FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Johannes Franke

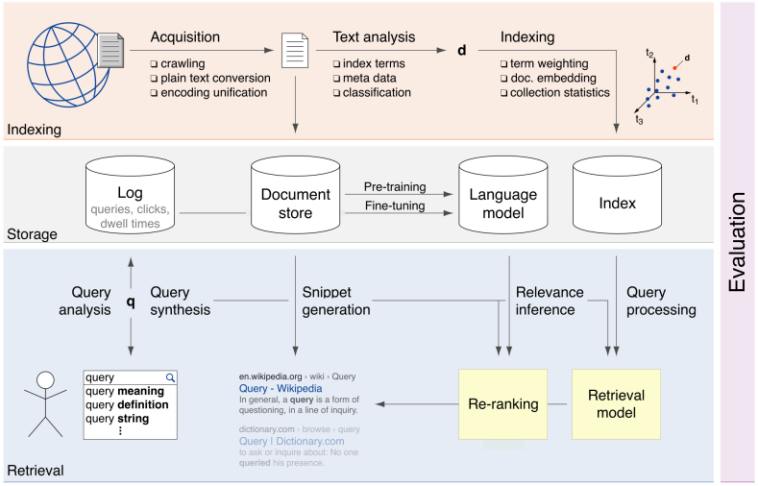
IR: Query Understanding WiSe 24/25

Query Segmentation

05.11.2024

Grundlagen

- Aufteilung in semantische Einheiten (Token) Sequenzen
- Query mit ***n*** Token hat ***n – 1*** Segmentierungen
- Segmentierungen haben Einfluss auf Bedeutung
- Einfachster Ansatz: Wörterbücher



Inspiration

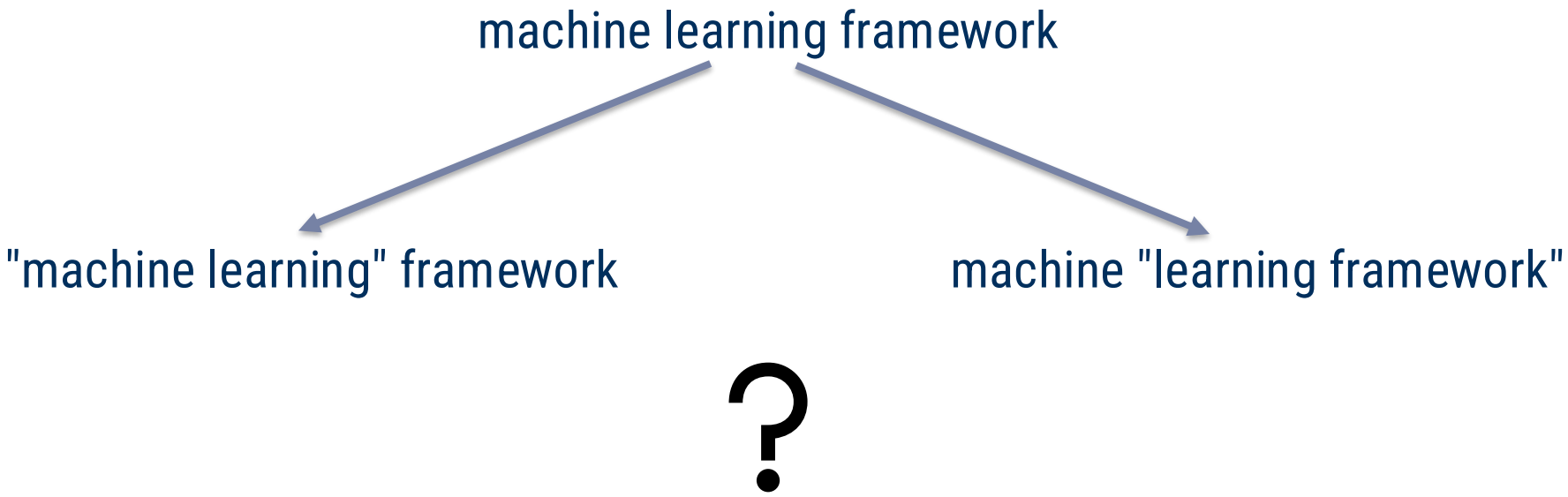
Ziele

- Verbesserung der Precision
- Erkennung korrekter semantischer Bedeutung
- Optimale Aufteilung und Kombination der Query

