize similarity values:

$y_l: 0 - 0.3$, $y_m: 0.3 - 0.75$, $y_h: 0.75 - 1$

Determine MPE sequence w.r.t. each category HMM on sequence of MPSCD similarity values.

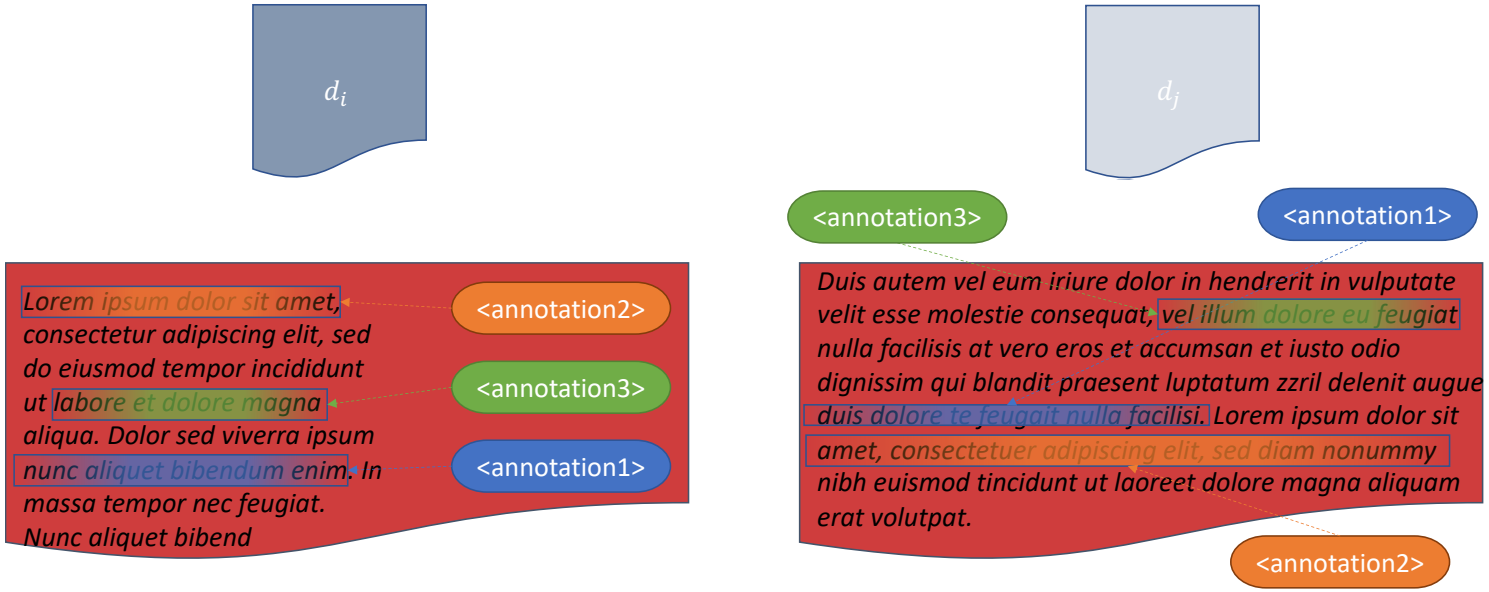
Take category of HMM with most-likely MPE sequence as classification

