LREC 2022 – Are Embedding Spaces Interpretable? Results of an Intrusion Detection Evaluation on a Large French Corpus

Thibault Prouteau
Nicolas Dugué
Nathalie Camelin
Sylvain Meignier
WORD EMBEDDING

- Representing words as dense vectors
- First building block of many NLP systems
- Dense vectors: convenient as input for neural nets
- Vectors supposed to encapsulate meaning of words

Distributional hypothesis

Firth 1957 — A word is characterized by the company it keeps.

→ Meaning of words emerging from their co-occurrences in the corpus
WORD EMBEDDING

Distributional hypothesis as a computational model

Using a big corpus:
the more words co-occur, the more one wants their embedding vectors to be similar

Let $U$ and $V$ be embedding matrices (parameters), $X$ the co-occurrence matrix, and $\cos$ the cosine similarity.
$\forall i, j \in Voc \times Voc$, one optimizes $U$ and $V$ such as $\cos(U_i, V_j) \approx f(X_{ij})$
**EXAMPLE : GLOBAL VECTORS (GloVe)**

\[ U_i^t \cdot V_j \approx f(X_{ij}) \]

Factorizing the co-occurrence matrix as the product of the embedding matrices

\[
\arg\min_{U,V} \sum_i \sum_j f(X_{ij})(U_i^t \cdot V_j + b_i^U + b_j^V - \log(X_{ij}))^2
\]
FIRST BUILDING BLOCK OF NLP SYSTEMS

- SNGS [3]
- GloVe [5]
- Bert [2]

→ Amazing perf on downstream tasks, but what about interpretability?
INTERPRETABILITY OF AI SYSTEMS?

- Understanding the decision
- Making possible for humans to adapt this decision
- Necessary to build trust-worthy systems
- One step towards auditable ai systems
LREC 2022 – Are Embedding Spaces Interpretable? Results of an Intrusion Detection Evaluation on a Large French Corpus

1. “Interpretable” approaches considered

2. An intrusion detection task

3. Results

4. Conclusion and perspectives
INTERPRETABILITY OF EMBEDDINGS

Interpretability?

Regarding word embedding, interpretability is the ability for humans to interpret each dimension of the space

| $\text{NNSE}_{1000}$ | inhibitor, inhibitors, antagonists, receptors, inhibition | bristol, thames, southampton, brighton, poole |
|                     | delhi, india, bombay, chennai, madras                  | pundits, forecasters, proponents, commentators, observers |
|                     | nosy, averse, leery, unsympathetic, snotty             |

*Figure: Murphy et al. 2012 [4]*
SPINE : SPARSE INTERPRETABLE NEURAL EMBEDDINGS
Subramanian et al. 2018 [8]

$k$: sparse overcomplete auto-encoder

- Sparse-code in a bigger than $300$ dimensional space
- Enforce vectors sparsity

Figure: From de Subramanian et al. 2018 [8]
SINR : SPARSE INTERPRETABLE NODE REPRESENTATIONS [6]

- Cooccurrence matrix as a graph
- Small dense clusters of this graphs: kind of *topics*
- A word is described by the distribution of its connections across these *topics*
  → Non-negative sparse vectors with tangible dimensions (dense clusters)
SINr : SPARSE INTERPRETABLE NODE REPRESENTATIONS IS NOT A SIN [6]
SINr: SPARSE INTERPRETABLE NODE REPRESENTATIONS IS NOT A SIN [6]

Figure: With Louvain [1] community detection
SINr: SPARSE INTERPRETABLE NODE REPRESENTATIONS IS NOT A SIN [6]

Figure: With Louvain [1] community detection
**SINr : SPARSE INTERPRETABLE NODE REPRESENTATIONS IS NOT A SIN [6]**

![Diagram with nodes and connections labeled 0 to 7, and two communities C0 and C1 with edge weights 0.33, 0.5, 0.5, 0.33, 0.17, 0.2, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4.]

**Figure : With Louvain [1] community detection**
SINr: SPARSE INTERPRETABLE NODE REPRESENTATIONS IS NOT A SIN [6]

Figure: With Louvain [1] community detection
**Hypothesis**: dimensions of the embedding space = communities $\rightarrow$ Interpretability?
LREC 2022 – Are Embedding Spaces Interpretable? Results of an Intrusion Detection Evaluation on a Large French Corpus

1. "Interpretable" approaches considered

2. An intrusion detection task

3. Results

4. Conclusion and perspectives
A dimension is said semantically coherent if a human can find the odd one (intruder) out of a set of words with the highest values on this dimension.

