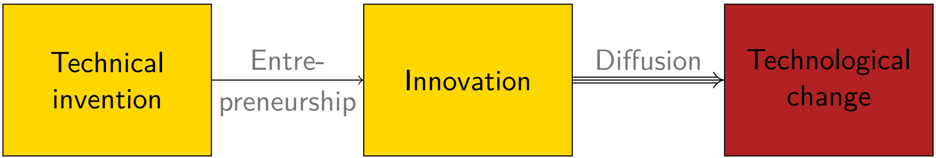
Mapping Technologies to Business Models: An Application to Clean Technologies and Entrepreneurship

published as part of the *the 26th International Conference on Science, Technology and Innovation Indicators (STI2022)* Conference Proceedings

Technological innovation and its measurement



Technological innovation and its measurement

***Patents** have become a surrogate for measuring the innovation process.*

Jaffe (2021)

Technological innovation and its measurement

***Patents** have become a surrogate for measuring the innovation process.*

Jaffe (2021)

***Patent** subclasses provide a [. . .] reliable picture of a firm's technological capabilities*

Aharonson et al. (2016)

Patents and start-ups

A measurement problem

Start-ups barely patent (Graham et al., 2008; Helmers et al., 2011):

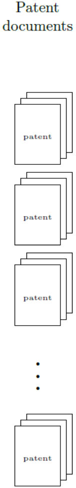
- distracting engineers/managers from key functions (Mann, 2005)
- costs of patenting/patent litigation too high (Graham et al., 2009)
- disclosure through patent allows 'design around' (Mann, 2005)
- patents impact VC decisions as property rights, not as signals of technology quality (Hoenig et al., 2015)

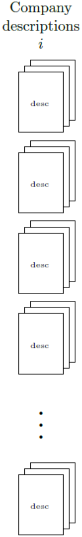
Research question I

How to capture the role of start-ups in the innovation process?

Textual innovation data

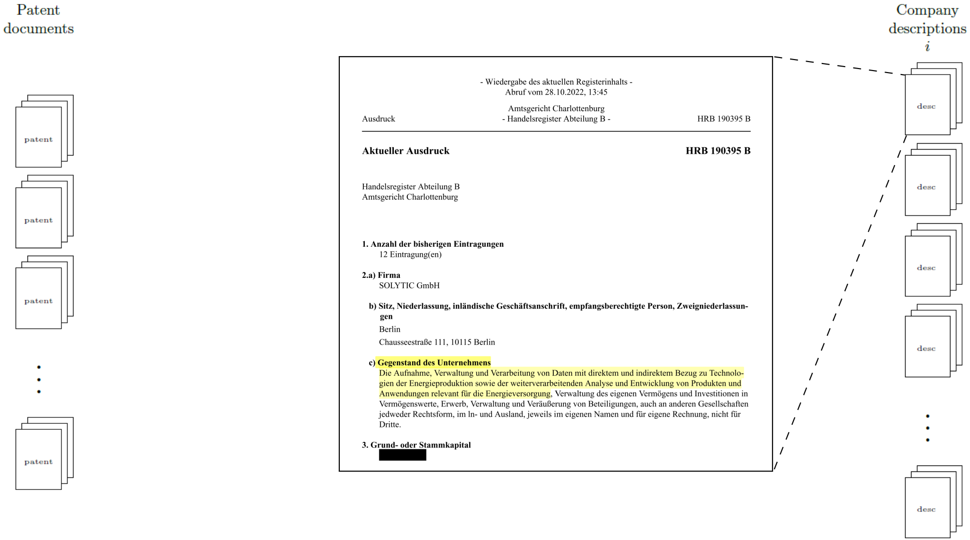
Patents as technology descriptions, business purpose at business registration as technology indication





Textual innovation data

Patents as technology descriptions, business purpose at business registration as technology indication



- Wiedergabe des aktuellen Registerinhalts -
Abruf vom 28.10.2022, 13:45

Ausdruck | Amtsgericht Charlottenburg - Handelsregister Abteilung B - | HRB 190395 B

Aktueller Ausdruck **HRB 190395 B**

Handelsregister Abteilung B
Amtsgericht Charlottenburg

1. Anzahl der bisherigen Eintragungen
12 Eintragung(en)

2.a) Firma
SOLYTIC GmbH

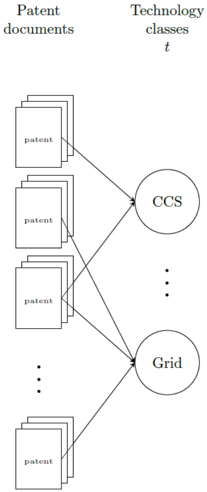
b) Sitz, Niederlassung, inländische Geschäftsanschrift, empfangsberechtigte Person, Zweigniederlassungen
Berlin
Chausseestraße 111, 10115 Berlin

c) Gegenstand des Unternehmens
Die Aufnahme, Verwaltung und Verarbeitung von Daten mit direktem und indirektem Bezug zu Technologien der Energieproduktion sowie der weiterverarbeitenden Analyse und Entwicklung von Produkten und Anwendungen relevant für die Energieversorgung, Verwaltung des eigenen Vermögens und Investitionen in Vermögenswerte, Erwerb, Verwaltung und Veräußerung von Beteiligungen, auch an anderen Gesellschaften jedweder Rechtsform, im In- und Ausland, jeweils im eigenen Namen und für eigene Rechnung, nicht für Dritte.

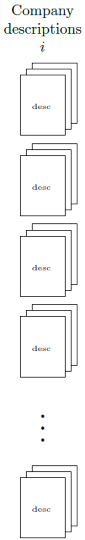
3. Grund- oder Stammkapital
[redacted]

From patents to technology descriptions

Clean technology fields by European Patent Office (EPO)

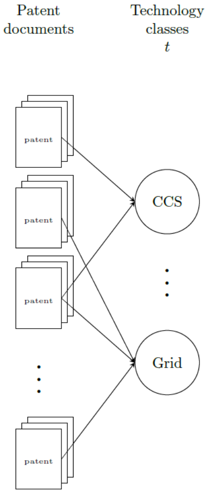


	Clean technology field
Adaption	Technologies for the adaption to climate change
Battery	Battery storage and fuel cell technologies
Biofuels	Biofuel technologies
CCS	Carbon capture, storage and sequestration
E-efficiency	Technologies improving energy efficiency
Generation	Renewable energy generation technologies
Grid	Grid and power conversion technologies
Materials	Low carbon materials and manufacturing
Mobility	Electric vehicles and low carbon mobility solutions
Water	Water and wastewater treatment

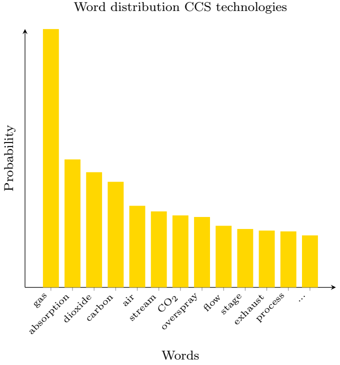


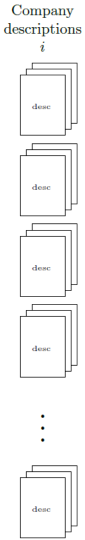
From patents to technology descriptions

L-LDA



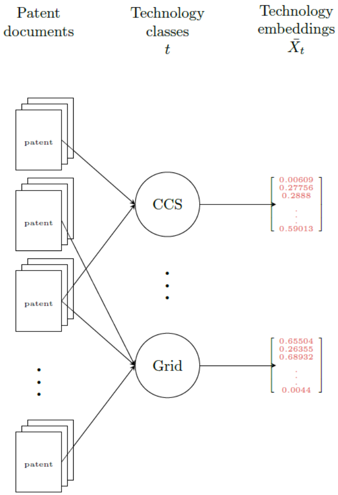
<u>Goal:</u> Derive technology-word distributions from expert-labeled corpus of patent docs





