| | |
|---|---|
| 氏名 | 横井　祥 |
| 研究科、専攻 | 東北大学大学院情報科学研究科（博士課程）　システム情報科学専攻 |
| 学位論文題目 | Computing Co-occurrence with Kernels<br>（カーネル法に基づく共起の計算） |

# 博士学位論文全文の要約

## 第１章　序論

ふたつの言語表現（単語・文・文書など）の間の**共起の強さ**（対応の良さ・くっつきやすさ）のモデル化と計算は，自然言語処理および計算言語学において常に立ちはだかる基本的問題である．たとえば “New York” のようにスペースで区切られているがひとかたまりになっていると考えられる語句を同定する際（collocation extraction），たとえば人間と対話するロボットが適切な回答候補を選択する際（dialogue response selection），たとえばウェブから自動で収集された機械翻訳の訓練データが対訳データとして適切かを判定する際（noisy parallel corpus filtering）など，共起の問題が現れる場面は枚挙にいとまがない．

これまで，共起のモデル化と計算には**自己相互情報量（PMI）**と呼ばれる情報論的尺度が一般的に利用されてきた．しかし PMI には疎な表現（文や文書）への適用可能性と計算効率の間にトレードオフがある．さらに自然言語処理のための統計的尺度には，小規模訓練データ（low resource language）への適用可能性，解釈性・説明可能性，また十分な性能が要請される．以上すべての要件を満たす共起の尺度はいまだ確立されておらず，言語データの情報処理の大きな課題のひとつである．論文を通してこの問題に取り組む．

## 第２章　カーネル法に基づく共起尺度 PHSIC の提案

PMI が「相互情報量へのインスタンスペアの貢献度」と捉えられることと対応付け，新しい共起尺度である **Pointwise HSIC（PHSIC）**を「カーネル法に基づく依存性尺度 HSIC へのインスタンスペアの貢献度」として提案する．

PHSIC のナイーブ推定量を考えると，実際に PHSIC は PMI をカーネルでスムージングした量となることがわかる．またこのことが，PHSIC がカーネル（類似度関数）を利用することで疎な表現への適用可能になる理由づけとなる．

また，近年盛んに研究されている単言語コーパスから学習された文エンコーダーを用いることで，性能の向上，さらに小規模訓練データへの適用可能性が高まることが期待される．

## 第３章　計算効率の高い PHSIC の推定量の導出

PHSIC をナイーブに推定すると，データ数に対して二乗のオーダーの学習時間を要する．本章では，線形カーネルやコサイン類似度等を用いる場合はカーネルトリックを “逆に” 使うことによってデータ数に対して線形のオーダーでの推定が可能となることを示す．

また RBF カーネル等の非線形カーネルを用いる場合は不完全コレスキー分解を介したグラム行列の近似をおこなうことによって，同様のオーダーを達成することを示す．

## 第４章　PHSIC の推定量への解釈性の付与

機械学習に基づく予測に対して解釈性の付与するための基本的な指針は，予測を特徴量や訓練データの線形和に近似することである．提案法はカーネル法に基づく手法であり，言い換えれば特徴空間での線形なデータ解析に基づく．前章で提案した高速な推定手法も，特徴量および訓練データについて双線形形式であるため，（計算量を支払うことで）近似抜きでのそれぞれの寄与度の評価が可能となる．以上を踏まえ，PHSIC に対する解釈性の付与について述べる．

## 第５章　実応用（リランキングおよびデータセレクション）での実証実験

提案尺度である PHSIC が，これまでの章で理論的に示した様々な要請を満たすことを，対話システムおよび機械翻訳という自然言語処理でもっとも盛んに取り組まれている実タスク・実データで実際に確認する．

実験結果として，PHSIC を対話の応答文選択に適用すると，性能を犠牲にせず PMI よりも約 1000 倍高速に学習できることを示した．また PHSIC は学習データが少ない設定でも性能の低下がほとんど見られないことを確認した．さらに PHSIC を機械翻訳の訓練データのフィルタリングに適用し，フィルタリング後の訓練事例を用いて学習された翻訳器の性能の観点で最良のフィルタリング手法に勝ることを確認した．

加えて，PHSIC を対話の応答文選択に用いる場合も機械翻訳の訓練データのフィルタリングに用いる場合も，予測された共起の強さを訓練データの和に分解することで高い解釈性を与えられることを確認した．

**第６章　カーネル法に基づく依存性の最大化による知識獲得**

さらなる応用タスクとして知識獲得に取り組む．一般に，実世界で観測される言語データ（たとえば“昨日友人とお気に入りの和食屋で夕飯を食べて，すっかり腹がふくれた”）は背景となる世界知識（たとえば〈夕飯を食べる，腹がふくれる〉）以外の様々な情報が付加されており，計算機が言語データから常識を理解・獲得するための障壁となっている．類似度に基づく共起の尺度である PHSIC 応用することでこの問題に教師なしで対処できることを示す．

手法として，知識の共起スコア（PHSIC スコア）の総和を最大化するように観測データの枝刈りをおこなう．ただし探索空間が極めて大きいため，マルコフ連鎖モンテカルロ法に基づく確率的な山登り法をおこなう．

実験では人工データおよび実データで以上の手法が効果的に働くことを確認した．

**第７章　結論**

論文全体のまとめを示す．