Comparing Features for Ranking Relationships Between Financial Entities based on Text

Tim Repke, Michael Loster, Ralf Krestel

FEIII Challenge 2017

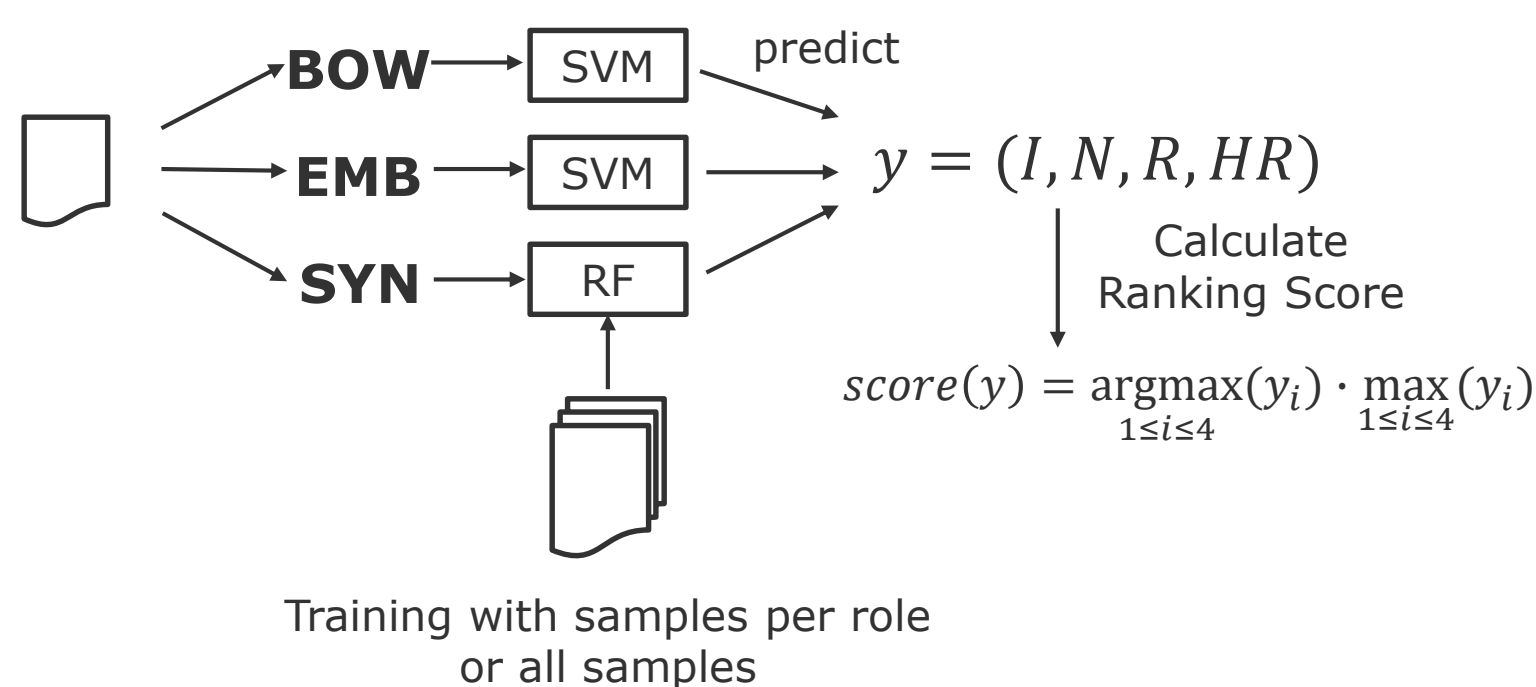Financial Entity Identification and Information Integration Challenge 2017

Given:
- set of tuples (financial entity A, role, financial entity B)
- text snippet the tuple was extracted from (10-K, 10-Q filings)
- tuples labelled by experts as (highly) relevant, neutral, irrelevant