Contextualized vector representations

BERT

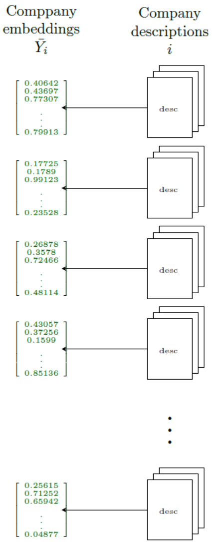


Goal: Derive contextualized vector representations of technology & business descriptions

$$\underset{(1\times Q)}{X_{CCS}} = \langle \text{gas, absorption, dioxide}, \dots, \text{scrub}, \dots \rangle$$

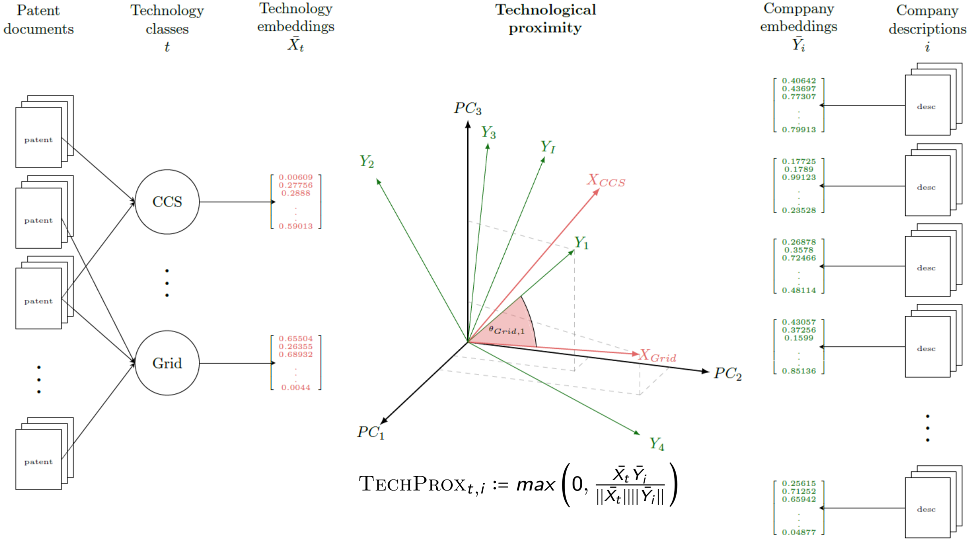
$$\Big\downarrow \text{SBERT}$$

$$\underset{(1\times 384\ \forall Q)}{X_{CCS}} = [0.006, 0.277, 0.288, \dots 0.590]'$$



Mapping framework

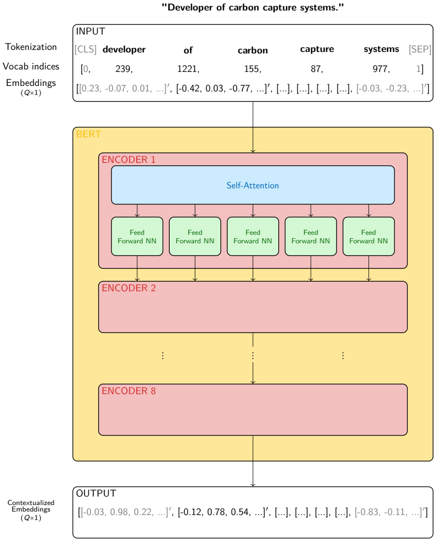
Cosine similarity as measure of a company's technological orientation



$$\text{TECHPROX}_{t,i} := max\left(0, \frac{\bar{X}_t \bar{Y}_i}{||\bar{X}_t||||\bar{Y}_i||}\right)$$

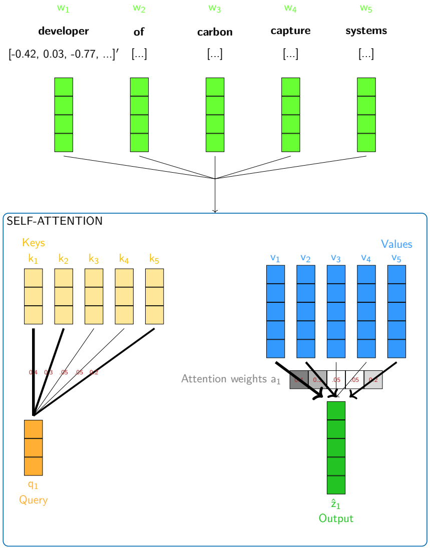
Excursus: BERT

Model architecture



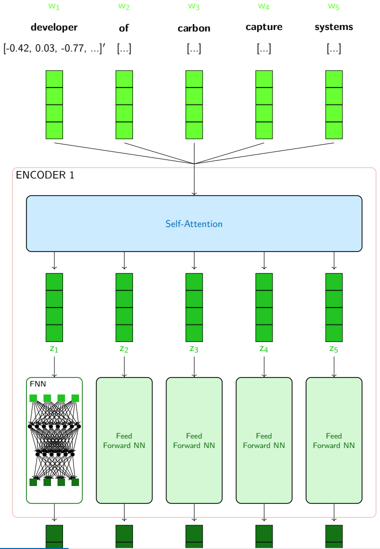
Attention Is All You Need (Vaswani et al., 2017)

Let tokens 'look around' the whole input, and decide how to update its representation based on on what it sees



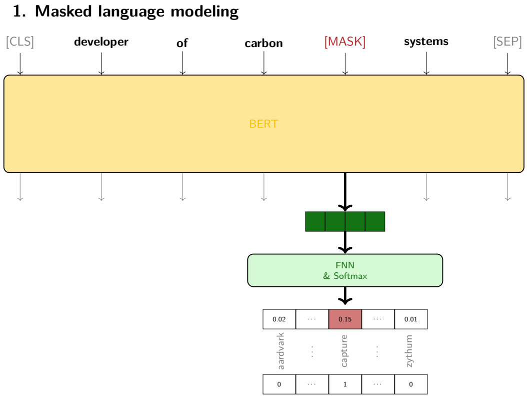
Encoder

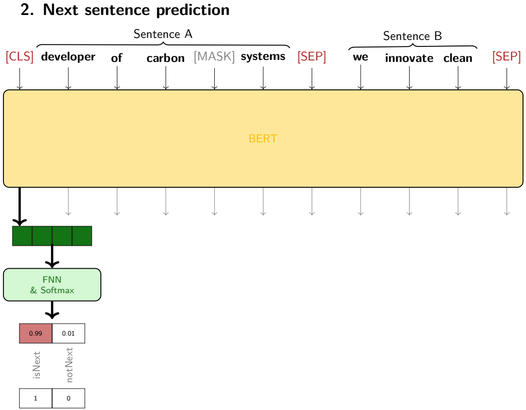
After Attention, each token pondering for itself about what it has observed previously



Training BERT

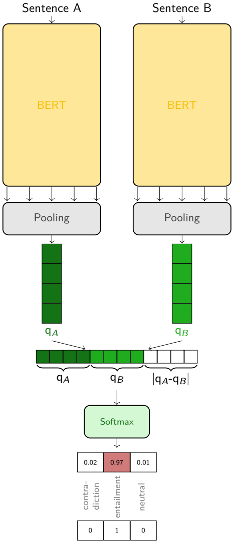
Self-supervised learning based on English Wikipedia





From BERT to SBERT: Contextualized sequence representations

Finetuning based on collection of sentence pairs labeled for entailment, contradiction, and semantic independence



