Panel Discussion: How do we validate patent metrics derived from semantic analysis

Sam Arts
Department of Management, Strategy and Innovation
Faculty of Economics and Business
KU Leuven
sam.arts@kuleuven.be

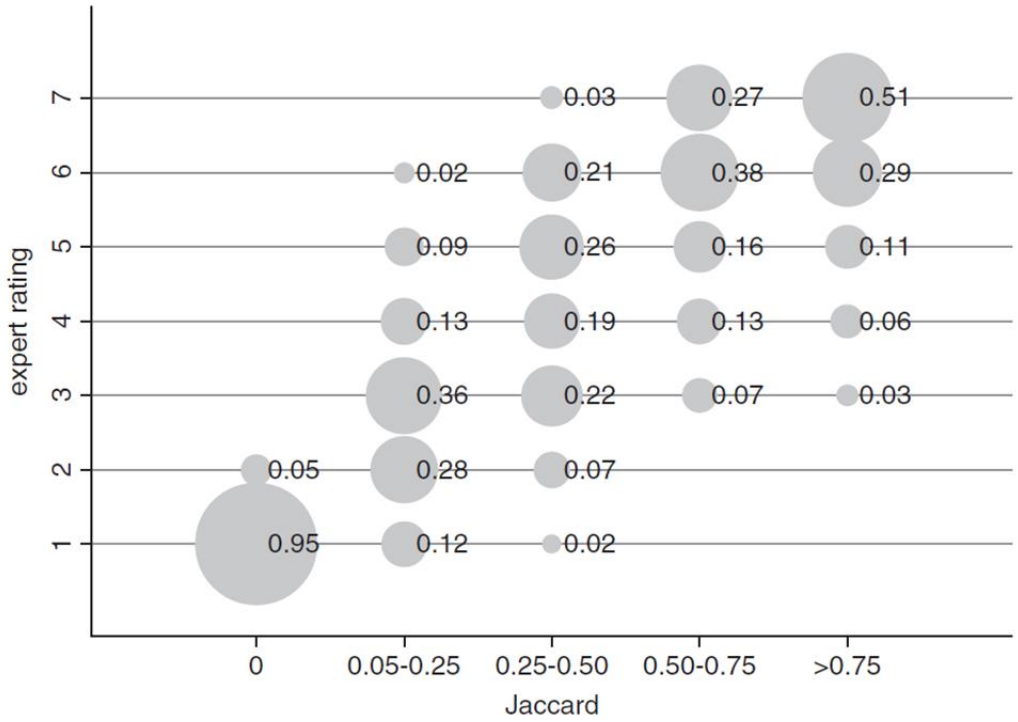
NBER I3 TWG meeting, December 3, 2022

My own experience

- Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1), 62-84.
  - Data: https://dataverse.harvard.edu/dataverse/patenttext
  - Code: https://github.com/sam-arts/smj_code
  - Objectives:
    1. Develop and validate text-based metric to measure similarity between patents
    2. Validate if text is better than patent classification to measure similarity

Validate use of text to measure similarity between patents

- Internal validity: expert ratings
  - Recruit 13 paid experts from 5 fields
  - For each field, randomly select baseline patents and 5 patents with varying degrees of similarity to baseline patent
  - Ask experts to rate similarity of patent pairs (Likert scale 1 to 7), 850 ratings
  - Consistency between ratings: inter-item correlation, Cronbach's alpha



Validate use of text to measure similarity between patents

- External validity

TABLE 3 Summary statistics for subsamples of text-matched patent pairs with varying degrees of Jaccard similarity

	(1) Jaccard = 0 Mean	(2) Jaccard ≥ 0.05 and < 0.25 Mean	(3) Jaccard ≥ 0.25 and <0.50 Mean	(4) Jaccard ≥ 0.50 and <0.75 Mean	(5) Jaccard ≥ 0.75 Mean
Jaccard	0.000	0.164	0.322	0.609	0.928
Binary: Same patent family	0.000	0.002	0.024	0.086	0.407
Binary: Same inventor(s)	0.000	0.083	0.496	0.867	0.927
Binary: Same assignee(s)	0.001	0.148	0.651	0.866	0.865
Binary: Citation link	0.000	0.008	0.044	0.085	0.085
N	4,386,405	3,426,228	601,947	137,551	220,679