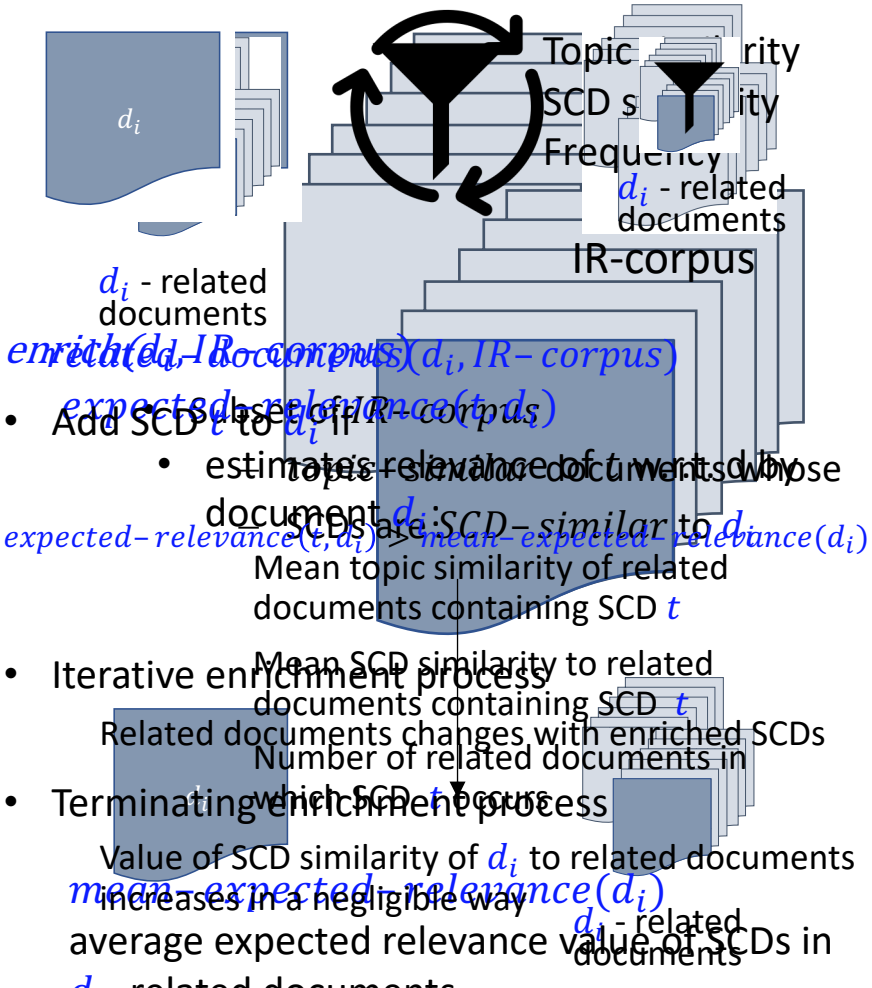
Focus on content of SCDs

Extend corpus based on document category and transfer SCDs above a threshold

[4]

Corpus-Driven Document Enrichment using SCDs





Goal: Enrich a document with *relevant* SCDs associated with other documents in an IR-corpus.

Fixed-point iteration procedure:

