Background & Motivation
What we do, how, and why

Patents & innovation performance: What we know...

- Technological & economic significance of patents varies broadly [Basberg, 1987].
- Consequently, the quality rather than number of patents more informative.

Patent Quality: What’s been done so far

- Number/composition of IPC assignments [Lerner, 1994].
- Lately, first attempts to introduce text (keyword) based indicators [Arts et al., 2017].

What we do instead...

1. Exploit rich textual information with semantic embedding techniques to capture technological signatures.
2. Relational mapping of similarity structures between patents (network analysis).
3. Temporal mapping of technological similarity between patents (lead-lag analysis).
5. Provide interactive visualization with high granularity (placecasting).

⇒ AKA: What (and where) will be “Europe’s Next Super Patent”?
Methods: Pipeline Overview
Our Approach in a Nutshell

Custom word embedding model
approx. 42m patent titles & abstract

Phrase embedding representation
approx. 10m patent titles & abstract transformed
TF-IDF weighted average embeddings

Similarity Matrix construction
Annoy Approximate Nearest Neighbours (Oh Yeah) matching

Lead-Lag
Analysing temporal similarity structures

Prediction
Auto-encoder anomaly detection approach

300-dimensional vector representations

target = "Cats are beautiful animals."
test = "Felines are gorgeous creatures."
test2 = "Dolphins are swimming mammals."
print(target.similarity(test))
print(target.similarity(test2))

Out: 0.9522696443950177
Out: 0.7822956256736615

interactive visualisation
Methods

For starters: Why to look at text?
Methods: 1 - Patent-to-Vector

Creating and validating

- Simple intuition: Counting keyword appearance → But what about synonyms, antonyms, analogies etc.?
- We instead use word embedding: Natural-language-processing technique that represents words as high dimensional vectors according to the context in which they tend to appear.
Methods: 2 - Vector-to-Similarity

Patent embedding & Technological distance

- We use a TF-IDF weighted average word embedding representations.
- Result: 300-dimensional patent embedding vector ⇒ technological signature.
- Embeddings subject to vector algebra ⇒ Distance between two patent embeddings = technological distance.

TF-IDF - weighted - embeddings

\[
\text{vec1} + \text{vec2} + \text{Vec3}
\]

electrical_connector characterised by a receptacle containing a plurality of female_contacts having redundant_contact portions and wiping_capabilities with respect to male_pins

target = "Cats are beautiful animals."
test = "Felines are gorgeous creatures."
test2 = "Dolphins are swimming mammals."

\[
\text{print(}\text{target}.\text{similarity(}\text{test}\text{))}
\]
\[
\text{print(}\text{target}.\text{similarity(}\text{test2}\text{))}
\]

Out: 0.9522696443950177
Out: 0.7822956256736615

First validation exercise

- Patent embeddings predict IPC3 Classification with 83% multi-class prediction accuracy. (out-of-sample).
- Patents which cite each others, are from the same applicant, inventor, patent family etc. have significantly lower technological distance.
Methods: 3 - Similarity-to-Quality

Temporal similarity: Intuition

- Semantic similarity independent of time.
- **Temporal similarity distribution** can be exploited
- Inspired by the lead-lag approach of Ramage et al. [2010]; Shi et al. [2010].

Temporal similarity: Types

**Similarity to past: Novelty**

- Exploitation of existing knowledge.
- High values might indicate backward orientation, low values indicate novelty.

**Similarity to present: Popularity**

- “Riding the wave”, indicates activity in a trending area.

**Similarity to future: Impact**

- Shaping the agenda, indicator of future impact.
- Also: Indicator of “Window-of-Opportunity”, high growth technological field.
Use-Case: Electromobility Technologies

- Electromobility related patents based on expert-advised IPC class selection.
- Further, all patents cited by “seed” also included (ca. 13k).

<table>
<thead>
<tr>
<th>IPC class</th>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>B60L 11/00</td>
<td>0</td>
<td>Electric propulsion</td>
</tr>
<tr>
<td>B60L 11/02</td>
<td>1</td>
<td>using engine-driven generators</td>
</tr>
<tr>
<td>B60L 11/04</td>
<td>2</td>
<td>using dc generators and motors</td>
</tr>
<tr>
<td>B60L 11/06</td>
<td>2</td>
<td>using ac generators and dc motors</td>
</tr>
<tr>
<td>B60L 11/08</td>
<td>2</td>
<td>using ac generators and motors</td>
</tr>
<tr>
<td>B60L 11/10</td>
<td>2</td>
<td>using dc generators and ac motors</td>
</tr>
<tr>
<td>B60L 11/12</td>
<td>2</td>
<td>with additional electric power supply</td>
</tr>
<tr>
<td>B60L 11/14</td>
<td>2</td>
<td>with provision for direct propulsion</td>
</tr>
<tr>
<td>B60L 11/16</td>
<td>1</td>
<td>using power stored mechanically</td>
</tr>
<tr>
<td>B60L 11/18</td>
<td>1</td>
<td>using power from primary cells</td>
</tr>
</tbody>
</table>
Use-Case: Electromobility Technologies

Aggregate Picture: Where is Novelty and Impact created?

[Graph showing a scatter plot with similarity to future and similarity to past on the axes, and countries represented as points on the graph.]