Note. 4,386,405 text-matched patent pairs for patents granted between 1976 and 2013. Each baseline patent is matched to the patent with the highest Jaccard index based on keywords and filed in the same year. In cases where there are multiple matches, patents are matched on approximate filing date. Patents with less than 10 keywords are excluded and a minimum Jaccard of 0.05 is imposed. Column 1 includes an additional set of 4,386,405 patent pairs with no overlap in keywords, i.e., Jaccard index of zero, filed in the same year, and matched on approximate filing date. Unreported *t* tests indicate significant differences in same patent family, same inventor(s), same assignee(s), and citation link across the five different subsamples (columns 1–5). All differences are significant at the 1% level. The only nonsignificant difference is in same assignee(s) and citation link for pairs with Jaccard ≥ 0.50 and <0.75 (column 4) versus pairs with Jaccard ≥ 0.75 (column 5).

Patents with higher text similarity more like to belong to same patent family (docdb), inventor(s), assignee(s), and are more likely to cite each other

My own experience

- Arts, S., Hou, J., & Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2), 104144.
  - Data: https://zenodo.org/record/3515985#.Y2oWf3bMIuUn
  - Code: https://github.com/sam-arts/respol_patents_code
  - Objectives:
    1. NLP to develop new measures for novelty and impact based on patent text
    2. Validate metrics and improvement over measures based on patent classification and citations
    3. Provide open access to code and data

Validation

- Patents that cover new technologies with major impact on technological progress
  - Patents linked to prizes (Nobel prize, …)
  - 393 patents
  - Control patents (matched on technical content and year)
  - Can metrics predict award patents: precision, recall, area under curve

- Patents lack novelty and little impact on technical progress
  - USPTO is perhaps granting too many weak or invalid patents that fail to meet the novelty requirement (Jaffe and Lerner, 2004; Lemley and Shapiro, 2005)
  - Granted by all
  - Control patents (granted by USPTO but rejected by EPO and JPO)
  - Can metrics predict grant: precision, recall, area under curve

United States Patent [19]

Goodenough et al.

[11] **Patent Number: 5,910,382**

[45] **Date of Patent: *Jun. 8, 1999**

[54] **CATHODE MATERIALS FOR SECONDARY (RECHARGEABLE) LITHIUM BATTERIES**

[75] Inventors: **John B. Goodenough**, Austin, Tex.; **Akshaya K. Padhi**, LaSalle, Ill.; **K. S. Nanjundaswamy**, Joplin, Mo.; **Christian Masquelier**, Boulogne Billancourt, France

[73] Assignee: **Board of Regents, University of Texas Systems**, Austin, Tex.

[21] Appl. No.: **08/840,523**

[22] Filed: **Apr. 21, 1997**

Related U.S. Application Data

[60] Provisional application No. 60/032,346, Dec. 4, 1996, and provisional application No. 60/016,060, Apr. 23, 1996.

[51] **Int. Cl.**[6] **H01M 4/58**

[52] **U.S. Cl.** **429/218.1**; 429/221; 429/224

[58] **Field of Search** 429/218, 221, 429/224; 29/623.1; 423/554

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,465,747	8/1984	Evans	429/194
4,526,844	7/1985	Yoldas et al.	429/30
4,959,281	9/1990	Nishi et al. .	
4,985,317	1/1991	Adachi et al.	429/191
5,514,490	5/1996	Chen et al.	429/191

FOREIGN PATENT DOCUMENTS

WO 98/12761 3/1998 WIPO .

OTHER PUBLICATIONS

International Search Report dated Aug. 29, 1997.

Delmas and Nadiri, *Mater. Res. Bull.*, 23, 65 (1988). no month.

Goodenough, Hong, Kafalas, *Mater. Res. Bull.* 11, 203 (1976). no month.

Guyomard and Tarascon, *J. Electrochem. Soc.*, 139, 937 (1992). Apr.

Long, Longworth, Battle, Cheetham, Thundathil, Beveridge, *Inorg. Chem.*, 18, 624 (1979). no month.

Manthiram and Goodenough, *J. Power Sources*, 26, 403 (1989). no month.