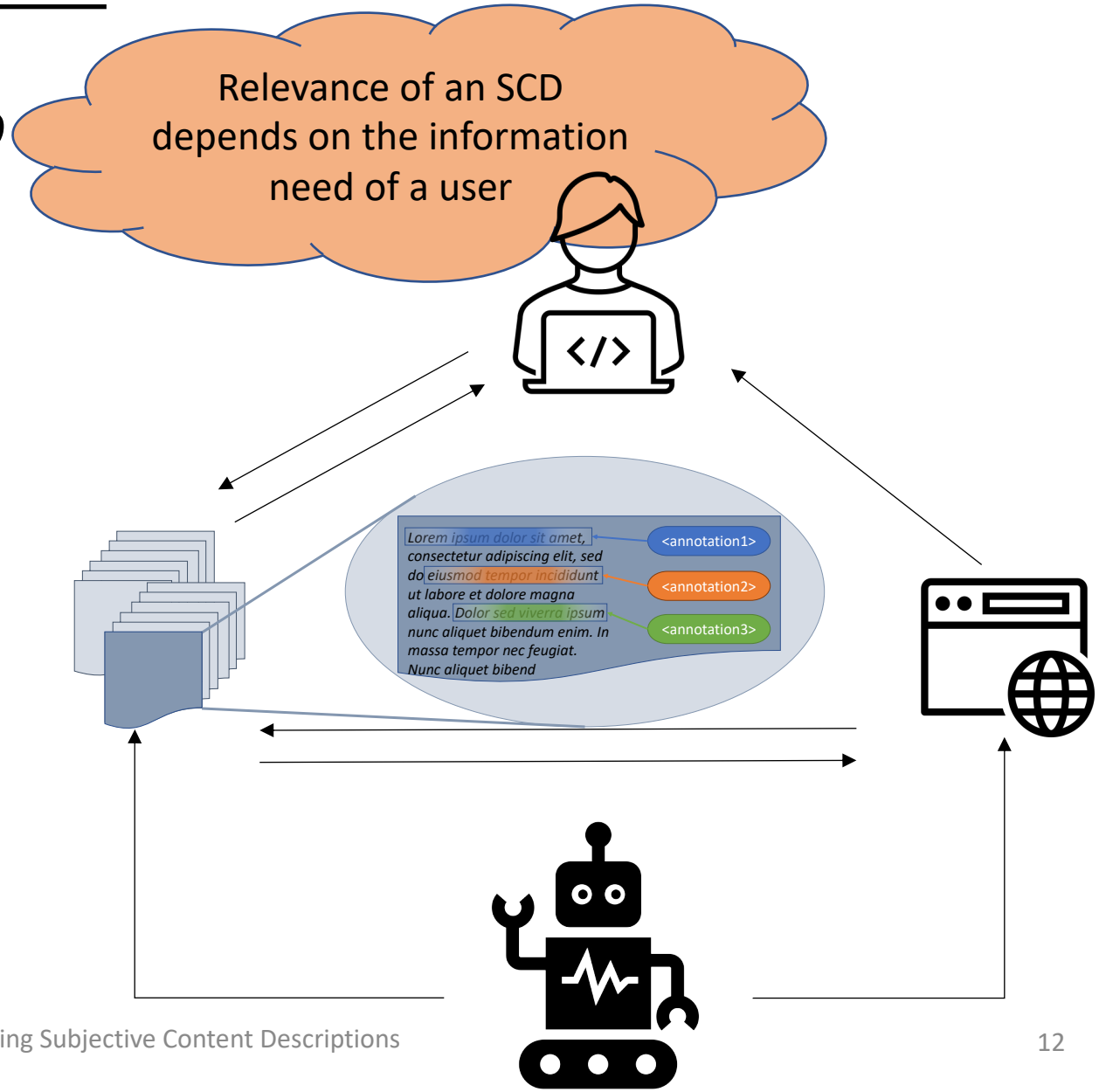
- determine the expected related documents in IR-corpus D,
- determine the set of SCDs T from D that are newly added to d, then
- determine the expected related documents D again, and so one
- until no more SCDs are assigned to document d.

- Add SCD t to d_i's set of IR-corpus $related\_documents(d_i, IR-corpus)$
  - estimate relevance of t w.r.t. d_i: $expected\_relevance(t, d_i)$
    - topic similarity of documents whose SCDs are similar to d_i
    - Related documents changes with enriched SCDs
    - Value of SCD similarity of d_i to related documents increases in a negligible way
- Iterative enrichment process
- Terminating enrichment process
  - $expected-relevance(t,d_i) > mean-expected-relevance(d_i)$
  - average expected relevance value of SCDs in d_i -related documents

[4] Felix Kuhr, Bjarne Witten, Ralf Möller: Corpus-Driven Annotation Enrichment. Proceedings of the 13th IEEE International Conference on Semantic Computing (ICSC-19), 2019

Expected Relevance Value

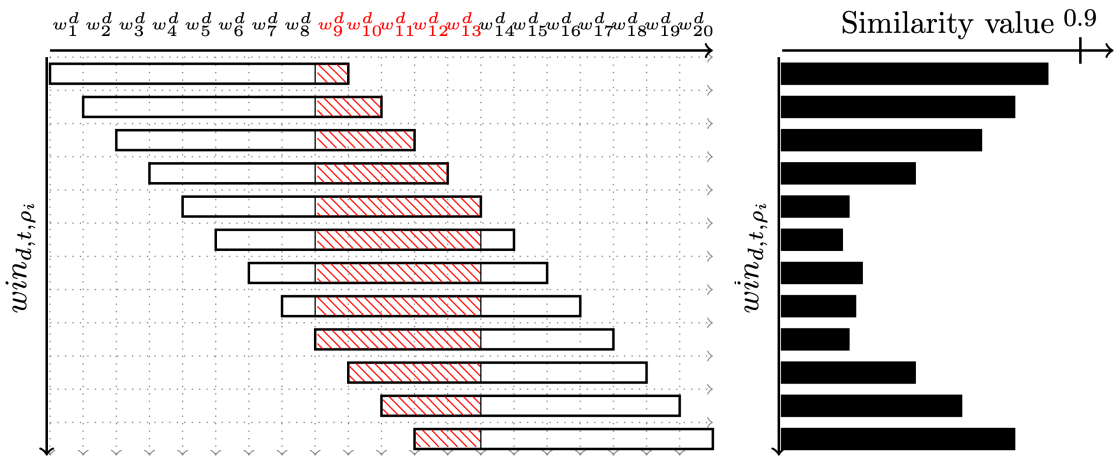
- Given: document d_j from IR-corpus D
- Question: What is the expected relevance value of an SCD associated to a related document?
- Some ways to adjust performance:
  - Similarity between documents
  - Similarity between SCDs
  - Frequency of SCDs