Challenge:
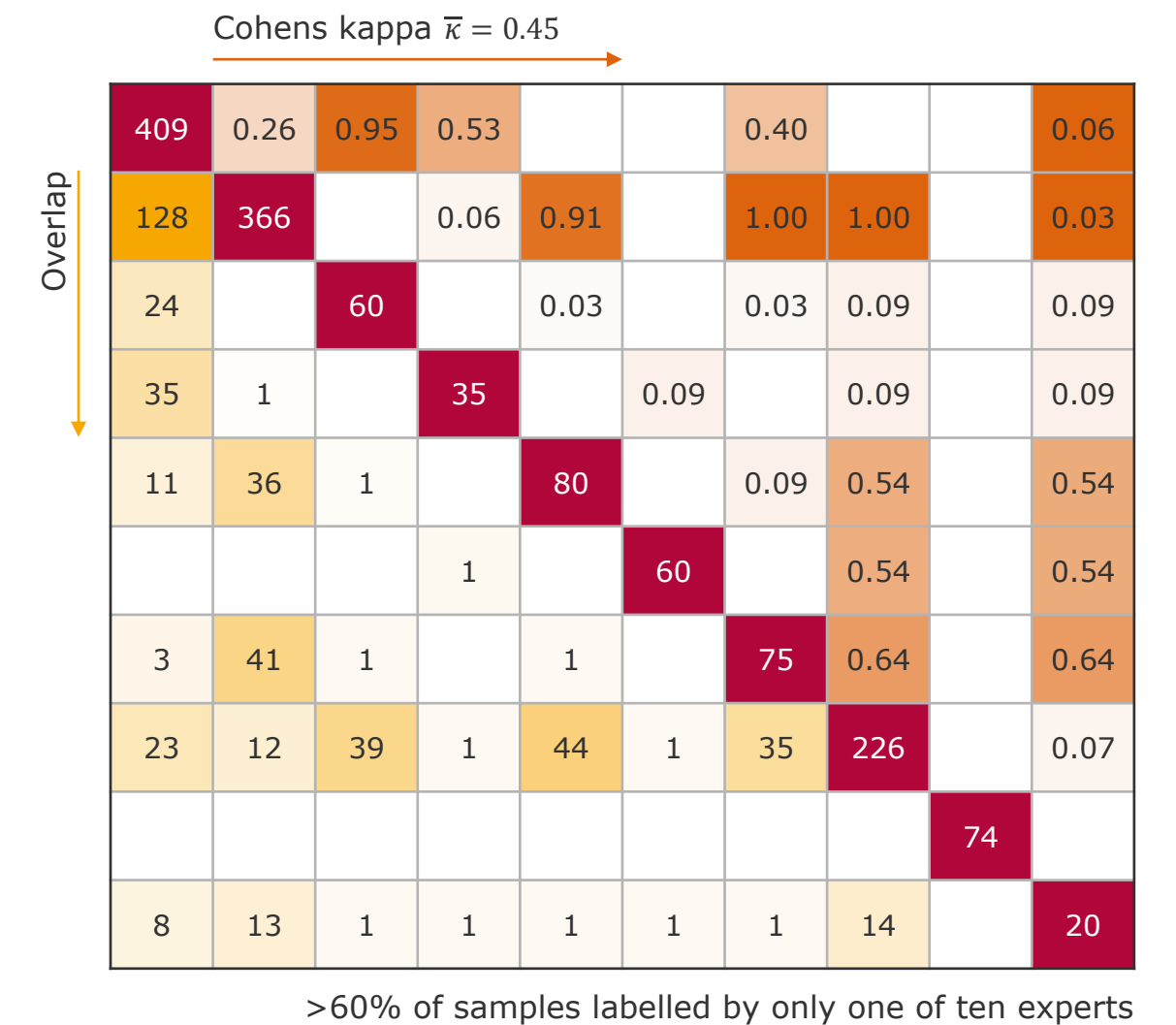- per role, rank relationship tuples by relevance

