STRATEGY COMPARISON FOR SEMANTIC ZERO-SHOT TAXONOMY FILTERS

OSSYM 2022 – 4th INTERNATIONAL OPEN SEARCH SYMPOSIUM

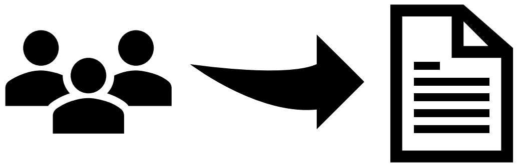
Andreas Hamm
DLR Institute for Software Technology

DLR

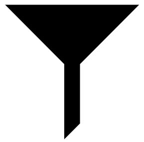
Searching vs Filtering

- Searching

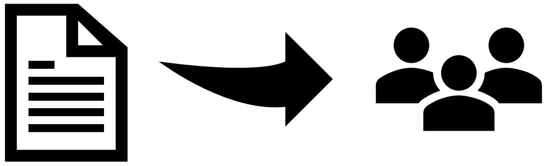
  - Information need formulated freely by users
  - Users know what they are looking for
  - Users find documents



- Filtering

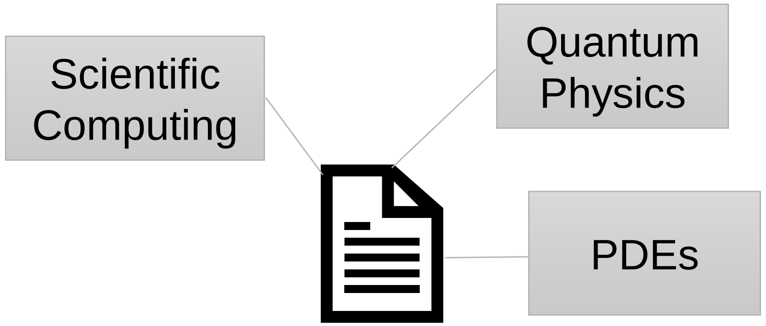


  - Categories / meta data / tags made available by information service provider
  - Users scan categories
  - Documents find users



Multi-Label Text Classification

- Tag a text with content-related labels from a predefined controlled vocabulary (label set)
  - Traditionally manual tasks requiring expert subject knowledge
  - Automation needed for large-scale document numbers and vocabulary sizes
    - Rule based approach via Boolean combination of search terms
    - ML approach with classifiers trained on labeled examples
      - State-of-the-art: Label-wise attention networks
    - These approaches require a lot of effort when introducing a new label set



- Aiming at a method that
  - Works with user-provided label set
  - Does not require label-set-dependent training
  - Works fast for large-scale situations

Zero-Shot Text Classification with Transformer Models

- Classifiers without explicit training on labeled examples
  - Use pretrained transformer-based models
- Zero-Shot Text Classification as sentence entailment problem (Yin, Hay, Roth 2019)
  - Use template „This is a text about …“ together with the class label as hypothesis
  - Use a transformer-based model to evaluate whether the text entails the hypothesis
  - Scales like N*M for N texts and M as size of the label set
- Zero-Shot Text Classification via sentence similarity
  - Use template „This is a text about …“ together with the class label as hypothesis
  - Use sentence-transformers (Reimers, Gurevych 2019) to transform sentences into vectors and calculate cosine similarity between text and hypothesis
  - Scales like N+M for N texts and M as size of the label set

Taxonomies

- Hierarchically structured label sets
- Wide-spread in many subject areas
- Examples used here (both with broad scope)
  - For scientific publications: OpenAlex concept hierarchy
    - Reduced version of the MAG concept hierarchy
    - 65k concepts on 6 levels
    - Example: *Mathematics* > *Geometry* > *Differential Geometry* > *Hyperbolic Geometry* > *Hyperbolic Triangle* > *Ultraparallel Theorem*
    - Many labels carry multilingual descriptions
    - Tested with samples from OpenAlex (English)
    - Base line: Attention-based classifier
  - For news articles: Media Topics of the International Press Telecommunications Council
    - All labels carry multilingual descriptions
    - 1350 categories on 5 levels
    - Example: *Politics* > *Government* > *Defense* > *Armed Forces* > *Military Service*
    - Tested with samples from Reuters (English) and APA (German language)
    - Base line: Rule-based classifier

