氏　　名	前田晃弘
学位の種類	博士（情報科学）
学位記番号	博情第 576 号
学位授与年月日	令和 8 年 3 月 25 日
論文題目	Laying Algebraic-Geometric Foundation for Probabilistic Models of Natural Languages
論文審査委員	日髙　昇平　北陸先端科学技術大学院大学　准教授
	上原　隆平　　同　教授
	吉高　淳夫　　同　准教授
	井之上　直也　同　准教授
	大関　洋平　東京大学　准教授

論文の内容の要旨

Human language exhibits a remarkable form of compositionality: complex expressions are systematically built from simpler parts, and this structure supports strong out-of-distribution generalization. Despite the empirical success of modern neural language models, no existing approach provides a principled probabilistic account of how such compositional structure is represented, computed, or learned. This dissertation addresses this gap by developing a mathematical framework that treats linguistic probability itself as an algebraic and geometric object.

The first contribution is the formulation of a *structural probability model*, grounded in algebraic statistics. Sentences are represented as joint probability tensors whose algebraic constraints correspond to geometric entities. This provides the first unified framework in which the compositional organization of language is expressed as algebraic system and appears as low-dimensional geometric structure in distributional representations.

The second contribution is the introduction of a new invariant unit of probabilistic structure, termed the *Minimum Invariant Constraint (MIC)*. An MIC is defined as the atomic algebraic component of a probability tensor, mathematically characterized by a vanishing 2×2 minor. MICs generalize the notions of independence; they are invariant under reparameterization, robust across corpora, and serve as the irreducible building blocks for complex structural patterns. This framework provides a principled explanation for the local rank-one patterns observed in PMI matrices and co-occurrence statistics.

The third contribution is the development of computational methods for discovering MICs and their compositions in empirical data. Two complementary approaches are introduced: (i) a divide–and–conquer method based on marginalization identities and cumulants, which interprets PMI heuristics as second-order cumulants; and (ii) a harmonic-analysis method using the Walsh–Hadamard transform and geometric algebra, enabling efficient detection of symmetric and low-rank structure in high-order tensors. Proof-of-concept experiments demonstrate that these methods recover invariant components that were previously inaccessible to conventional tensor decompositions.

Together, these contributions establish a new algebraic–geometric foundation for linguistic compositionality. They show that the internal organization of language emerges as invariant algebraic constraints on probability distributions, offering both theoretical insight into the structure of language and practical tools for analyzing modern language models.

Keywords: language model, algebraic statistic, invariant constraint, compositionality, vanishing binomial

論文審査の結果の要旨

本論文は、自然言語の確率的モデル（特に、言語モデル）を対象として、言語の「構成性」に着目した上で、大規模言語モデルの成功の背景にある代数的・幾何学的な基盤を提案している。既存の言語モデルでは「意味構造」と「数理構造」の対応（構成性）が理論的に説明されていないという問題意識に基づき，言語の対称性（交換可能性）から必然的に生じる確率分布上の不変構造を最小不変制約(MIC: Minimum Invariant Constraint)として定義し，それを確率テンソル・共起行列から同定・学習できることを理論・アルゴリズム・実証の 3 つの観点から示している。

具体的に、本論文は言語を表現したモデルの確率分布の構造を MIC の組み合わせとして分析する方法を提案した。これによりコーパス全体を代数的な分析が可能になり、単語間の関係を表す共起行列に潜在する MIC の構造を抽出することで意味付け可能な単語群を特定する数理実験も示されている。

主な貢献として、第一に、代数学に基づく「構造的確率モデル」を定式化し、文を確率テンソルとして扱うことで、言語の構成性を分布表現上の幾何学的構造として定義する統一的フレームワークを構築した。第二に、確率構造の不変単位として独立性の概念を一般化した「最小不変制約（MIC）」を導入し、単語の PMI(Pointwise Mutual Information)共起行列等に見られる局所的なランク 1 パターンの理論的根拠を明らかにした。第三に、MIC 発見のための計算手法として、キュムラントに基づく分割統治法や、幾何代数を用いた調和解析手法を開発し、実データから高次テンソル内の低ランク構造を効率的に検出できることを実証した。

本研究の発想は自然であるが，これを最小不変制約で理論的に明確に定式化している点に独創性と新規性がある。さらに素朴な計算方法では計算コストが高くなるところを，様々なアルゴリズム的な工夫により効率化しており，情報科学的観点からも新規性と実用的な意味での有用性が示されている。実言語規模の応用には依然として解決すべき技術的な課題が複数残っているものの、構成性・対称性を帰納バイアスとして備えた次世代言語モデルに応用できるという点で自然言語処理における有用性が認められる。

加えて、本博士学位論文は、自然言語処理のトップ国際会議である ACL や人工知能学会の論文誌である「人工知能」で出版した論文に基づくなど、業績的な観点からも学術的水準は十分である。従って、本論文は博士（情報科学）の学位論文として十分価値あるものと認めた。