Contrat Doctoral — ED Galilée

Titre du sujet : Logical and Geometric Structures Induced by Statistical Data

- ➢ Unité de recherche : LIPN – UMR7030
- ➢ Discipline : Informatique
- ➢ Direction de thèse : Thomas Seiller (CR CNRS, HdR), co-encadrant : Juan-Luis Gastaldi (ETH Zurich)
- ➢ Contact : thomas.seiller@cnrs.fr
- ➢ Domaine de recherche : Informatique, Logique, Théorie des Catégories
- ➢ Mots-clés : Logique linéaire, Théorie des catégories

Projet Scientifique détaillé (en anglais) :

The ability of Large Language Models (LLMs) to produce syntactically, semantically, and grammatically correct sentences demonstrates that significant structure of natural language can be extracted from statistical data on large textual corpora. Modern neural architectures produce coherent and plausible text while relying almost exclusively on statistical learning procedures. Despite their practical success, however, the formal structures underlying these systems remain poorly understood.

Current interpretability research explains neural behavior after training, via probing, attribution, or causal intervention. These approaches remain tied to implementation details and offer no principled account of why statistical learning produces linguistic and inferential structure. This project takes a different approach : rather than analysing trained models, it derives formal structure directly from corpus statistics, independently of any neural implementation.

Word embeddings, the first layer of neural networks used for linguistic tasks, provide a classical example where statistical data and geometric structure can be explicitly related. Levy and Goldberg [14] showed that classical embedding models perform an implicit matrix factorization of a shifted PMI matrix [1] computed on words and contexts.

This result opens the way to a more systematic question : can one recover richer structure — compositional, logical — from corpus statistics, without importing linear algebraic assumptions extraneous to the data ? Bradley, Gastaldi, and Terilla [7] proposed a categorical approach based on profunctors and enriched adjunctions to address precisely this. Starting from a profunctor

$$M : \mathbf{C}^{\mathrm{op}} \times \mathbf{D} \to \mathbf{Set}$$

representing co-occurrence statistics, one defines an adjunction $M_* \dashv M^*$. The *nucleus* of this adjunction, i.e. pairs (A, B) satisfying $M^*(A) = B$ and $M_*(B) = A$, generalize several classical constructions such as eigenspaces, Galois correspondences, formal concepts [18], and singular value decompositions. Crucially, the construction can be performed in the real-valued setting by defining real-valued presheaves $A : D \to \mathbf{R}$ to represent word vectors without assuming linearity or additional algebraic structure.

Together with Jarvis, Gastaldi, and Terilla, I observed that the PMI measure satisfies the following trefoil (2-cocycle) identity under concatenation :

$$[\![w_1 \cdot w_2, w_3]\!] + [\![w_1, w_2]\!] = [\![w_1, w_2 \cdot w_3]\!] + [\![w_2, w_3]\!],$$

1. Pointwise mutual information (PMI) measures in/dependence. It is computed as $\frac{p(ab)}{p(a)p(b)}$ where p is a probability distribution. The value equals 1 when a and b are independent and departs from 1 depending on positive or negative dependence.

mirroring a central property originally identified in the Interaction Graphs framework [16, 17]. This observation implies that words together with concatenation and the PMI measure define what we call a *linear realizability situation*. Using constructions inspired by realizability and linear logic, this allows the definition of a linear realizability model in which types (equivalently, logical formulas) correspond exactly to points of the nucleus.

Importantly, this construction induces logical structure directly from corpus statistics [10]. The induced logic resembles, although it differs from, the logic of Lambek categories [12] ; logical connectives correspond to a notion of *derived nuclei* whose complex structure evaded purely categorical methods. Moreover, understanding types in this way clarified several geometric properties of the nucleus : it defines a tropical projective space together with a cell complex structure whose wall-crossing behavior and further decomposition properties were studied in recent work [9].

The PhD project will contribute to this larger research program through three closely related research directions.

1. **Logical structures arising from corpora.**

   The central objective of this axis is to understand what kinds of inferential and compositional structures are naturally induced by statistical organization in corpora, and to establish characterization theorems relating the internal logic of enriched nuclei to known categorical structures. Several families of such theorems are known : Grothendieck topoi correspond to categories of models of geometric theories ; elementary topoi provide models of higher-order intuitionistic logic. The question is whether analogous results hold for the enriched setting arising from the nucleus construction.

   Preliminary results already point in this direction. The induced logic and the geometry-logic correspondence described above provide a starting point ; several concrete questions remain open. The distance between points a, b in the nucleus corresponds to a value in the type

   $$(T(a) \multimap T(b)) \odot (T(b) \multimap T(a)),$$

   where $T(x)$ denotes the type of point x and $\odot$ is a non-commutative conjunction. Several questions follow from this : whether the geometric decomposition of [9] has a logical counterpart ; whether the remaining values in $(T(a) \multimap T(b)) \odot (T(b) \multimap T(a))$ admit geometric interpretations ; and what the geometric role of the tensor unit is. A further open question concerns the extent to which the induced logic extends to richer substructural systems : the types induced by the PMI matrix are expected to carry additive connectives, since the types, seen as presheaves, are formal sums of pairs (a, α), and moving from pairs to formal sums is the standard extension technique from Multiplicative Linear Logic to MALL, which would ensure the existence of categorical products and coproducts on the nucleus.

