

TraininG towards a society of data-saVvy inforMation
prOfessionals to enable open leadership INnovation



Using Titles vs. Full-text as Source for Automated Semantic Document Annotation

Lukas Galke Florian Mai Alan Schelten Dennis Brunsch
Ansgar Scherp

Kiel University, Germany

Knowledge Capture, December 4th-6th, 2017, Austin TX

www.moving-project.eu

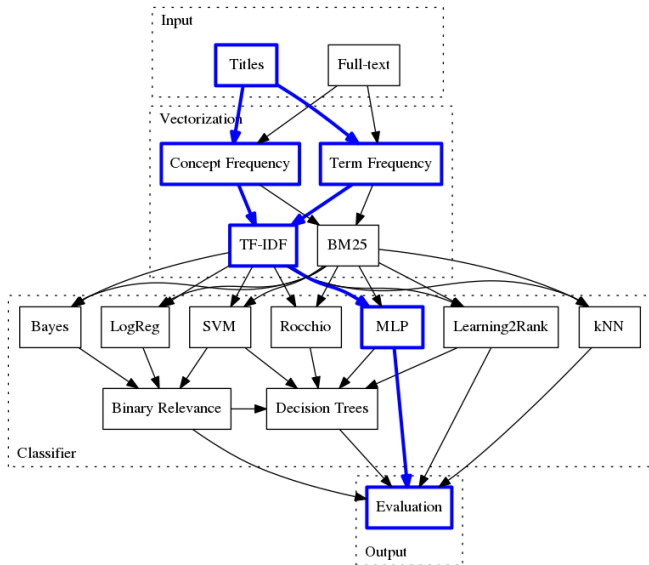
Consider semantic annotation as **Multi-Label Classification** task

- ▶ Full-text not freely available (copyright issues)
- ▶ Full-text not easily accessible (technical issues)
- ▶ Many studies on classification of either short or full-text, but no systematic comparison

Main Research Question

To what extent can automated semantic document annotation be performed solely based on the titles of documents?

- ▶ Multi-label k-nearest neighbors (kNN) (Zhang and Zhou 2007)
- ▶ Generalized linear models optimized by SGD (Bottou 2010)
- ▶ kNN + Learning2Rank (Huang, Névél, and Lu 2011)
- ▶ Revisiting Neural Networks (Nam et al. 2014)
- ▶ Concept extraction + kNN (Große-Bölting, Nishioka, and Scherp 2015)



Three-Step Procedure

1. Comparison of text vectorization methods
2. Comparison of text classification methods
3. Compare the results of titles vs full-text

Evaluation

- ▶ Sample-based F_1 -score. Compute precision and recall for each document individually, then average over the documents.
- ▶ 10-fold cross validation
- ▶ Fixed method-specific hyper-parameters across all experiments

Basic Dataset Statistics

	Econ.	Polit.	RCV1	NYT
Documents	63k	28k	100k	100k
Classes	4.7k	7.2k	101	6.8k
Labels / Document ₅₀	4	5	14	2
Documents / Class	70.8	32.6	3174.9	37.1
Words / Title	7.07	8.13	12.21	4.46

- ▶ **Preprocessing:** Stop word removal, lowercase, WordNet lemmatization
- ▶ No pruning of rare classes
- ▶ No mercy for out-of-training set classes

F₁-score of different vectorization methods with using kNN classifier

Input	Vectoriz.	Econ.	Polit.	RCV1	NYT
Full-text	TF-IDF	.406	.269	.758	.394
Full-text	BM25	.370	.230	.740	.370
Full-text	CF-IDF	.402	.266	.451	.367
Full-text	BM25C	.296	.161	.423	.236
Full-text	CTF-IDF	.411	.272	.761	.406
Full-text	BM25CT	.377	.231	.742	.379
Titles	TF-IDF	.351	.201	.709	.238
Titles	BM25	.349	.196	.687	.230
Titles	CF-IDF	.303	.183	.275	.105
Titles	BM25C	.304	.172	.193	.073
Titles	CTF-IDF	.368	.212	.717	.242
Titles	BM25CT	.364	.208	.693	.239

F_1 -score of different classifiers with using CTF-IDF features

Input	Classifier	Econ.	Polit.	RCV1	NYT
Full-text	kNN (<i>baseline</i>)	.411	.272	.761	.406
Full-text	SVM	.481	.319	.852	.554
Full-text	Log. Regr.	.485	.322	.851	.556
Full-text	L2R	.431	.328	.727	.435
Full-text	MLP	.519	.373	.857	.569
Titles	kNN	.368	.212	.717	.242
Titles	SVM	.426	.272	.804	.325
Titles	Log. Regr.	.429	.274	.803	.326
Titles	L2R	.419	.296	.699	.296
Titles	MLP	.472	.309	.812	.332

⋮

Retained F_1 -score of MLP with respect to full-text

	Econ.	Polit.	RCV1	NYT
Δ	.047	.064	.045	.237
%	91%	83%	95%	58%

Remark: Mean title length of NYT articles is 4.46 words

Assumption: Lower bound on available information given by number of words

Main Result






- ▶ On average: > 7 words $\rightarrow > 80\%$ retained F_1 -score
- ▶ Apart from NYT, 90% F_1 -score could be retained.

Side Results

- ▶ Concatenation of textual features and extracted concepts is preferable over one of them alone.
- ▶ MLP is capable of outperforming previously prominent kNN-based approaches on multi-label classification.

Code available on github.com/Quadflor/quadflor
Technical report on arxiv.org/abs/1705.05311

MOVING is funded by the EU Horizon 2020 Programme under the project number INSO-4-2015: 693092

-  Bottou, Léon (2010). “Large-scale machine learning with stochastic gradient descent”. In: *Proceedings of COMPSTAT'2010*. Springer.
-  Große-Bölting, Gregor, Chifumi Nishioka, and Ansgar Scherp (2015). “A comparison of different strategies for automated semantic document annotation”. In: *Knowledge Capture*. ACM.
-  Huang, Minlie, Aurélie Névéol, and Zhiyong Lu (2011). “Recommending MeSH terms for annotating biomedical articles”. In: *Am. Medical Informatics Association* 18.5.
-  Nam, Jinseok et al. (2014). “Large-scale Multi-label Text Classification. Revisiting Neural Networks”. In: *Machine Learning and Knowledge Discovery in Databases*. Springer.
-  Zhang, Min-Ling and Zhi-Hua Zhou (2007). “ML-KNN: A lazy learning approach to multi-label learning”. In: *Pattern recognition* 40.7.

Binary Relevance

Train one binary classifier for each class.

Thresholds

Impose fixed thresholds on ranking-based or probabilistic classifiers.

Learned Thresholds

Learn the thresholds from training data. Regression from training examples to thresholds.

Decision Tree Module

Learn a decision tree for each class on top of base classifier.

Inverse Document Frequency (IDF)

Re-weight term frequencies (t, d) by $\log \frac{|D|}{|\{d \in D | t \in d\}|}$. Discount terms that appear in many documents.

MLP Hyper-Parameters

One hidden layer of size 1000, Adam optimizer with learning rate 10^{-3} , ReLU activation (tanh for NYT), dropout $p = .5$
Fix threshold for prediction: 0.2