Application

Role of start-ups in clean technology diffusion

Fight against climate change requires new technological pathways and radical innovations (*inter alia* European Commission (2019))

- but: technological path dependencies and system/innovation inertia among incumbents (**Patel1996**; Aghion et al., 2016)
- costly: delay in redirecting innovation towards clean technologies (Benner, 2009; Dijk et al., 2016; Sick et al., 2016)

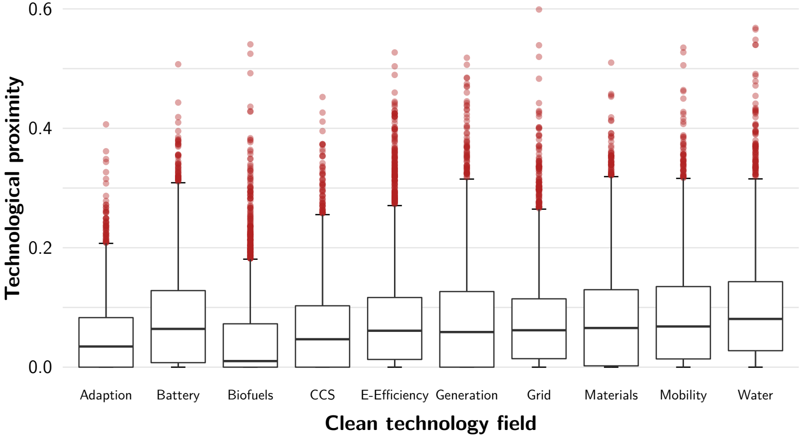
⇒ special role of new (path-independent!) ventures in driving clean technology change

Research question II

Which role do start-ups play in the technological transition to higher levels of decarbonization?

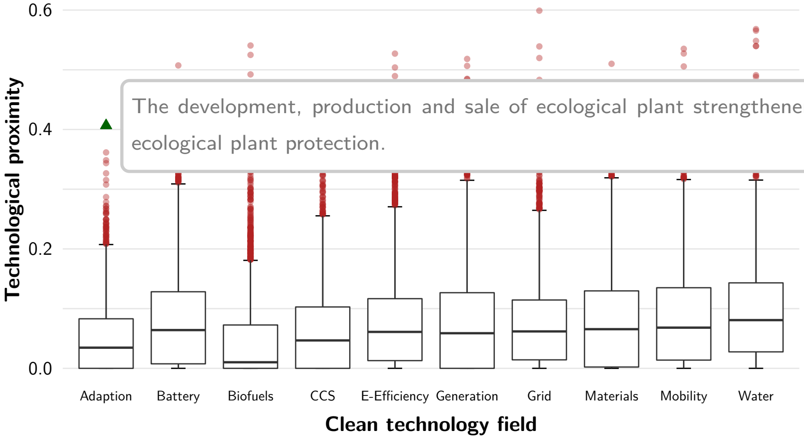
Application

TechProx in survey of German start-ups



Application

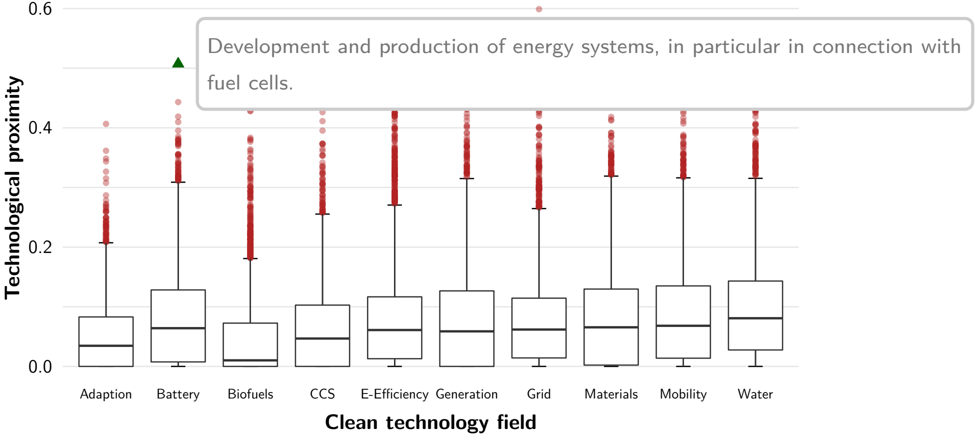
A glance at the 'outliers'



The development, production and sale of ecological plant strengtheners and ecological plant protection.

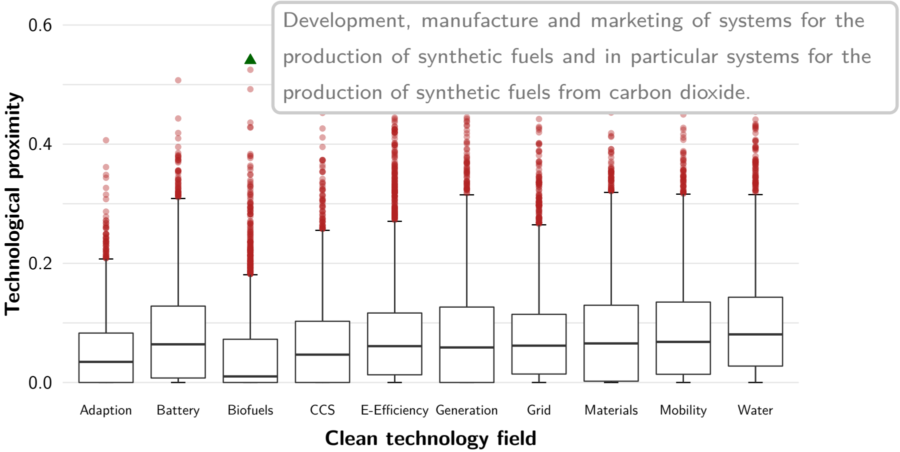
Application

A glance at the 'outliers'



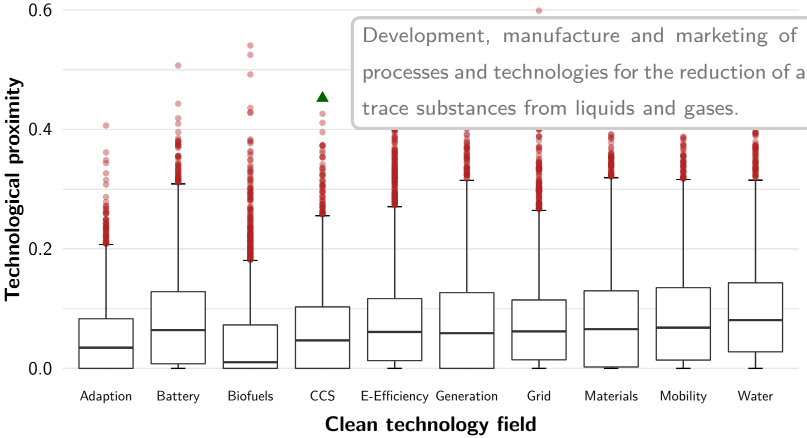
Application

A glance at the ‘outliers’



Application

A glance at the 'outliers'

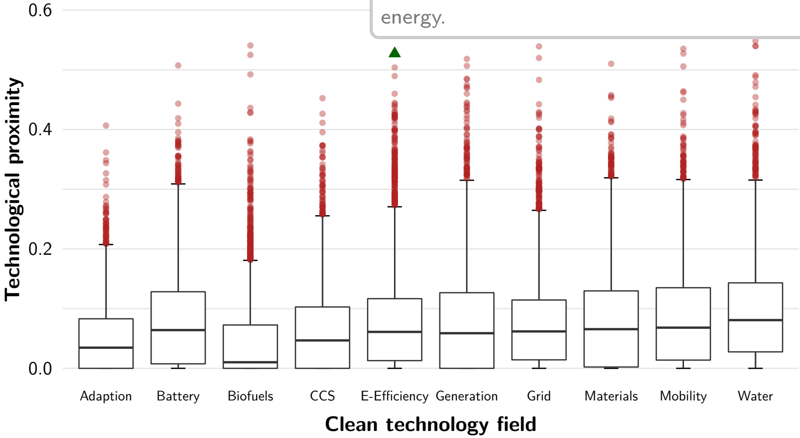


Development, manufacture and marketing of devices, processes and technologies for the reduction of anthropogenic trace substances from liquids and gases.