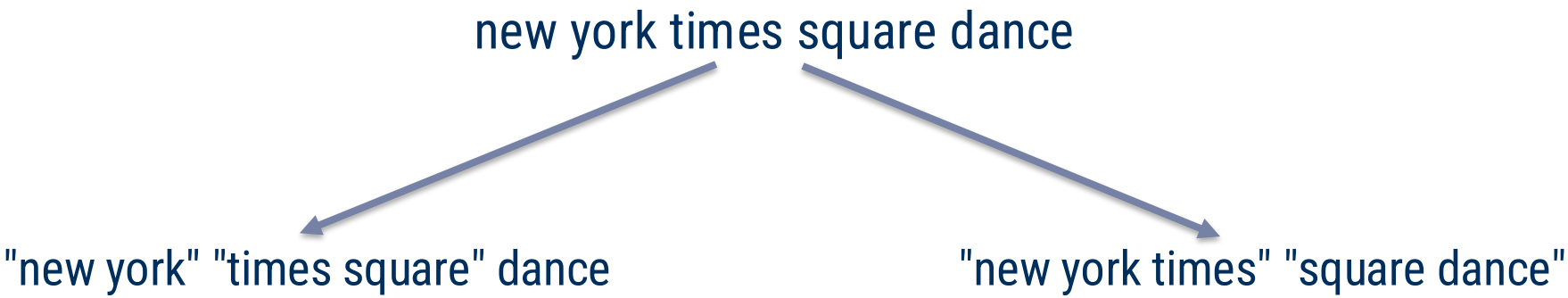
Auswirkungen

- Nutzer erhalten gezieltere, relevantere Dokumente
- Verkürzt Zeit & Erhöht Zufriedenheit beim Retrieval

Beispiele


machine learning framework
"machine learning" framework
machine "learning framework"
?

Beispiele


new york times square dance
"new york" "times square" dance
"new york times" "square dance"
?

Paper

- *"Query Segmentation for Web Search"*
  – Risvik et al., 2003
- *"Learning Noun Phrase Query Segmentation"*
  – Bergsma et al., 2007

Query Segmentation for Web Search

- Ansatz: Definition einer "guten" Token-Sequenzen *S*
  - 2- bis 4-gramme
  - trifft "häufig" in Corpus auf
  - "gute" Transinformation (MI)

$$\text{conn}(S) = \text{freq}(S) \cdot \text{I}(w_1 \ldots w_{n-1}, w_2 \ldots w_n)$$

```
34259:  (msdn library)[5110] (visual studio)[29149]
29149:  msdn[47658] library[209682] (visual studio)[29149]
5110:  (msdn library)[5110] visual[23873] studio[53622]
41:  (msdn library visual studio)[41]
7:  msdn[47658] (library visual studio)[7]
0:  msdn[47658] library[209682] visual[23873] studio[53622]
```

Learning Noun Phrase Query Segmentation

Idee

- Ansatz mittels Support Vector Machines
  - inkl. Feature Engineering + Decision Boundary
- Dataset: AOL search query database (2006)

Evaluation

- Aufteilung in Training/Validation/Test (je 500 Queries)
- manuelle Annotatoren stellten Ground Truth
- Evaluationsmaß: Seg-Acc. & Qry-Acc.

Table 1: Indicator features.

Name	Description
is-the	token x = "the"
is-free	token x = "free"
POS-tags	Part-of-speech tags of pair x_{L0} x_{R0}
fwd-pos	position from beginning, i
rev-pos	position from end $N - i$

$\{\ldots, w_{L2}, w_{L1}, w_{L0}, w_{R0}, w_{R1}, w_{R2}, \ldots\}$

Table 3: Segmentation Performance (%)

Feature Type	Feature Span	Test Set		Intersection Set	
		Seg-Acc	Qry-Acc	Seg-Acc	Qry-Acc
MI	Decision-Boundary	68.0	26.6		
Basic	Decision-Boundary	71.7	29.2		
Basic	Decision-Boundary, Context				
Basic	Decision-Boundary, Context, Dependency				
All	Decision-Boundary				
All	Decision-Boundary, Context				
All	Decision-Boundary, Context, Dependency				

Takeaways & Fragen

- Aufteilung der Query in Token-Sequenzen
- Ziel: semantische Bedeutung (Informationsbedürfnis) erkennen & Suchergebnisse verbessern
- Viele verschiedene Ansätze
  - Wörterbücher
  - MI-basiert
  - ML-basiert
  - etc.

Quellen

- https://queryunderstanding.com/query-segmentation-2cf860ade503
- https://www.algolia.com/blog/product/query-understanding-101/