Our Approach

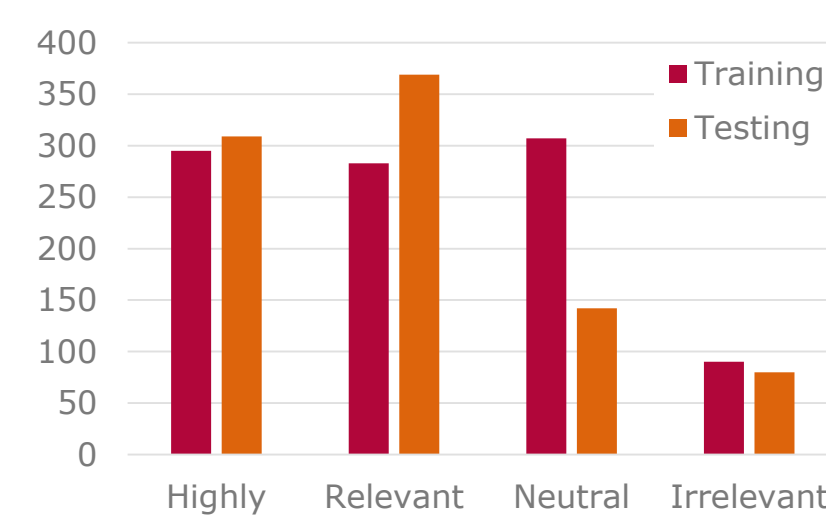$$y = (I, N, R, HR)$$

Calculate Ranking Score

$$score(y) = \underset{1 \le i \le 4}{\operatorname{argmax}}(y_i) \cdot \underset{1 \le i \le 4}{\max}(y_i)$$

Training with samples per role or all samples
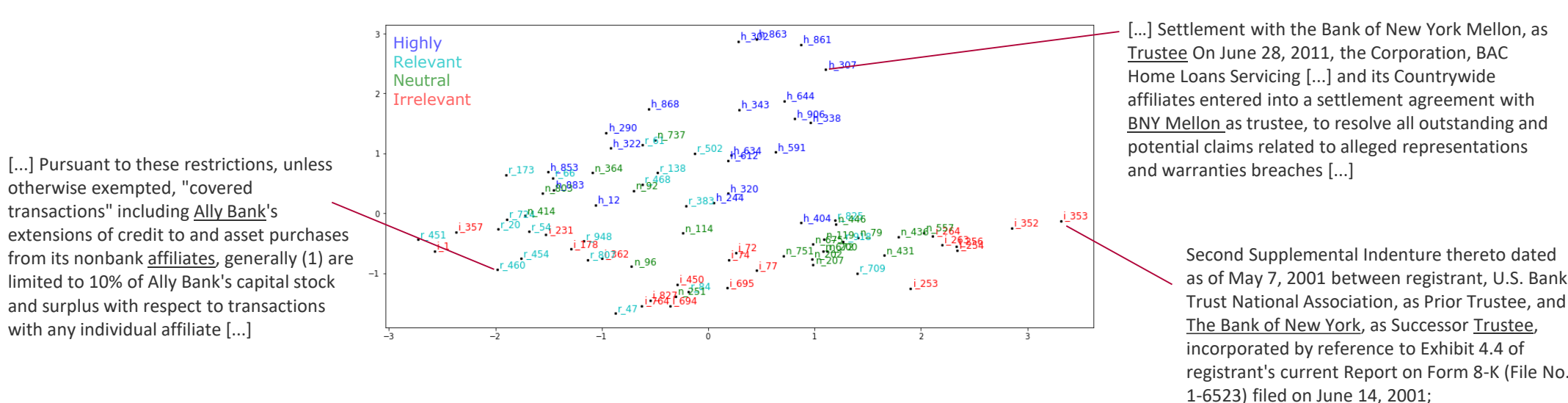
Features

Bag-of-Words (BOW)