Application

A glance at the 'outliers'

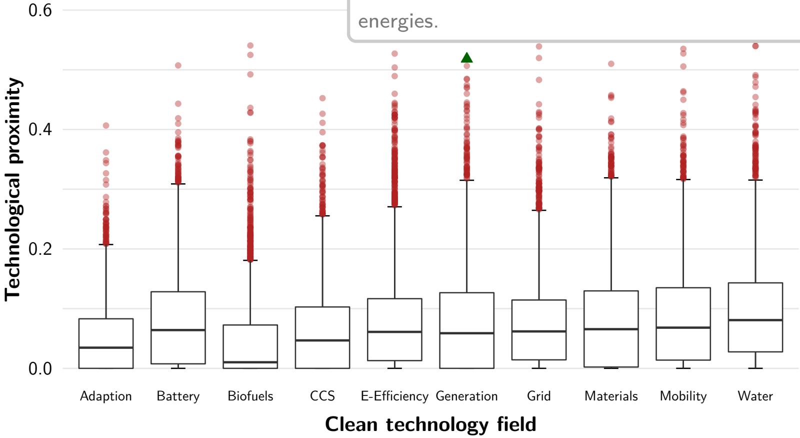
Manufacture, sale, maintenance and repair of heat pumps and other technical equipment, in particular for generating thermal energy.



Application

A glance at the 'outliers'

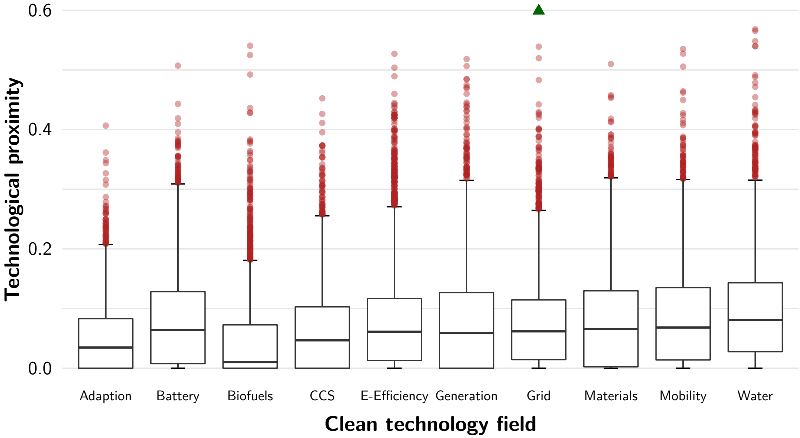
The construction, distribution and sale of turnkey photovoltaic systems and other activities of all kinds in the field of renewable energies.



Application

A glance at the 'outliers'

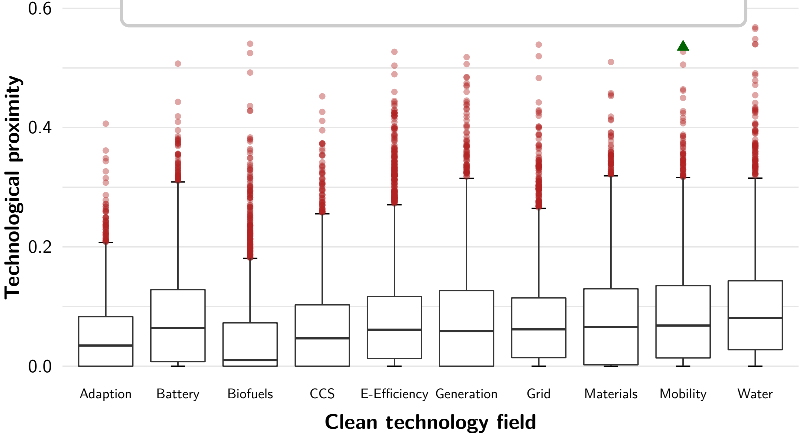
Manufacture of electrode foils, lithium accumulators and energy storage systems and the provision of services in this area.



Application

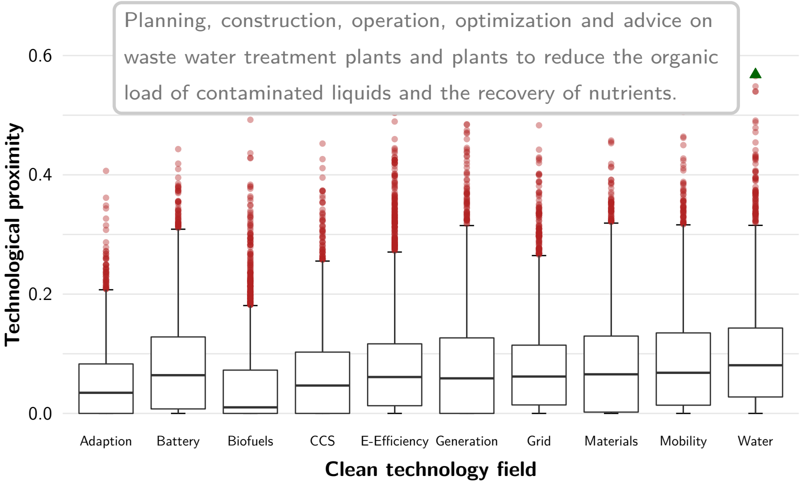
A glance at the ‘outliers’

Engine and software optimization on motor vehicles of all kinds.



Application

A glance at the 'outliers'



Characteristics of clean technology start-ups

Cleantech start-ups show a higher propensity to eco-innovate

	EInno					
	(1)	(2)	(3)	(4)	(5)	(6)
TECHPROX	1.015*	1.014*	1.013*	1.013*	1.012*	1.014***
log(size)		1.190***	1.140***	1.125***	1.186***	1.175***
age		1.001	1.010	1.001	1.005	1.012
subsidy			1.317***	1.353***	1.413***	1.456***
R&D			1.427***	1.434***	1.605***	1.675***
R&D intensity			0.780	0.910	0.904	0.815
returns				1.743***	1.633**	1.551**
break even				1.295***	1.226**	1.237**
team size					0.899**	0.887**
university					0.614***	0.627***
Sector controls	Y	Y	Y	Y	Y	Y
Product type controls	N	N	N	N	N	Y
N	3,269	3,269	3,269	3,192	3,192	2,774
Pseudo R^2	0.022	0.026	0.030	0.033	0.041	0.047

Note: Environmental innovation questions were asked on a Lickert scale with three response possibilities: (1) No environmental innovation; (2) environmental innovation with moderate environmental effect; (3) environmental innovation with substantial environmental effect. Coefficient estimates reported as proportional odds ratios reflecting the factor by which an increase in TECHPROX of one index point (0.01) corresponds to an increase in the odds of having introduced a innovation with at least a moderate environmental effect compared to having introduced no environmental innovation (c.p.). Change in observation numbers due to item non-response. Significance levels: *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$

Summary

- Latest evolutions in the field of NLP allow fine granular determination of a firm's technological profile
- Legal obligation to publish a business purpose makes the technology mapping possible for start-ups even w/o traditional innovation data

- Leveraging the introduced technology mapping to the field of clean technologies suggests:
  - a high propensity of cleantech start-ups to introduce eco-innovations
  - supporting their special role in the transition to a green economy derived from theory
  - both by virtue of their business models as well as a high propensity to adopt additional environmental innovations

Aghion, P., Dechezleprêtre, A., Hémous, D., Martin, R., & van Reenen, J. (2016). Carbon taxes, path dependency, and directed technical change: Evidence from the auto industry. Journal of Political Economy, 124(1), 1–51. https://doi.org/10.1086/684581

Aharonson, B. S., & Schilling, M. A. (2016). Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. Research Policy, 45(1), 81–96. https://doi.org/10.1016/j.respol.2015.08.001

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3, 1137–1155. https://doi.org/10.1080/1536383X.2018.1448388

Benner, M. J. (2009). Dynamic or static capabilities? Process management practices and response to technological change. Journal of Product Innovation Management, 26(5), 473–486. https://doi.org/10.1111/j.1540-5885.2009.00675.x

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022. https://doi.org/10.1145/2133806.2133826

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5arXiv 1607.04606, 135–146. https://doi.org/10.1162/tacl_a_00051

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Dijk, M., Wells, P., & Kemp, R. (2016). Will the momentum of the electric car last? Testing an hypothesis on disruptive innovation. Technological Forecasting and Social Change, 105, 77–88. https://doi.org/10.1016/j.techfore.2016.01.013

European Commission. (2019). The European Green Deal. https://doi.org/10.2307/j.ctvd1c6zh.7

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. http://annabellelukin.edublogs.org/files/2013/08/Firth-JR-1962-A-Synopsis-of-Linguistic-Theory-wfihi5.pdf

Graham, S. J., Merges, R. P., Samuelson, P., & Sichelman, T. (2009). High Technology Entrepreneurs and the Patent System : Results of the 2008 Berkeley Patent Survey Author (s): Stuart J . H . Graham , Robert P . Merges , Pam Samuelson and Ted Sichelman Published by : University of California , Berkeley , School of Law S. Berkeley Technology Law Journal, 24(4), 1255–1327.

Graham, S. J., & Sichelman, T. (2008). Why Do Start-ups Patent? Berkeley Technology Law Journal, 23(3), 1063–1097.

Helmers, C., & Rogers, M. (2011). Does patenting help high-tech start-ups? Research Policy, 40(7), 1016–1027. https://doi.org/10.1016/j.respol.2011.05.003

Hoenig, D., & Henkel, J. (2015). Quality signals? the role of patents, alliances, and team experience in venture capital financing. Research Policy, 44(5), 1049–1064. https://doi.org/10.1016/j.respol.2014.11.011

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In Proceedings of the 56th annual meeting of the association for aomputational linguistics. https://doi.org/10.48550/arXiv.1801.06146

Jaffe, A. B. (2021). Patent Metrics for Innovation Research. (September).

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. Proceedings of the 15th conference of the European chapter of the association for computational linguistics, 2(Short Papers), 427–431. https://doi.org/10.1176/appi.ps.201500423

Mann, R. J. (2005). Texas Law Review. Texas Law Review, 83(4).

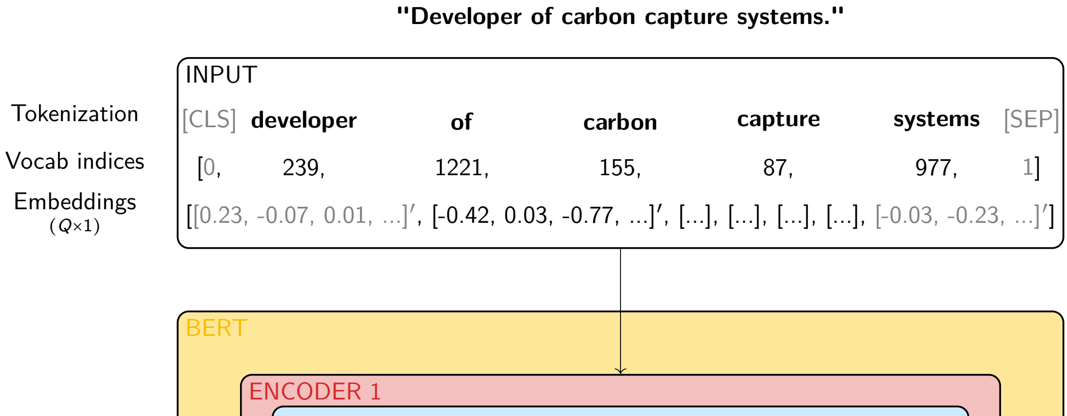
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations ofWords and Phrases and their Compositionality. Advances in neural information processing systems, arXiv 1606.08359, 3111–3119. https://doi.org/10.18653/v1/d16-1146

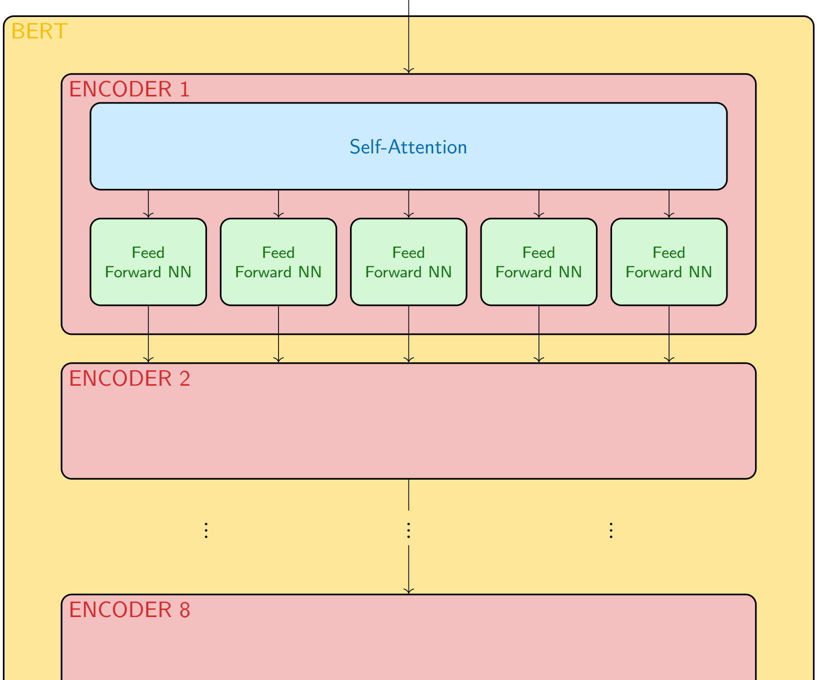
Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation Jeffrey. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532–1543. https://doi.org/10.1080/02688697.2017.1354122

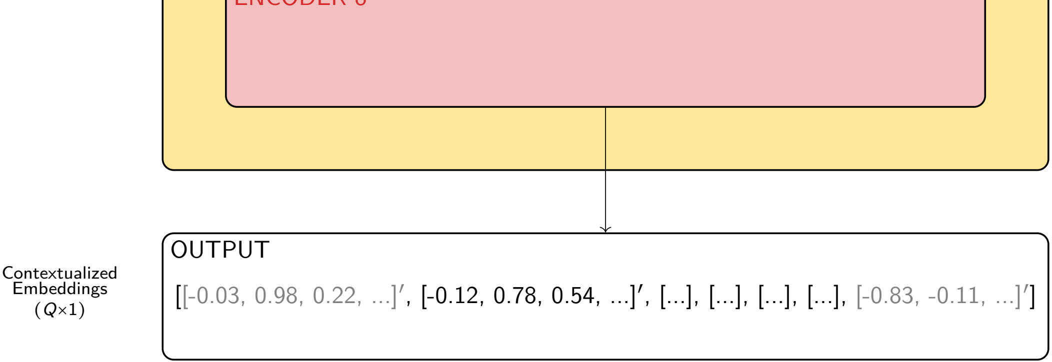
Peters, M. E., Neumann, M., Iyyer, M., & Gardner, M. (2018). Deep contextualized word representations. Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: 1(Long Paper), arXiv 1802.05365, 2227–2237.
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. https://doi.org/10.4310/HHA.2007.v9.n1.a16
Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 248–256. http://www.aclweb.org/anthology/D09-1026
Sick, N., Nienaber, A. M., Liesenkötter, B., vom Stein, N., Schewe, G., & Leker, J. (2016). The legend about sailing ship effects – Is it true or false? The example of cleaner propulsion technologies diffusion in the automotive industry. Journal of Cleaner Production, 137, 405–413. https://doi.org/10.1016/j.jclepro.2016.07.085
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), arXiv 1706.03762. http://arxiv.org/abs/1706.03762
Wang, Y., Hou, Y., Che, W., & Liu, T. (2020). From static to dynamic word representations: a survey. International Journal of Machine Learning and Cybernetics, 11(7), 1611–1630. https://doi.org/10.1007/s13042-020-01069-8

