Thibault Prouteau

Nicolas Dugué

Nathalie Camelin

Sylvain Meignier

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

Le Mans Université

LIUM
Laboratoire d'Informatique
Université du Mans

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.

$\rightarrow$ 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

$$\underset{U,V}{argmin} \sum_i \sum_j f(X_{ij})(\mathbf{U_i^t} \cdot \mathbf{V_j} + b_i^U + b_j^V - \mathbf{log}(\mathbf{X_{ij}}))^2$$

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

INTERPRETABILITY OF EMBEDDINGS

Interpretability ?

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

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

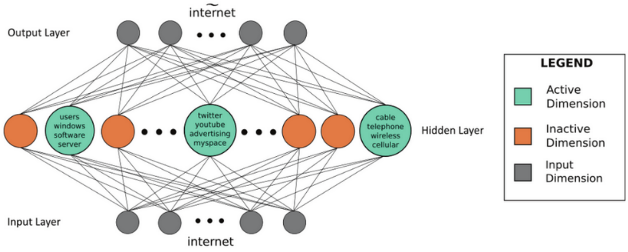


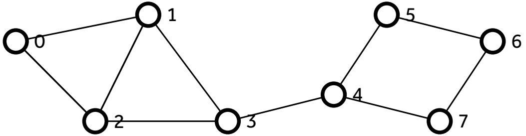
Figure : From de Subramanian et al. 2018 [8]

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*

$\rightarrow$ 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]

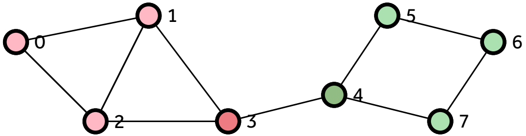


Figure : With Louvain [1] community detection

SINr : SPARSE INTERPRETABLE NODE REPRESENTATIONS IS NOT A SIN [6]

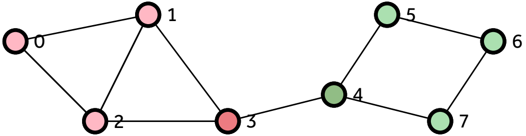
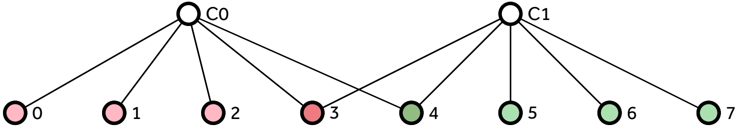


Figure : With Louvain [1] community detection



SINr : SPARSE INTERPRETABLE NODE REPRESENTATIONS IS NOT A SIN [6]

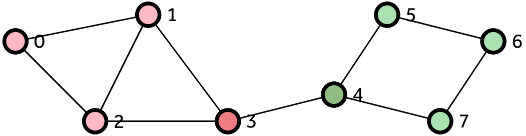
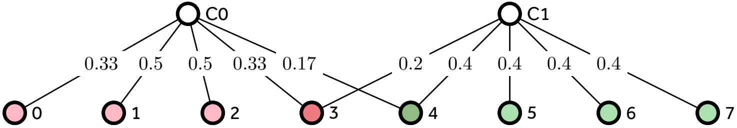


Figure : With Louvain [1] community detection



SINr : SPARSE INTERPRETABLE NODE REPRESENTATIONS IS NOT A SIN [6]

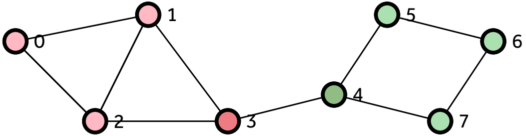
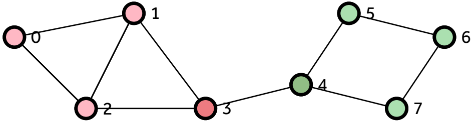
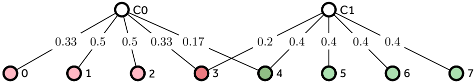


Figure : With Louvain [1] community detection

3	0.33	0.66	0.2	0.33
4	0.17	0.33	0.4	0.66
7	0	0	0.4	1

SPARSE INTERPRETABLE NODE REPRESENTATIONS





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

WORD INTRUSION TASK

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.

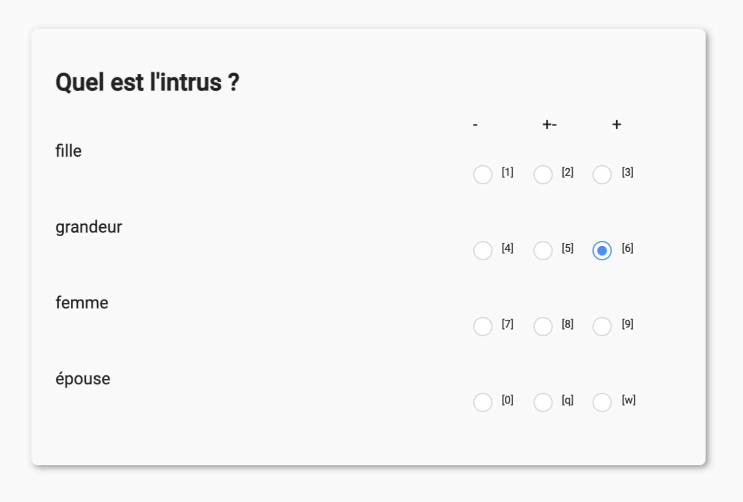
- Three models : Spine [8] with 1000 dimensions, SINr with $\approx$ 22k dimensions, and $\approx$ 5k dimensions
- French news corpus (Le monde + AFP) : 330M tokens, vocabulary of 323k words
- 19 human annotators : students with knowledges in NLP
- Dimensions are sampled at random, 200 for each embedding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

WORD INTRUSION TASK

Model	Top Words			Intruder
`SPINE`	suffrage	urne *(ballot box)*	législative *(legislative)*	colmatage *(sealing)*
	tramway	ferroviaire *(rail)*	rail	orientation
`SINr-1`	monospace *(multipurpose vehicle)*	véhicule *(vehicle)*	voiture *(car)*	remarquer *(to notice)*
	droit *(law)*	mort *(death)*	peine *(penalty)*	fortune
`SINr-2`	réseau *(network)*	chaîne *(channel)*	groupe *(group)*	déclencher *(trigger)*
	Intel	microprocesseur *(microprocessor)*	processeur *(processor)*	garder *(to keep)*

Table : Examples of tasks extracted for each model [7]

WORD INTRUSION TASK



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

RUNTIME

Model	Runtime
SNGS+ SPINE	17.2
SINr-1	1.3
SINr-2	**1**

Table : Runtime of each model in hours.

RESULTS

	SPINE	SINr-1	SINr-2
IntruderOK	**36%**	31%	35%
+ HesitateOK	56%	53%	**60%**
+ Consistent	57%	**58%**	**62%**

Table : Positive results of the intrusion detection task.

RESULTS

Overall	SPINE	SINr-1	SINr-2
56%, 17%	**58%, 21%**	55%, 17%	55%, 13%

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

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

REFERENCES

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre.
Fast unfolding of communities in large networks.
J. Stat. Mech. : Theory Exp., 2008(10) :P10008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
Bert : Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv :1810.04805, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.
Efficient estimation of word representations in vector space.
In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 2013.
arXiv : 1301.3781.

REFERENCES

Brian Murphy, Partha Talukdar, and Tom Mitchell.
Learning effective and interpretable semantic models using non-negative sparse embedding.
In *COLING*, pages 1933–1950, 2012.

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.

REFERENCES

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.