Strategies for Improving Classification Results (1)

- Use label descriptions when generating hypotheses
  - *Differential geometry (branch of mathematics dealing with functions and geometric structures on differentiable manifolds)*
  - *Defense (anything involving the protection of one‘s own country)*
- Break down text into individual sentences
  - Do not aggregate sentence embedding vectors
  - Calculate similarity scores of labels for each sentence individually
  - Aggregate label scores, but with saturation (cf. BM25 ranking)
  - Consider all labels surpassing a score threshold
- Put higher weight on first sentence (typically the title)

Strategies for Improving Classification Results (2)

- Make use of hierarchical taxonomy structure
  - Proceed top-down
    - ☹ Relies on taxonomy quality
    - ☹ Relies on complete coverage by children
  - Take account of distance of labels in the hierarchy graph
    - ☹ Blurs semantic details on finer levels
  - Aggregate similarity scores bottom-up
    - ☺ Prefer labels along paths originating from highest scored labels on top levels
    - ☺ Eliminates misclassifications caused by homonyms
- Try several pretrained sentence transformer models

Assessing Multi-Label Classification Quality

- Benchmarking multi-label text classification is notoriously problematic
  - Impossible to decide about **the** correct labeling
  - Impossible to provide complete coverage of all labels
- Here: Mean precision (∅P) , mean recall (∅R), mean F1 (∅F1)
  - Compute per document the precision, recall, and F1 of predicted vs. „true“ labels
  - Average these over a sample of documents

Preliminary Assessment

OpenAlex	Conventional			Label			Description			Sentences			Hierarchy		
Model	∅P	∅R	∅F1	∅P	∅R	∅F1	∅P	∅R	∅F1	∅P	∅R	∅F1	∅P	∅R	∅F1
all-MiniLM-L6-v2	**60.6**	37.1	44.9	40.0	22.4	27.6	51.7	31.1	37.4	32.0	48.4	37.1	47.4	**52.4**	**47.1**
paraphrase-multilingual-MiniLM-L12-v2	**60.6**	37.1	**44.9**	27.5	14.6	18.6	30.3	14.2	18.8	28.9	15.9	18.6	25.7	**37.8**	30.6

Reuters	Conventional			Label			Description			Sentences			Hierarchy		
Model	∅P	∅R	∅F1	∅P	∅R	∅F1	∅P	∅R	∅F1	∅P	∅R	∅F1	∅P	∅R	∅F1
all-MiniLM-L6-v2	51.3	44.4	43.9	41.8	18.9	24.7	**59.5**	25.0	33.2	47.2	30.4	34.3	47.5	**45.9**	**45.5**
paraphrase-multilingual-MiniLM-L12-v2	**51.3**	44.4	**43.9**	26.6	27.5	24.9	26.5	24.1	23.5	26.1	30.8	24.8	30.3	**45.1**	34.0

APA	Conventional			Label			Description			Sentences			Hierarchy		
Model	∅P	∅R	∅F1	∅P	∅R	∅F1	∅P	∅R	∅F1	∅P	∅R	∅F1	∅P	∅R	∅F1
paraphrase-multilingual-MiniLM-L12-v2	**83.0**	51.7	**61.3**	14.5	11.7	10.5	31.0	32.8	29.9	26.1	35.2	27.3	33.0	**61.0**	40.4

Observations and Summary

- Time
  - Entailment-based zero-shot classification is too slow for large-scale label sets
  - Similarity-based zero-shot classification runs much faster
  - Not possible to speed up further by Approximate Nearest Neighbor search because of risk of missing labels
- Quality
  - Using descriptions, sentence aggregation with saturation, and hierarchical consistency can enhance pretrained zero-shot classification close to the performance of more elaborate classifiers
    - Clearly better recall, slightly less precision
    - This is true only when using the best-suited pretrained English language models
    - Pretrained multilingual models are less suitable (still slightly better recall but much lower precision)