Masquelier, Tabuchi, Ado, Kanno, Kobayashi, Nakamura, Goodenough, *J. Solid State Chem.*, 123, 255 (1996). no month.

Mizushima, Jones, Wiseman, Goodenough, *Mater. Res. Bull.*, 15, 783 (1980). no month.

Nanjundaswamy, Padhi, Goodenough, Okada, Ohtsuka, Arai, Yamaki, *Solid State Ionics*, 92, 1 (1996). no month.

Okada, Nanjundaswamy, Manthiram, Goodenough, *Proc. 36th Power Sources Conf.*, at Cherry Hill, New Jersey (Jun. 6–9, 1994).

Schöllhorn and Payer, *Agnew. Chem. (Int. Ed. Engl.)*, 24, 67 (1985). no month.

Sinha and Murphy, *Solid State Ionics*, 20, 81 (1986). no month.

Thomas David, Goodenough, Groves, *Mater. Res. Bull.*, 20, 1137 (1983).

Thackeray, Johnson, de Piciotto, Bruce, Goodenough, *Mater. Res. Bull.*, 19, 179 (1984). no month.

Thackeray, David, Bruce, Goodenough, *Mater. Res. Bull.*, 18, 461 (1983). no month.

Wang and Hwu, *Chem. of Mater.* 4, 589 (1992).

CA 115:238022, Petit et al. abstract only, 1991 (month N/A).

Primary Examiner—Maria Nuzzolillo

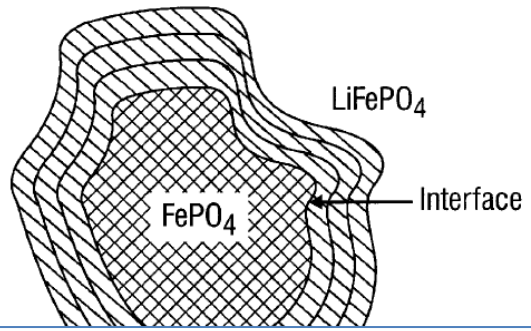
Assistant Examiner—Carol Chaney

Attorney, Agent, or Firm—Arnold White & Durkee

[57] **ABSTRACT**

The invention relates to materials for use as electrodes in an alkali-ion secondary (rechargeable) battery, particularly a lithium-ion battery. The invention provides transition-metal compounds having the ordered-olivine or the rhombohedral NASICON structure and the polyanion $(PO_4)^{3-}$ as at least one constituent for use as electrode material for alkali-ion rechargeable batteries.

9 Claims, 10 Drawing Sheets



- John Goodenough 2018 Benjamin Franklin Medal
- Rechargeable battery
- US5910382
  - “lifepo4” (lithium iron phosphate)
    - reused by 260 patents
  - “batteri lifepo4”
    - reused by 211 patents