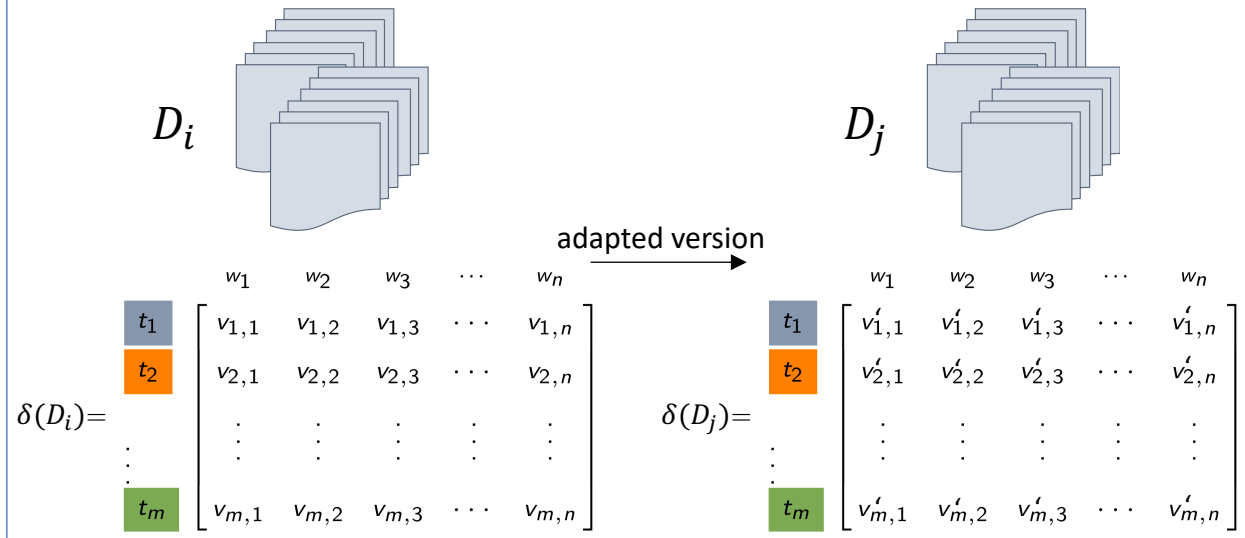
Bootstrap Approaches for SCDs

Inline SCDs [5]

- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text
- Estimate MPSCDs and use trained HMM to analyse sequence of corresponding SCD similarity values
  - Small similarity values → different content → new inline-SCDs
  - Inline-SCD = Content of window
  - Inline-SCD represent new row in SCD word matrix



Adapting SCD word distribution from another IR-corpus [6]



- Adapt SCD word distribution from IR-corpus D_i to documents in D_j
  - Analyze difference in word distributions of documents in corpus D_i and D_j
  - Reweight word distribution for each SCD in $\delta(D_i)$ s.t. distribution fits for D_j
  - Remove documents from D_i not relevant for D_j

[5] Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)