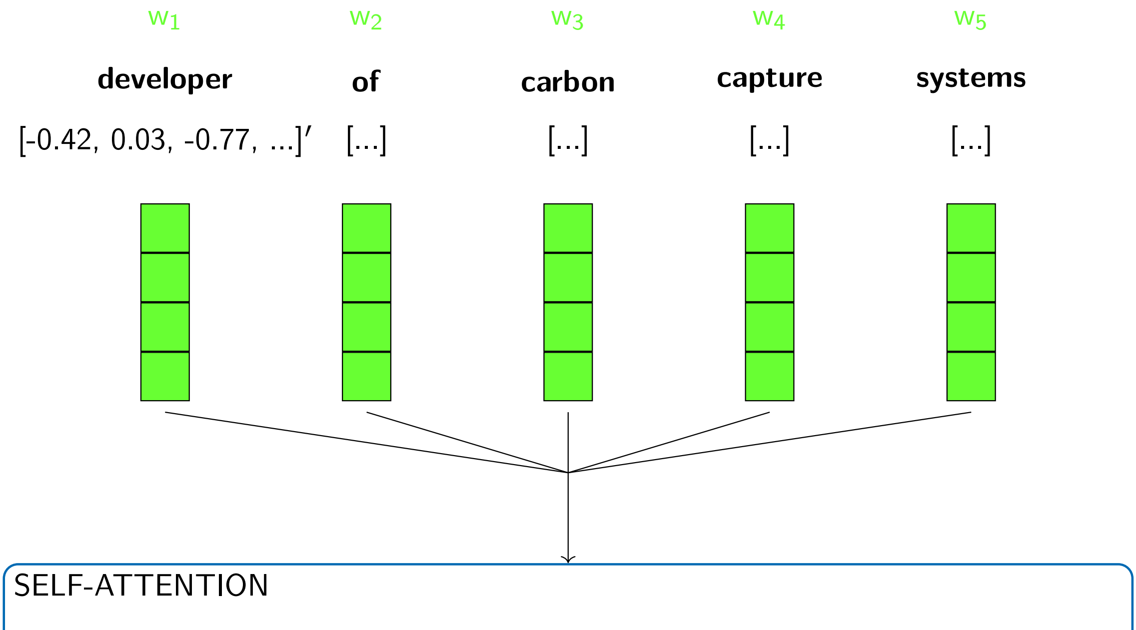
Appendix

Technology-company mapping framework

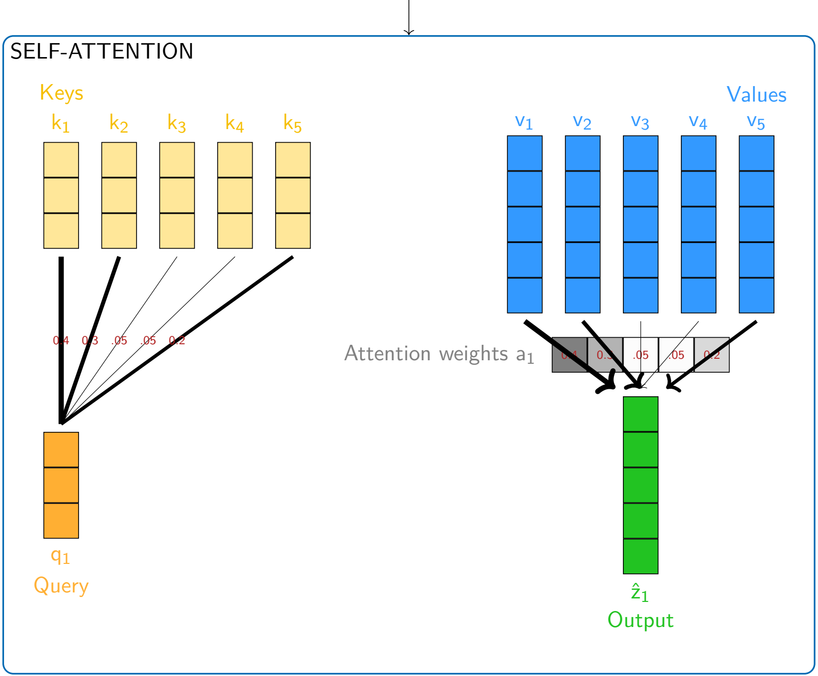

"Developer of carbon capture systems."
INPUT
Tokenization
[CLS] developer of carbon capture systems [SEP]
Vocab indices
[0, 239, 1221, 155, 87, 977, 1]
Embeddings
(Q×1)
[[0.23, -0.07, 0.01, ...]′, [-0.42, 0.03, -0.77, ...]′, [...], [...], [...], [...], [-0.03, -0.23, ...]′]
BERT
ENCODER 1

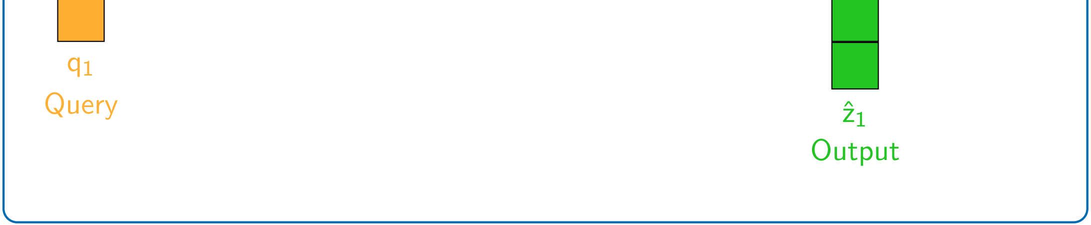

BERT
ENCODER 1
Self-Attention
Feed Forward NN
Feed Forward NN
Feed Forward NN
Feed Forward NN
Feed Forward NN
ENCODER 2
ENCODER 8


OUTPUT
Contextualized Embeddings (Q×1)
[[-0.03, 0.98, 0.22, ...]′, [-0.12, 0.78, 0.54, ...]′, [...], [...], [...], [...], [-0.83, -0.11, ...]′]



w1
w2
w3
w4
w5
developer
of
carbon
capture
systems
[-0.42, 0.03, -0.77, ...]′
[...]
[...]
[...]
[...]
SELF-ATTENTION


SELF-ATTENTION
Keys
k1 k2 k3 k4 k5
Values
v1 v2 v3 v4 v5
0.4 0.3 .05 .05 0.2
Attention weights a1
q1
Query
ẑ1
Output



1. Attention weights $a_{1:5}$ are query-key similarities:

   $\hat{a}_i = \mathbf{q}_i \times \mathbf{k}_i$

   Normalized via softmax: $a_i = e^{\hat{a}_i} / \sum_j e^{\hat{a}_j} \;\in [0, 1]$