Utilise wording that correlates with relevance
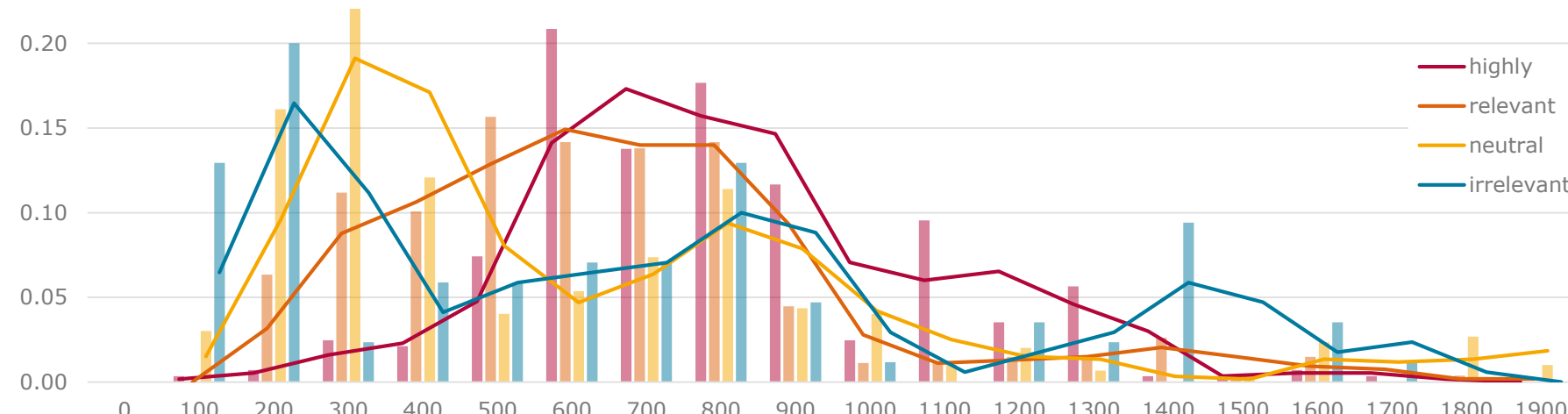

Paragraph Embeddings (EMB) [Mikolov, 2013]

Dense vector representations of sentences with embeddings trained on full-length filings

[...] Pursuant to these restrictions, unless otherwise exempted, "covered transactions" including Ally Bank's extensions of credit to and asset purchases from its nonbank affiliates, generally (1) are limited to 10% of Ally Bank's capital stock and surplus with respect to transactions with any individual affiliate [...]

[...] Settlement with the Bank of New York Mellon, as Trustee On June 28, 2011, the Corporation, BAC Home Loans Servicing [...] and its Countrywide affiliates entered into a settlement agreement with BNY Mellon as trustee, to resolve all outstanding and potential claims related to alleged representations and warranties breaches [...]

Second Supplemental Indenture thereto dated as of May 7, 2001 between registrant, U.S. Bank Trust National Association, as Prior Trustee, and The Bank of New York, as Successor Trustee, incorporated by reference to Exhibit 4.4 of registrant's current Report on Form 8-K (File No. 1-6523) filed on June 14, 2001;

Syntax Features (SYN)

Manually derived features independent of wording on character, token, and POS-tag level

Example: Histogram of snippet length (characters)

The Dataset

Role Distribution

Role-Keyword	Training	Testing
Affiliate	186	129
Agent	61	40
Counterpart	64	108
Guarantor	34	28
Insurer	19	47
Issuer	129	98
Seller	20	49
Servicer	21	57
Trustee	420	304
Underwriter	21	40
All	**975**	**900**

Inter Annotator Agreement

Cohens kappa $\bar{\kappa} = 0.45$

>60% of samples labelled by only one of ten experts

Relevance Distribution

Implications on NDCG

- Baseline (**random**): average NDCG of 100 random sorts
- Baseline (**worst**): NDCG of inverse perfect order

ideal	4	4	3	3	3	2	2	1
worst	1	2	2	3	3	3	4	4

Results

Approach	NDCG (std)	F1-Score
Random	0.88 (0.03)	--
Worst	0.72 (0.06)	--
BOW	0.88 (0.05)	0.34 (0.13)
EMB	0.89 (0.04)	0.24 (0.18)
SYN	0.94 (0.04)	**0.44** (0.11)
Vote	**0.95** (0.04)	0.43 (0.12)