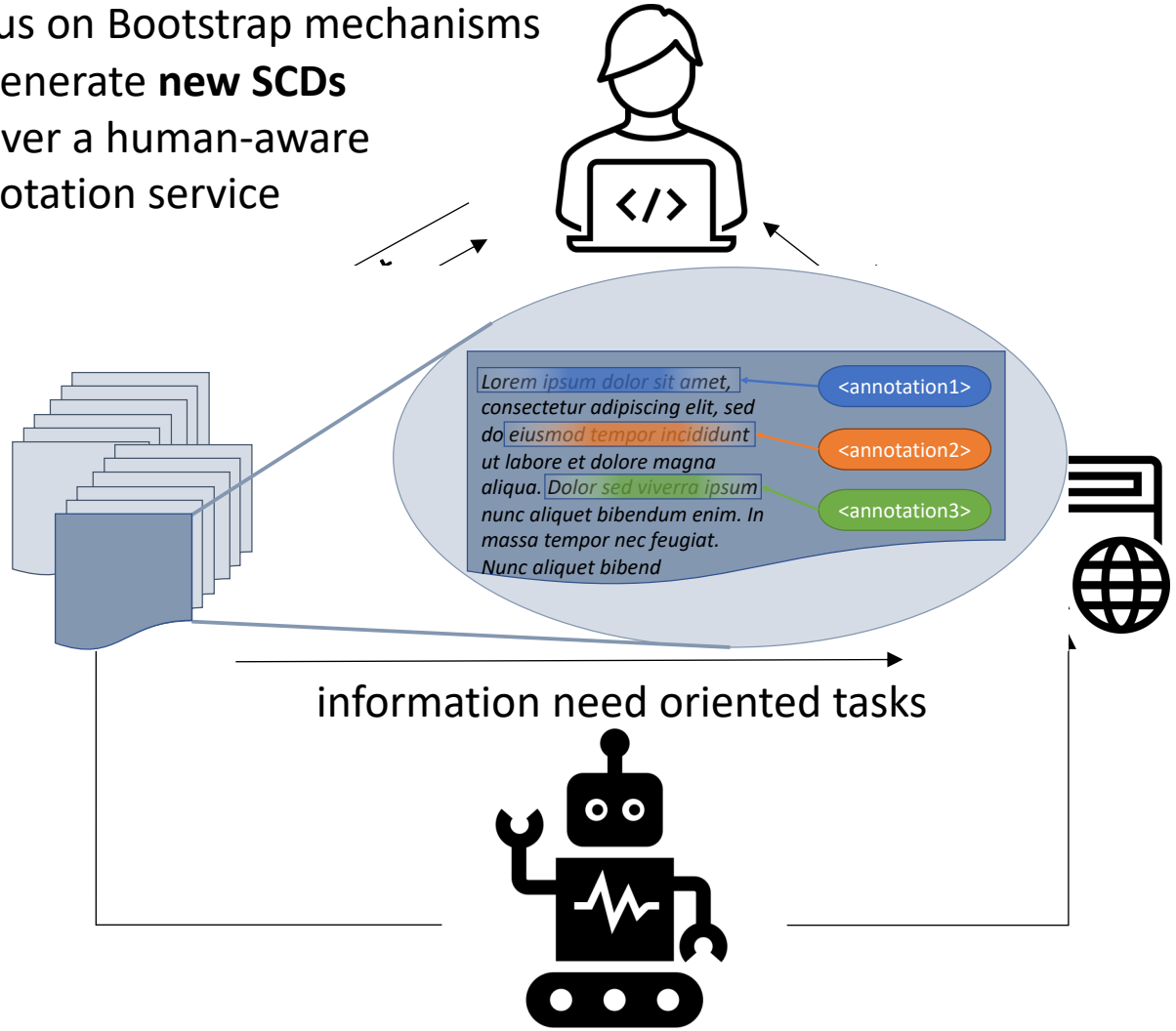
[6] Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)

Conclusion

- Human-aware information retrieval considering not only content of documents and queries
- Fully automated annotation approach considering the human information need represented by a corpus and SCDs
- Approach for the bootstrap problem considering inline-SCDs

Future Work:

- Focus on Bootstrap mechanisms to generate **new SCDs**
- Deliver a human-aware annotation service



Focus on human-aware AI approaches:
→ Data linking **services** in a fashion that takes into aware **human expectations**