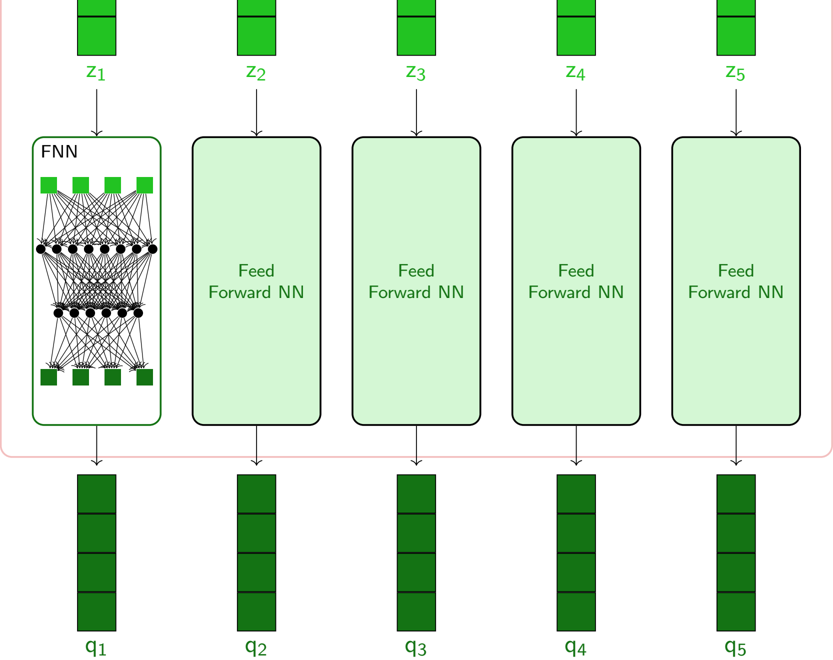
2. Output $\underset{(1\times Z)}{\widehat{\mathbf{z}}_i}$ is attention-weighted average of value vectors $\mathbf{v}_{1:5}$:

   $\widehat{\mathbf{z}}_i = \sum_i a_i \mathbf{v}_i$

3. $\mathbf{k}$, $\mathbf{v}$ and $\mathbf{q}$ are derived from the entire input $\mathbf{w}$:

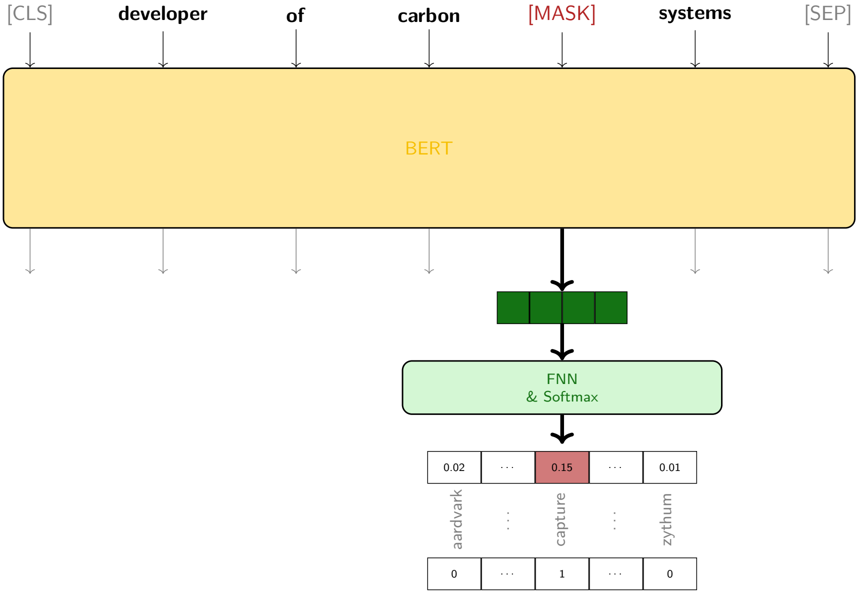
   $\mathbf{k} = W_k \times \mathbf{w} \qquad \mathbf{v} = W_v \times \mathbf{w} \qquad \mathbf{q} = W_q \times \mathbf{w}$

Note: Self-attention is repeated H times (multi-head attention) and the resulting vectors are concatenated along the feature dimension. Multiplying with a weight matrix $\underset{(Q\times HZ)}{W_z}$ yields the final output vector that is passed to the FNN.

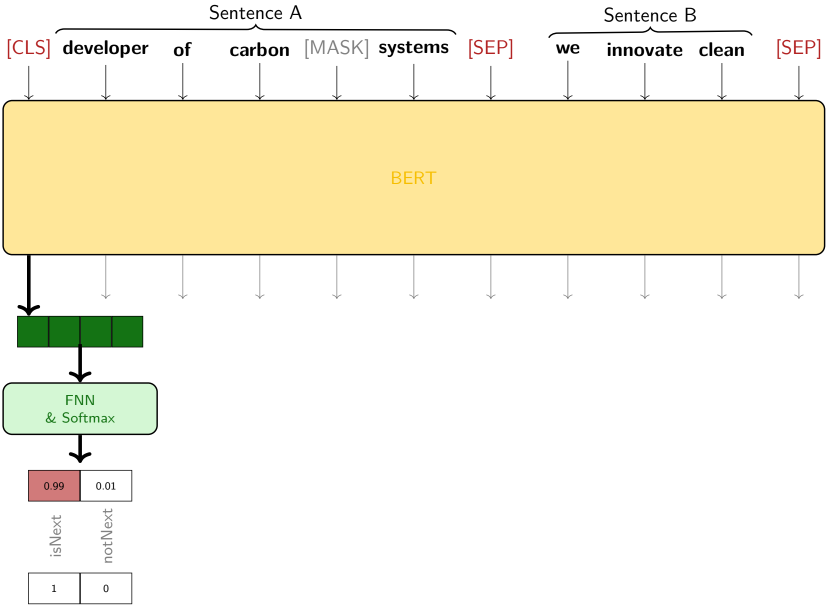


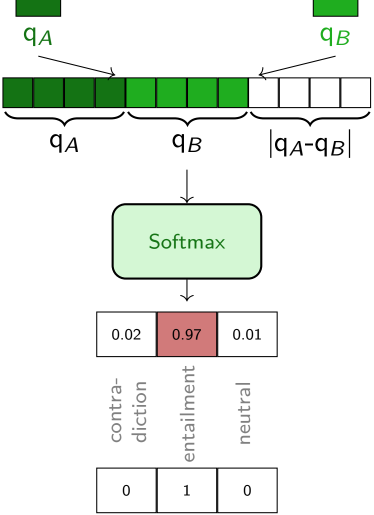
z_1
z_2
z_3
z_4
z_5
FNN
Feed Forward NN
Feed Forward NN
Feed Forward NN
Feed Forward NN
q_1
q_2
q_3
q_4
q_5

1. Masked language modeling



2. Next sentence prediction





q_A
q_B
q_A
q_B
$|q_A$-$q_B|$
Softmax
0.02
0.97
0.01
contra-
diction
entailment
neutral
0
1
0

Text preprocessing

1. translation of non-English texts to English
2. Part of Speech (PoS) tagging
    2.1 remove punctuation, numbers and unknown tags
    2.2 lemmatization
3. stop word deletion

A labeled corpus of patent abstracts

Patent	Technology class	Abstract
1	B, C, Y02C, Y02P	Catalyst, comprising one or more compounds of the perovskite-type as catalytically active component, is new, where the catalytically active component in the form of at least one layer is applied on a support body from an open cell foam ceramic material ...
2	A, Y02A, Y02C, Y02E	Absorber fluid, comprises a carbon dioxide binding absorbent and an ionic additive in a concentration, which is greater than a minimum concentration, so that the activity of the products formed by the connection of carbon dioxide to the absorbent is reduced ...
⋮	⋮	⋮
P	B, F, Y02C	The invention relates to a power plant for generating electrical energy, comprising a combustion chamber for producing steam, at least one waste gas purification stage that is connected downstream, a separation stage for CO2 ...