- **Vote** (BOW+EMB+SYN): soft vote of BOW, EMB, and SYN
- 5-fold cross-validation
- Trained on ~900 training samples, leaving out samples from 5 documents
- Tested on all testing samples, evaluation per role and aggregated

Confusion Matrix

- Good performance for highly relevant samples
- Predictions biased towards relevant
- Weighting during training doesn't resolve bias
- Classification performance only secondary, using predictions to calculate ranking score partly resolves bias

annotated \ predicted	Highly	Relevant	Neutral	Irrelevant
Highly	0.73	0.23	0.04	0.00
Relevant	0.25	0.39	0.34	0.02
Neutral	0.21	0.56	0.17	0.06
Irrelevant	0.09	0.54	0.21	0.16

Challenge Results

Name	gt1	gt2	gt3	gt4	gt5	SUM	Ranking
P1	16 (0.66)	16 (0.84)	16 (0.52)	17 (0.8)	15 (0.71)	102	17
P2	13 (0.69)	12 (0.87)	14 (0.54)	13 (0.81)	14 (0.71)	74	12
P3 scored_full_all	8 (0.74)	8 (0.89)	7 (0.6)	8 (0.84)	9 (0.74)	39	6
P4 scored_full_bow	15 (0.67)	15 (0.86)	15 (0.54)	12 (0.82)	5 (0.74)	85	16
P5 scored_full_emb	10 (0.72)	11 (0.87)	12 (0.56)	11 (0.82)	16 (0.7)	74	12
P6 scored_full_syn	**3 (0.82)**	**2 (0.92)**	**4 (0.65)**	**2 (0.9)**	**3 (0.76)**	**16**	**2**
P7 scored_role_all	12 (0.69)	13 (0.86)	11 (0.56)	16 (0.81)	13 (0.72)	80	15
P8 scored_role_bow	7 (0.78)	7 (0.9)	2 (0.65)	7 (0.86)	4 (0.75)	40	7
P9 scored_role_emb	11 (0.72)	10 (0.87)	10 (0.57)	14 (0.81)	17 (0.7)	76	14
P10 scored_role_syn	6 (0.79)	6 (0.9)	1 (0.65)	4 (0.88)	2 (0.78)	33	4
P11	17 (0.62)	17 (0.84)	17 (0.51)	10 (0.82)	1 (0.78)	72	11
P12	4 (0.81)	4 (0.92)	6 (0.61)	5 (0.88)	12 (0.72)	45	8
P13	14 (0.68)	14 (0.86)	13 (0.55)	15 (0.81)	11 (0.73)	68	10
P14	9 (0.74)	9 (0.88)	8 (0.6)	9 (0.84)	8 (0.74)	49	9
P15	2 (0.83)	3 (0.92)	3 (0.65)	3 (0.89)	7 (0.74)	19	3
P16	5 (0.79)	5 (0.91)	9 (0.6)	6 (0.87)	10 (0.73)	33	4
P17	1 (0.85)	1 (0.93)	5 (0.64)	1 (0.9)	6 (0.74)	9	1

gt1: H+V=4; H=3; R+V,N, I = 0.
gt2: H+V=4; H=3, R+V=2; R=2; N=1; I = 0.
gt3: H+V=4; H, R, R+V, N, I = 0.
gt4: H+V=4; H=3; R+V=3.5; R, N, I = 0.
gt5: H+V=4; H=3; R+V=3.5; H, R, N, I = 0.

Contact

Tim Repke (tim.repke@hpi.de)
Web: https://hpi.de/naumann
Rudolf-Breitscheid-Str. 187
14482 Potsdam, Germany

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jerey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In Conference on Neural Information Processing Systems 2013.

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam