Hierarchical Classification for Economic News Articles

Fernando Benites Mark Cieliebak

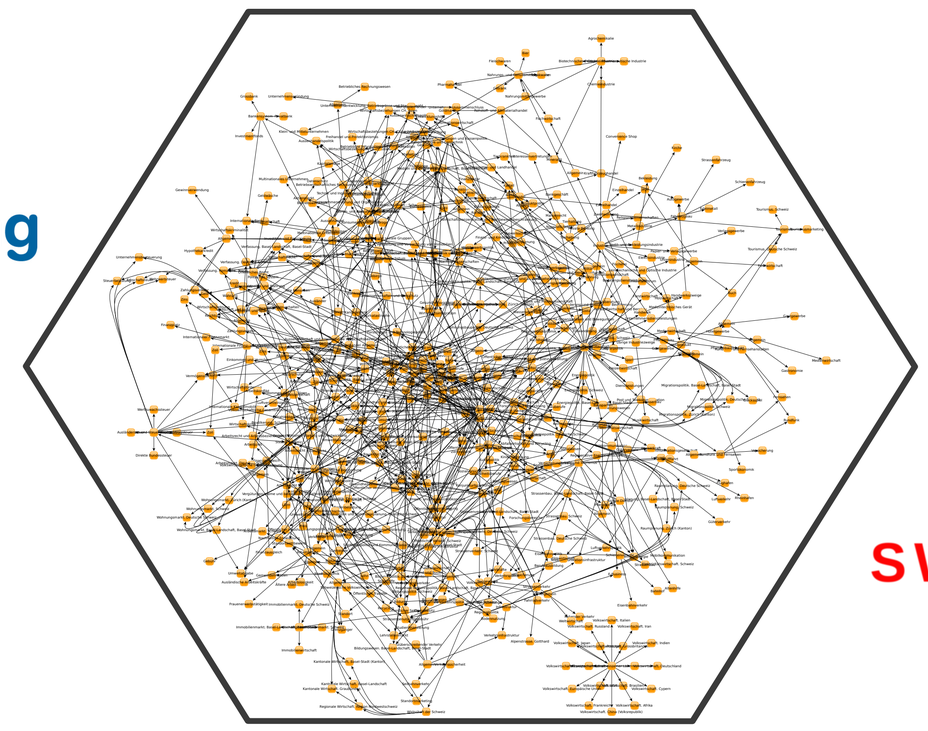
Zurich University of Applied Sciences

School of Engineering

University of Basel

SPINNINGBYTES

SWA SCHWEIZERISCHES WIRTSCHAFTSARCHIV



Motivation

Organizing **large collections** of documents is a challenge known to every large company or institution. Information scientists organize these documents in **ontologies** that are **intuitive and easy** to navigate. However, with greater amount of documents to classify and labels to assign to, the task is not trivial. **Subjectivity, experience, and state of mind** are variables that humans are facing when performing such tasks and consequently may influence the **quality of document collections** and search results.

The "**Swiss Economic Archive**" has a large number of **news articles** (since 1910 approx. 2.5M) which it embeds into a **large hierarchy** based on economic topics such as Insurance, Economic Sanctions, or Advertising. In a joint research project, **ZHAW** and **SpinningBytes AG** implemented a system to support and partially automatize this process.

Multi-label Classification Specs:

Data

Assigned Classes subset of total (multi-label):