Note: Corpus comprises $P \sim 560,000$ patents (all patents filed by German firms after 1990) and a vocabulary size of $V \sim 370,000$ (after text preprocessing).

Vertical differentiation in technology classes

Classification system of the European Patent Office using the example of **carbon capture and storage technologies**:

CPC	COOPERATIVE PATENT CLASSIFICATION
Y	New technological developments
Y02	Climate change mitigation technologies
Y02C	Carbon capture and storage technologies
Y02C20	Capture and disposal of greenhouse gases
Y02C20/10	- of N_2O

Latent Dirichlet Allocation

Core idea in Blei et al. (2003) seminal work on Latent Dirichlet Allocation (LDA):
Model the generative process that led to the creation of a text corpus incorporating both:

- the observed words in the corpus' documents
- *and* the hidden topic structure within the corpus

in the imaginary data generating process.

The latter includes the distribution of topics over documents and the word distributions over topics.

Latent Dirichlet Allocation

Core idea in Blei et al. (2003) seminal work on Latent Dirichlet Allocation (LDA):
Model the generative process that led to the creation of a text corpus incorporating both:

- the observed words in the corpus' documents
- *and* the hidden topic structure within the corpus

in the imaginary data generating process.

The latter includes the distribution of topics over documents and the word distributions over topics.

L-LDA (Ramage et al., 2009) extents upon LDA by taking into consideration document labelsin the generative process.

L-LDA in patent corpus:

- document $\hat{=}$ patent, p
- labels/topics $\hat{=}$ technology classes, t
- word distributions over topics $\hat{=}$ semantic technology description, $p(\delta_t)$

Statistical Learning in L-LDA

Patent corpus D consisting of P distinct patent abstracts each of length N_p, generative process can be modeled as follows:

1. For each technology class $t \in \{1, \dots, T\}$: generate word distribution $\delta_t \sim Dir(\beta)$
2. For each patent $p \in \{1, \dots, P\}$: generate technology class distribution $\lambda_p \sim Dir(\alpha_p)$
3. For each of the word positions p, n, with $p \in \{1, \dots, P\}$ and $n \in \{1, \dots, N_p\}$:
   3.1 generate technology class assignment $z_{p,n} \sim Multinomial(\lambda_p)$
   3.2 and choose word $w_{p,n} \sim Multinomial(\delta_{z_{p,n}})$

$$p(\delta_{1:T}, \lambda_{1:P}, z_{1:P}, w_{1:P}) = \prod_{t=1}^{T} p(\delta_t) \prod_{p=1}^{P} p(\lambda_p) \left(\prod_{n=1}^{N_p} p(z_{p,n}|\lambda_p) p(w_{p,n}|\delta_{z_{p,n}}) \right)$$

Goal: Derive posterior distribution $p(\delta_t)$ from joint distribution $p(\delta_{1:T}, \lambda_{1:P}, z_{1:P}, w_{1:P})$

Importance of capture contextual meaning of words

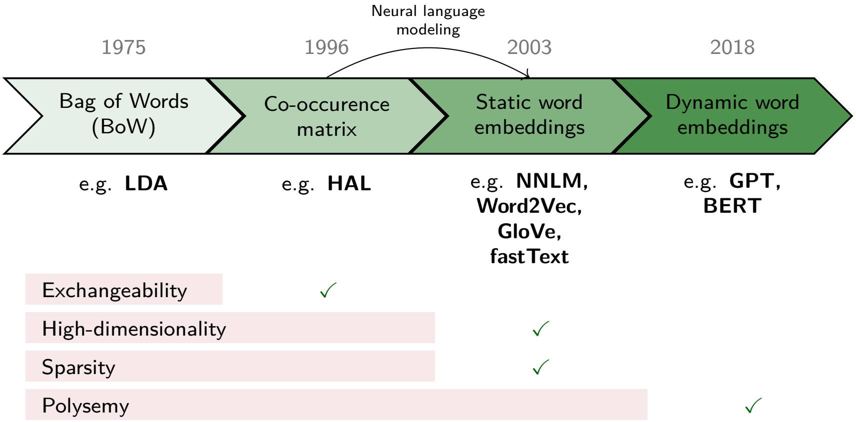
- **technical terms** in technology descriptions:
  $X_t = \langle$ gas, absorb, carbon, dioxide, desorption $\ldots \rangle$
- **non-technical terms** in company descriptions:
  '*Developer of direct air capture technology that safely and permanently removes CO2 from the air.*
  $\rightarrow Y_c =$ (developer, direct, air, technology, safe, permanent, remove, co2)'
- **But**: high semantic overlap between x_t and y_c as captured by token embeddings
  $\bar{X}_t(carbon) \approx \bar{Y}_c(co2)$
  $\bar{X}_t(absorb) \approx \bar{Y}_c(remove)$
- **Goal**: Exploit these relations to capture adopters of a technology

Classification performance of TechProx

Table: Performance of TechProx in distinguishing cleantech from non-cleantech firms

Label	Precision	Recall	F1-Score	Support
Cleantech	0.87	0.86	0.86	284
Non-cleantech	0.83	0.84	0.83	233
			0.85	517

Note: Performance measured on random test set with optimal values of $Q = 15$ and $\text{TechProx}_{min} = 0.27$. Optimal values for Q and TechProx_{min} have been determined on the validation set by tuning F1-Score.


Neural language modeling
1975
1996
2003
2018
Bag of Words (BoW)
Co-occurence matrix
Static word embeddings
Dynamic word embeddings
e.g. **LDA**
e.g. **HAL**
e.g. **NNLM, Word2Vec, GloVe, fastText**
e.g. **GPT, BERT**
Exchangeability ✓
High-dimensionality ✓
Sparsity ✓
Polysemy ✓

Word embeddings (1)

You shall know a word by the company it keeps!

Firth (1957)

General idea: exploit information on co-occurrence of words in large text corpora in order to learn the semantic meaning of a word as represented by a low-dimensional, dense vectors ($E \ll V$).

Natural Language Processing (NLP) as highly active field of research with major advances in recent years (see Wang et al. (2020)):

Neural Network Language Models

- 'distributed representation for words' (Bengio et al., 2003)
    - learn model that predicts next word given previous words
    - word embeddings carrying semantic meaning of a word as by-product

Word embeddings (2)

Static word embeddings

- Word2Vec (Mikolov et al., 2013)
  - neural network architecture specifically designed to learn word embeddings
  - Continuous Bag-of-Words (CBOW): predict word given its surrounding context words
  - Skipgram: predict context words given central word
- GloVe (Pennington et al., 2014)
  - direct exploitation of co-occurence statistics from large text corpora
- fastText (Bojanowski et al., 2017; Joulin et al., 2017)
  - learning embeddings for character n-grams and representing words as the sum of the n-gram embeddings (towards multi-language models)

Word embeddings (3)

Contextualized word embeddings

Tackle the issue that words have different meanings in different contexts (polysemy)

- ELMo (Peters et al., 2018)
  - use bidirectional LSTM to capture whole sentence (context!) in order to model embeddings of words in sentence
- ULMFit (Howard et al., 2018)
  - introduce a general language model and a process to fine-tune to domain-specific NLP tasks
- GPT (Radford et al., 2018)
  - use transformer network architecture to learn linguistic long-term dependencies
- BERT (Devlin et al., 2018)
  - Consider bidirectional contexts and relation of sentence pairs based on transformer encoders