Some general remarks

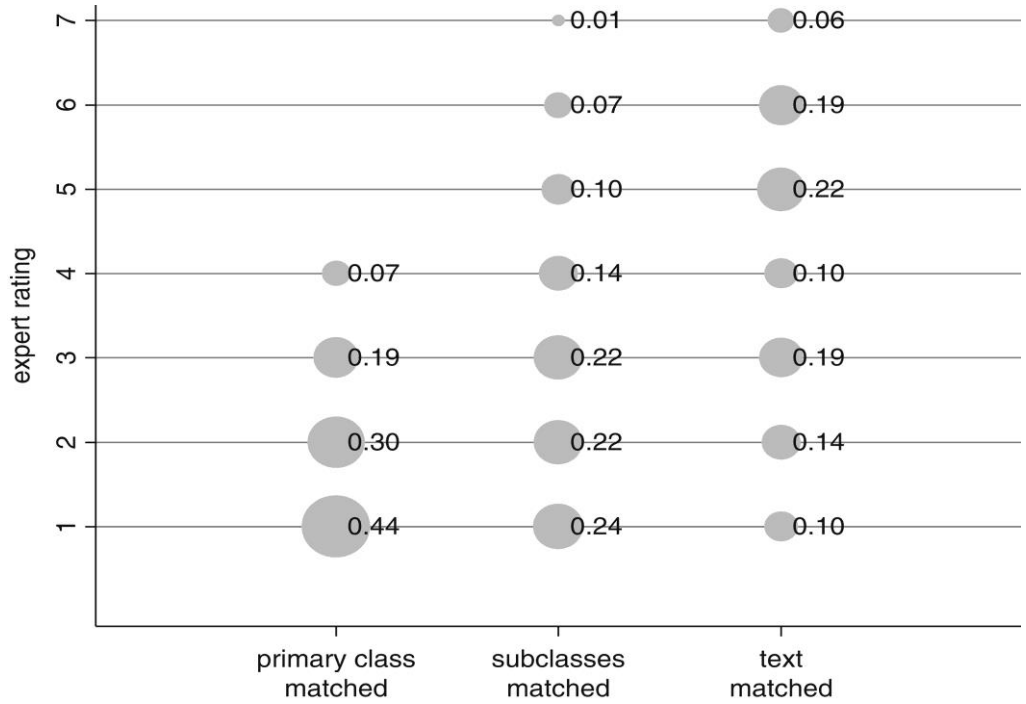
- Preprocessing of text seems to make a big difference
  - Which parts to include (title, abstract, claims, description), stemming, lemmatization, stop words, …
  - Provide open access to raw data and all code so everything can be replicated
- Trade off between simple and more advanced approaches
  - Advanced approaches not necessarily better for every application (different NLP tasks).
  - Simple approaches often work well and are easy to understand and explain to non computer scientists
  - Advanced approaches are often black boxes (what do they measure?)
- Explain why and show that text works better than traditional metrics
- What do you want to measure?
  - Novelty, disruptiveness, originality, private value, social value, impact, importance, diffusion, …
- Little validation, no standardized way to validate the metrics

Thank you!

sam.arts@kuleuven.be

2nd objective: validate if text is better than patent classification to measure similarity

- Internal validity: expert ratings
  - Recruit 13 paid experts from 5 fields
  - For each field, randomly select baseline patents and, for each baseline patent, the corresponding closest text-matched, primary-class matched, and subclasses-matched patent.
  - Ask experts to rate similarity of patent pairs (Likert scale 1 to 7), 300 ratings



2nd objective: validate if text is better than patent classification to measure similarity

- External validity

TABLE 6 Summary statistics for text-matched, primary-class-matched, primary-subclass-matched, and subclasses-matched patent pairs

	(1)	(2)			(3)			(4)		
	Text-matched patent pairs (n = 4,386,405)	**Primary-class-matched patent pairs (n = 4,279,839)**			**Primary-subclass-matched patent pairs (n = 3,492,480)**			**Subclasses-matched patent pairs (n = 4,229,647)**		
	Mean	**Mean**	**|t|**	**Pr(|T| > |t|)**	**Mean**	**|t|**	**Pr(|T| > |t|)**	**Mean**	**|t|**	**Pr(|T| > |t|)**
Jaccard	0.238	0.054	2200.000	0.000	0.092	1800.000	0.000	0.097	1900.000	0.000
Binary: Jaccard = 0	0.000	0.120	−760.000	0.000	0.043	−390.000	0.000	0.040	−420.000	0.000
Binary: Same patent family	0.028	0.005	308.439	0.000	0.012	255.467	0.000	0.013	239.195	0.000
Binary: Same inventor(s)	0.207	0.037	889.674	0.000	0.079	690.830	0.000	0.085	677.012	0.000
Binary: Same assignee(s)	0.276	0.059	992.268	0.000	0.114	752.643	0.000	0.118	743.565	0.000
Binary: Citation link	0.019	0.002	267.639	0.000	0.008	165.982	0.000	0.013	92.972	0.000

Note. t tests assess the mean difference between the text-matched pairs and the primary-class-matched, primary-subclass-matched, and subclasses-matched pairs in columns 2, 3, and 4, respectively. Only the subset of baseline patents for which both a text-matched and a primary-class-matched patent are found are used in the paired t test in column 2. Only the subset of baseline patents for which both a text-matched and a primary-subclass matched patent are found are used in the paired t test in column 3. Only the subset of baseline patents for which both a text-matched and a subclasses-matched patent are found are used in the paired t test in column 4.

Patents matched on text more like to belong to same patent family (docdb), inventor(s), assignee(s), and are more likely to cite each other