- CN
- FR
- IL
- TW
- JP
- IT
- DE
- NL
- CH
- CA
- KR
- SE
- FI
- GB
- US

Similarity to future:
- 0.08
- 0.09
- 0.10
- 0.11
- 0.12

Similarity to past:
- 0.1
- 0.2

internal

N Patents:
- 5000
- 10000
- 15000
Use-Case: Electromobility Technologies

(Less) Aggregate Picture

- Different levels of aggregation deliver different insights.
- Enables nuanced and disaggregated analysis where, by whom, and when novelty and impact is produced.

**Figure: Firm Level**

**Figure: Technology Level**
Use-Case: Electromobility Technologies

Dynamics over time: Capturing Technology Life-Cycles

▶ Reveals global technology life-cycles and “windows of opportunity”.
▶ Highlights different entry strategies and catching-up dynamics by latecomers.
Use-Case: Electromobility Technologies

Geography of Inventive Activity

- Providing granular insights in quality of inventive activity across regions.
- Facilitating smart specialization policies.
Methods: 4 - Nowcasting

Who would care so far?

- Academia: Interesting for historical and theoretical analysis.
- Policy: Not really. The state 5 years ago not so helpful for actions today...

⇒ Need for nowcasting (prediction)

A Note on Predictive Modelling

- Econometric modelling: Given a set of carefully selected variables of interest, how to identify causal effects on an outcome of interest?.
- Predictive modelling (aka machine learning): Given all available information, what is the best possible prediction of an outcome of interest ($\hat{y}$ rather than $\hat{\beta}$).

![Graph showing the trade-off between insight gain and amount of data with different learning capacities and interpretability.](image)

<table>
<thead>
<tr>
<th>Amount of data</th>
<th>Learning capacity</th>
<th>Interpretability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional techniques (e.g., linear, logistic regression)</td>
<td>SVM</td>
<td>Neural Nets</td>
</tr>
<tr>
<td>Nonparametric machine learning techniques (e.g., Support Vector Machines, Regression Trees)</td>
<td>Random Forest</td>
<td>Deep Learning Techniques</td>
</tr>
<tr>
<td>Deep Learning Techniques (e.g., Deep Neural Networks, Auto Encoders)</td>
<td>Regression Trees</td>
<td></td>
</tr>
</tbody>
</table>
(Ex-Ante) Predicting Patent Quality

- Forecasting of patent quality measures (novelty, impact, citations etc.) with modern ML rather easy.
- More interesting: Significant rare events → Europes Next Superpatent?
- Task: Identifying breakthrough patents (top-1%) [Ahuja and Lampert, 2001]

Rare event (anomaly) prediction

- **Deep Neural Autoencoder**: self-supervised model that aims at reproduction of its inputs
- Train on “boring normality” (non-breakthrough patents)
- High reproduction-error when facing anomal inputs → “something is wrong”.
- Results so far: Very nice AUC (>0.8), high accuracy (0.87) and sensitivity (0.81) out of sample.
Methods: 5 - Placecasting: The Global Patent Explorer
The power of visualization and data-narratives

- So far so good, but after all we just produce numbers.
- Complex data pipelines are of little use without producing a narrative.

⇒ We went a step further, and provide interactive visualizations of geolocations, granular geographical networks of knowledge flows, ad further indicators.¹

¹ As a goodie, many traditional patent measures [cf. Squicciarini et al., 2013].
Some central questions remain...

1. How to understand and trust predictions?

2. How to evaluate and improve predictions?
Modern predictive models (e.g., deep learning) are incredibly complex and nuanced. Result often:
What’s next? 1: Explain Models

Challenge 1: Explain model prediction

- **Global** model mechanics often to complicated for human annotation.
- **Local** model decision criteria can be approximated.
- One approach: “Local Interpretable Model-Agnostic Explanations” (LIME) [Ribeiro et al., 2016].
- Enables questioning and correcting model decisions.
- Can be used to increase fairness of models, and our trust in them.
What’s next? 2: Improve Models

Challenge 2: Improved and more nuanced predictions

- While our text-based method for technological similarity replicates commonly used approaches well, believes in superior performance are yet mainly technical.
- Even more prevalent when moving beyond similarity towards functional relationship mapping (eg., complements, substitutes, enabler, platforms).
- Ground truth still Human Intelligence.
- Computer Science approach to such hard problems: Produce a large annotated benchmark dataset → community challenges to push state-of-the-art.
- Example Computer Vision: IMAGENET - Enormeous (ca 1.2M train, 100k test) human annotated (1k classes) image dataset, annual competition. In 2010 unthinkable task → 2015: Solved (96.4% classification accuracy).
- Our (first) approach: Establish benchmark dataset of patent-similarity, joint effort of many cooperating POs.
Wrapping up

Some (Optimistic) Take-Away’s

1. Natural language processing (particularly: embedding) techniques are powerful tools to map and understand relationships in large bodies of text data.
2. Use-Cases are by far not limited to patent descriptions (eg., Policy reports, media debates, H2020 project descriptions, reports and meeting summaries).
3. Predictive modeling (ML) techniques have high potential to improve timely, granular, and precise forecasts of outcomes of interest (nowcasting & placecasting), and rare events (eg., breakthrough patents, unicorn start-ups).

Some more (Critical) Take-Away’s

1. ML models crucially depend on data (amount & details), and corresponding labels.
2. ML model mechanics tend to be opaque, but there are promising developments to change that.
3. Need of modern means of outcome-communication which are interactive (facilitates own insight generation), engaging (create data narratives), and selective (more not always better).
4. Collaborative effort needed to establish benchmarks, scrutinize and validate.
5. Open method and data workflows and requirements crucial for progress.
6. Cross-disciplinary efforts of Computer & Social Science necessary.
The End

Fin.