2. **Nuclei as generalized decomposition methods.**

   The nucleus construction generalizes SVD : in the linear setting, the nucleus of the adjunction $M_* \dashv M^*$ coincide with the eigenspaces of the operator defined by M, recovering the usual singular value decomposition as a special case. The general objective of this axis is to make the nucleus construction available as a practical replacement for SVD, applicable wherever SVD is currently used for data analysis. This requires addressing two problems, whose difficulty makes this axis more exploratory than the first.

   The first is dimension reduction. SVD is useful in practice not merely as a decomposition but because one can truncate it, retaining only the top k singular values to obtain a low-rank approximation. The analogue of this truncation in the nucleus setting is not yet understood. In the tropical setting to which the nucleus construction naturally belongs, low-rank approximation is a genuinely difficult problem :

there are at least four competing definitions, most of which are NP-complete, and it is an active area of research, notably by Gaubert and collaborators [11, 15]. However, we already have candidate algorithms for dimension reduction in our specific setting, and we expect the particular structure of the realizability framework — in which the tropical matrices arise from PMI statistics and carry additional algebraic constraints — to make the problem substantially more tractable than in the general case studied by Gaubert. A key task is therefore to understand whether these candidate algorithms can be given a mathematical characterization as performing some form of optimal reduction.

The second problem is computational feasibility. The project will develop efficient algorithms and data structures for computing nuclei on realistic corpora and datasets, with the aim of producing implementations that can serve as drop-in replacements for SVD-based pipelines in NLP and data analysis.

3. **Applications to structuralist data analysis.**

   The third direction provides a concrete validation testbed for the nucleus construction : correspondence analysis (CA), a dimensionality-reduction technique closely related to SVD applied to contingency tables, which should therefore naturally be subsumed by the nucleus construction. It is the standard method in social science, with published datasets and well-established results [1, 2, 8, 13, 5, 3, 4, 6], making it an ideal benchmark : the ground truth is known, the data is available, and any structural gain over CA is directly interpretable. The nucleus generalizes CA in three mathematically precise ways : it admits a hierarchical organization of factors allowing overlap and inclusion rather than enforcing orthogonality ; it carries a product structure on the nucleus that represents interactions between data points as genuinely new positions rather than mere aggregates ; and it provides a type-theoretic representation that complements the geometric picture with algebraic and logical structure [10].

   The empirical objective is to apply the nucleus construction to the datasets underlying the classical sociological studies cited above, first to reproduce the CA results and then to identify novel structure. This provides a controlled validation setting : the ground truth is known, the data is published, and we expect any structural gain to be interpretable. Contemporary DNN-based analyses of similar data will be reinterpreted as implicit instances of this framework, making precise a theoretical dimension that is otherwise absent.

Références

[1] J.-P. Benzécri. *L'analyse des données. 1 La taxinomie.* Bordas, Paris, 1973. Et coll.

[2] J.-P. Benzécri. *L'analyse des données. 2 L'analyse des correspondances.* Bordas, Paris, 1973. Et coll.

[3] L. Boltanski. *Les Cadres : La formation d'un groupe social.* Le sens commun. Les Éditions de Minuit, 1982.

[4] L. Boltanski, Y. Darré, and M.-A. Schiltz. La dénonciation. *Actes de la recherche en sciences sociales*, 51 :3–40, 1984.

[5] P. Bourdieu. *La distinction. Critique sociale du jugement.* Le sens commun. Les 'Editions de Minuit, Paris, 1979.

[6] P. Bourdieu. Une révolution conservatrice dans l'édition. *Actes de la recherche en sciences sociales*, 126-127 :3–28, 1999.

[7] T. D. Bradley, J.-L. Gastaldi, and J. Terilla. The structure of meaning in language. *Notices of the American Mathematical Society*, Feb 2024.

[8] P. Cibois. *L'analyse factorielle*. Que sais-je ? Presses Universitaires de France, Paris, 1983.

[9] J.-L. Gastaldi, S. Jarvis, T. Seiller, and J. Terilla. Geometric structures in **R**-enriched adjunctions. submitted, https://hal.science/view/index/docid/5452748.

[10] J.-L. Gastaldi, S. Jarvis, T. Seiller, and J. Terilla. Linear realizability and structures in **R**-enriched adjunctions. in preparation.

[11] S. Gaubert, W. M. McEneaney, and Z. Qu. Curse of dimensionality reduction in max-plus based approximation methods : Theoretical estimates and improved pruning algorithms. *IEEE Conference on Decision and Control and European Control Conference*, pages 1054–1061, 2011.

[12] J. Lambek. Bicategories in algebra and linguistics. *Linear logic in computer science, London Mathematical Society Lecture Note Series*, 316, 2004.

[13] B. Le Roux and H. Rouanet. *Geometric Data Analysis : From Correspondence Analysis to Structured Data Analysis*. Springer Netherlands, 2006.

[14] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27, 2014.

[15] O. Saadi. *Zero-sum repeated games : accelerated algorithms and tropical best-approximation*. PhD thesis, 2021. Thèse de doctorat dirigée par Gaubert, Stéphane et Akian, Marianne Mathématiques appliquées Institut polytechnique de Paris 2021.

[16] T. Seiller. Interaction graphs : Additives. *Annals of Pure and Applied Logic*, 167 :95 – 154, 2016.

[17] T. Seiller. Mathematical informatics, 2024. Habilitation thesis.

[18] R. Wille. Restructuring lattice theory : an approach based on hierarchies of concepts. In *International conference on formal concept analysis*, pages 314–339. Springer, 2009.