STRATEGY COMPARISON FOR SEMANTIC ZERO-SHOT TAXONOMY FILTERS

A. Hamm*, German Aerospace Center (DLR), Cologne, Germany

Abstract

This contribution extends and tests ideas from sentence-transformer-based zero-shot text classification to the problem of building a taxonomy filter which can be used for assigning large quantities of documents to user-defined thematic groups.

EXTENDED ABSTRACT

Motivation

In information retrieval, categorised filtering is a way of supporting users in formulating their information needs in an efficient way: Instead of having to provide an explicit specification of the search request, the user can select from an existing list of available categories or tags.

Many collections of documents provided by publishers, libraries, or archives are structured in terms of subject-specific categories that can be used within their domains. Traditionally, assigning documents to categories is an effortful manual process. Progress in machine learning classification algorithms has made it possible to automatize this task in a generally acceptable manner, provided a sufficient number of labelled example documents from all categories is put into the training process.

The latter requirement, however, is a serious obstacle for a flexible use over a broad range of domains and in areas with limited amount of training data available.

It is therefore attractive to explore how the recently proposed method of transformer-based zero-shot text classification [1] can be applied to building taxonomy filters.

Objective

The aim of this contribution is to suggest and compare methods with the following characteristics, which will be explicated below:

1. The method should work with any user-provided commented taxonomic category system.
2. The method should not require taxonomy-specific training.
3. Time-consuming pre-processing steps for each document should not depend on the individual taxonomy categories.

A taxonomy for documents is a hierarchical tree-like system of categories (groups of documents) which covers a domain of interest. Taxonomies are abundant in scientific, economic, and normative classification schemes. Formalised, they are part of the W3C-recommended Simple Knowledge Organization System (SKOS). A commented taxonomy adds a short description to each taxonomy label.

If users want to set up specialised taxonomies for their purposes, it is much easier to provide a short description than to collect a sufficiently large set of labelled examples. General purpose language models which are able to detect semantic similarity can then be used to match documents to taxonomic descriptions.

However, if this matching for N documents and M categories requires the encoding of NM text sequence pairs it becomes inefficient. For large scale applications it is therefore important to employ the bi-encoder strategy of sentence transformers [2] which needs just $N+M$ encodings – individually on documents and on category descriptions.

Method

A simple baseline implementation of a taxonomy filter obeying the first two objectives can be easily realised by directly using the *zero-shot classification pipeline* of *Hugging Face* [3]. The third objective can be achieved by confining to its *sentence similarity pipeline*.

This contribution reports on ongoing work regarding the refinement of taxonomy filters through category assignments which go beyond a gross similarity score by

a) Breaking down the documents into single sentences and computing a weighted aggregated category vote,
b) Taking into account hierarchical consistency as a criterion for category assignment.

Various variants of such taxonomy filters will be tested and compared on the basis of datasets and taxonomies from different domains.

These examples also show the degree to which approximate nearest neighbourhood computations in the underlying embedding space can replace exact calculations as a further performance improvement for large scale applications.

REFERENCES

[1] W. Yin, J. Hay, and D. Roth. "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3914–3923. doi:10.18653/v1/D19-1404

[2] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410

[3] https://github.com/huggingface/transformers

* Email: andreas.hamm@dlr.de

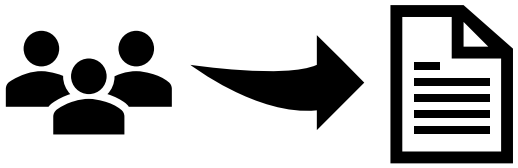
STRATEGY COMPARISON FOR SEMANTIC ZERO-SHOT TAXONOMY FILTERS

OSSYM 2022 – 4th INTERNATIONAL OPEN SEARCH SYMPOSIUM

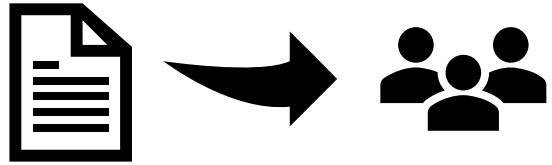
Andreas Hamm
DLR Institute for Software Technology

Searching vs Filtering

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

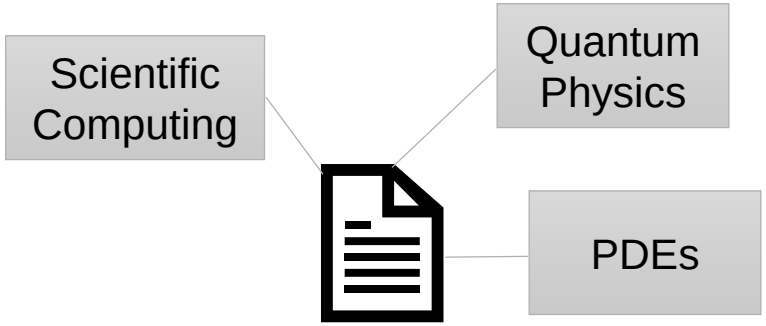


- Filtering 
  - Categories / meta data / tags made available by information service provider
  - Users scan categories
  - Documents find users



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)