- Three models: Spine [8] with 1000 dimensions, SINr with $\approx 22k$ dimensions, and $\approx 5k$ dimensions.
- French news corpus (Le monde + AFP): 330M tokens, vocabulary of 323k words.
- 19 human annotators: students with knowledges in NLP.
- Dimensions are sampled at random, 200 for each embedding approach.
- Intruder is sampled in the bottom 30 of values in the dimensions at hand.
### WORD INTRUSION TASK

<table>
<thead>
<tr>
<th>Model</th>
<th>Top Words</th>
<th>Intruder</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPINE</td>
<td>suffrage</td>
<td>colmatage (sealing)</td>
</tr>
<tr>
<td></td>
<td>urne (ballot box)</td>
<td>législative (legislative)</td>
</tr>
<tr>
<td></td>
<td>tramway</td>
<td>rail orientation</td>
</tr>
<tr>
<td>SINr-1</td>
<td>monospace (multipurpose vehicle)</td>
<td>remarquer (to notice)</td>
</tr>
<tr>
<td></td>
<td>véhicule (vehicle)</td>
<td>voiture (car)</td>
</tr>
<tr>
<td></td>
<td>droit (law)</td>
<td>peine (penalty)</td>
</tr>
<tr>
<td>SINr-2</td>
<td>réseau (network)</td>
<td>déclencher (trigger)</td>
</tr>
<tr>
<td></td>
<td>chaîne (channel)</td>
<td>groupe (group)</td>
</tr>
<tr>
<td></td>
<td>Intel</td>
<td>garder (to keep)</td>
</tr>
<tr>
<td></td>
<td>microprocesseur (microprocessor)</td>
<td>processeur (processor)</td>
</tr>
</tbody>
</table>

**Table:** Examples of tasks extracted for each model [7]
**WORD INTRUSION TASK**

*Quel est l’intrus ?*

- fille
  - [1]
  - [2]
  - [3]

- grandeur
  - [4]
  - [5]
  - [6]

- femme
  - [7]
  - [8]
  - [9]

- épouse
  - [10]
  - [11]
  - [12]
LREC 2022 – Are Embedding Spaces Interpretable? Results of an Intrusion Detection Evaluation on a Large French Corpus

1 "Interpretable" approaches considered

2 An intrusion detection task

3 Results

4 Conclusion and perspectives
“Interpretable” approaches considered

An intrusion detection task

Results

Conclusion and perspectives

**RUNTIME**

<table>
<thead>
<tr>
<th>Model</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNGS + SPINE</td>
<td>17.2</td>
</tr>
<tr>
<td>SINr-1</td>
<td>1.3</td>
</tr>
<tr>
<td>SINr-2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table: Runtime of each model in hours.
RESULTS

<table>
<thead>
<tr>
<th></th>
<th>SPINE</th>
<th>SINr-1</th>
<th>SINr-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>IntruderOK</td>
<td>36%</td>
<td>31%</td>
<td>35%</td>
</tr>
<tr>
<td>+ HesitateOK</td>
<td>56%</td>
<td>53%</td>
<td>60%</td>
</tr>
<tr>
<td>+ Consistent</td>
<td>57%</td>
<td>58%</td>
<td>62%</td>
</tr>
</tbody>
</table>

Table: Positive results of the intrusion detection task.
RESULTS

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>SPINE</th>
<th>SINr-1</th>
<th>SINr-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>56%, 17%</td>
<td><strong>58%, 21%</strong></td>
<td>55%, 17%</td>
<td>55%, 13%</td>
</tr>
</tbody>
</table>

Table: Inter-annotator agreements across all models presented and overall for the Word Intrusion Task. For each model, the first value is the percentage of tasks where at least two evaluators annotated similarly. The second value is the percentage of tasks where the three evaluators annotated similarly.
LREC 2022 – Are Embedding Spaces Interpretable? Results of an Intrusion Detection Evaluation on a Large French Corpus

1 "Interpretable" approaches considered

2 An intrusion detection task

3 Results

4 Conclusion and perspectives
CONCLUSION

- SINr runs fast and is on-par with Spine on the intrusion task
- Results show that interpretability is only reached in 50 to 60% of the cases
- Hard task with quite low inter-annotator agreement
Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 
Fast unfolding of communities in large networks. 

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 
Bert: Pre-training of deep bidirectional transformers for language understanding. 

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 
Efficient estimation of word representations in vector space. 
*arXiv: 1301.3781.*
Brian Murphy, Partha Talukdar, and Tom Mitchell. 
Learning effective and interpretable semantic models using non-negative sparse embedding. 

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 
Glove: Global vectors for word representation. 
In EMNLP, pages 1532–1543, 2014.

Thibault Prouteau, Victor Connes, Nicolas Dugué, Anthony Perez, Jean-Charles Lamirel, Nathalie Camelin, and Sylvain Meignier. 
Sinr: Fast computing of sparse interpretable node representations is not a sin! 
In Advances in Intelligent Data Analysis XIX, number 12695, pages 325–337, 2021.
Thibault Prouteau, Nicolas Dugué, Nathalie Camelin, and Sylvain Meignier.
Are embedding spaces interpretable? results of an intrusion detection evaluation on a large french corpus.
In LREC, 2022.

Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy.
Spine : Sparse interpretable neural embeddings.
In AAAI Conference on Artificial Intelligence, 2018.