# classes:	∅ classes assigned
- 2500 total	- 9.96 total
- 748 in dataset	- 1.7 at level 4

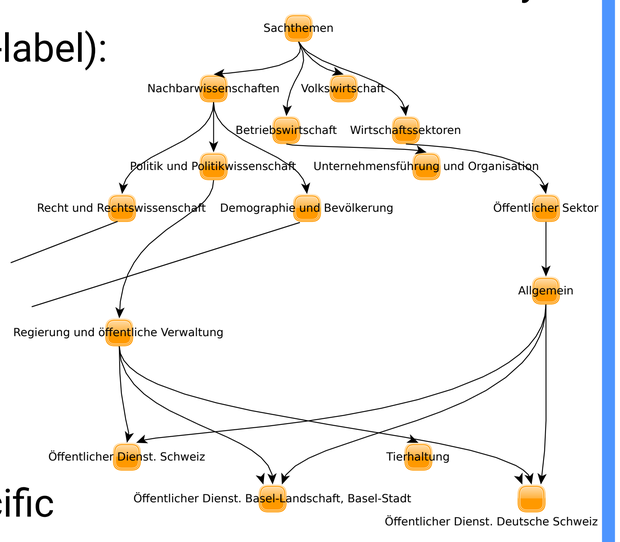
A Hierarchy is also available (4 levels):

- 4, 61, 200, 483 nodes per level
- Level 4 can have multiple
- Level 1-3 have one parent
- Only level 4 is manually assigned
- Topics very close and often geo specific

Number of documents:

- 31,664 with 58,328,594 words from which 463,680 are unique

Slice of the hierarchy



Algorithms

Hierarchical

- Parent-wise: each parent decides if their children are relevant
- Per level parallelizable
- Parents needs only data relevant to their children
- Preprocessing and parameters optimized per parent

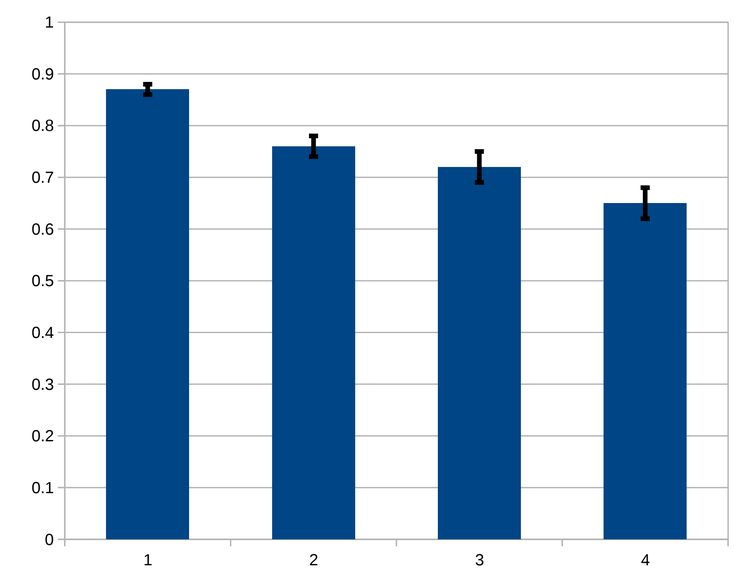
Binary Relevance (BR)

- 1 vs All: only level 4 is evaluated (no hierarchy information)
- Highly parallelizable
- Requires lots of memory
- Categories with many documents tend to score better
- Categories have large amounts of negative samples

Experiments

10-Fold CV

		Micro F-1	One-Error
Level 4	Hierarchical	0.65 ± 0.03	0.92 ± 0.02
	BR	0.46 ± 0.03	0.62 ± 0.04
All levels	Hierarchical	0.75 ± 0.02	-



Micro F-1 per Level for Hierarchical

Qualitative Analysis on 110 samples

27 perfect scores vs only 11 wrong assignments
35 missing assignments

Surprising helpful Assingments (exemplary):

	TRUE	Additional	Missing
Twint erobert die Region - Basler Kantonalbank und Bank Coop bieten Bezahl-App an	Payment Transactions	Mobile Communication	
Zwei Jahre im Armeedienst. Israelinnen bei Training in der Negev-Wüste. Sollen Schweizerinnen Armeedienst leisten können? Ja. Sollen sie es müssen? Nein.	Military	Women Politics	
RECHT Bei Online-Hypotheken	Mortgage Credit	Bank	Electronic Banking

This research has been funded by the Commission for Technology and Innovation (CTI) project no. 18832.1 PFES-ES and by SpinningBytes AG, Switzerland.

Contact Fernando Benites benf@zhaw.ch
Mark Cieliebak mc@spinningbytes.com