中国語における帰納的推論の計算モデルの構成 —名詞と形容詞及び名詞と動詞の関係を用いて— The Construction of Computational Model for Inductive Reasoning in Chinese-Using the Relationships between Nouns and Verbs also Nouns and Adjectives-

張寓杰[1]，孫星越[2]，菊地賢一[1]，中川正宣[2]
Yujie Zhang, Xingyue Sun, Kenichi Kikuchi, Masanori Nakagawa

[1]東邦大学理学部情報科学科，[2]東京工業大学大学院社会理工学研究科
Toho University, Tokyo Institute of Technology
zhang@sci.toho-u.ac.jp

Abstract

In previous studies, we constructed a computational model of inductive reasoning based on the probabilistic concept structure estimated by the statistical analysis of large scale Chinese language data. However, in this model, only the relationship between nouns and verbs were used. In order to improve the precision of the model, the relationships between nouns and adjectives should also be analyzed and included in the model.

In this study, "adjectives" refer to Chinese noun modifiers that function as adjectives. Using these relationships, the new computational model of inductive reasoning is constructed based on the statistical analysis of large scale Chinese language data. Then, the validity of the model is verified using the psychological experiment.

Furthermore, for the comparison of simulation results between the previous model and the present one, we input the same positive premises and negative premises in the both model. The results show that erroneous outputs of the previous model are improved successfully in the present one.

Keywords — Inductive Reasoning, Statistical Language Analysis, Computational Model

1. 研究目的

帰納的推論とは，初期の観察や命題に対して意味情報を増加させる結論を導く思考である．つまり，いくつかの個別知識から，一般法則を導き出す推論を意味する．帰納的推論は単に科学的推論に限らず，広く日常生活でも用いられることが多い, 極めて基本的な人間の思考過程の一つである．たとえば，楠見孝(1998)によれば，類推も厳密な論理規則に基づく演繹に対して，類似性に基づく柔軟な推論としても位置づけられ，知識を拡張したり，形成する推論として，広義の帰納的推論の一つとして位置づけられる．

心理学や認知科学の分野では特定の推論形式を用いた帰納的推論の実験が広く行われてきた．たとえば，以下のような推論形式が広く一般的に用いられる(Rips,1975; Osherson, Smith, Wikie, Lopez, and Shafir,1990; Sakamoto &Nakagawa 2007,2008,2010)．

A さんはステーキが好きである．(正事例前提)
A さんはうどんが好きではない．(負事例前提)
A さんはハンバーグが好きである．(結論)

A さんはステーキが好きである．(正事例前提)
A さんはうどんが好きではない．(負事例前提)
A さんはそばが好きである．(結論)

この形式では，線分の上部が前提命題で，線分の下部が結論命題である．Osherson(1990)は，この種の形式における推論を「カテゴリに基づく帰納的推論(category-based induction)」という仮説に基づき考察した．この例では，A さんはハンバーグが好きであるという結論のもっともらしさは，かなり高いと判断される．反対に，A さんはそばが好きであるという結論のもっともらしさはかなり低いと言える．この場合を Osherson のカテゴリ仮説を簡略化して説明すると,「A さんはステーキが好き」という一つの事例から,「A さんの好きなもの」がステーキの属する「洋食」というカテ

ゴリに一般化され，A さんは，同じ洋食というカテゴリに属するハンバーグが好きという結論のもっともらしさが高くなる．同じように，A さんはうどんが好きではないので，和食が好きではないと一般化され，和食の一種であるそばが好きであるという結論のもっともらしさが低くなると説明されるわけである．

一方, Hadijichristidis， Sloman, Stevenson, Over(2004)は上記と同じ形式における推論を「属性 の帰納的推論(property induction)」という立場から，主に類似性に基づく帰納的推論の仮説を考察した．この仮説に基づく場合，上記の例は，ステーキとハンバーグは「材料は肉である」,「フライパンで焼く」等の属性を共有することにより類似性が高くなり，その結論のもっともらしさが高くなると説明される．

帰納的推論の心理学的メカニズムを説明するために，今までにさまざまな計算モデルが提案されてきた(Rips, 1975; Osherson, 1990; Sloman, 1993; Sanjana, 2002)．しかし，これらのモデルは，全て共通して，非常に限られた知識領域のみを対象とした帰納的推論以外は検証していないという問題点を含んでいる．

坂本(Sakamoto & Nakagawa 2007, 2008, 2010)は以上の既存のモデルの問題点に対して，心理実験による評定に依存せず，大規模言語データの統計解析を用いて数万語を含む確率的言語知識構造を構成し，より広範な概念についての予測が可能な帰納的推論の計算モデルを構築した．ただし,坂本の研究はすべて日本語に限られており，日本語以外での計算モデルの可能性については，全く考慮されていない．この問題点を考慮し，張ら(2013)は日本語と中国語大規模言語データの統計解析を用いて，両言語の帰納的推論の計算モデルを構築し，両言語の背景にある文化や社会システムの共通性と差異を比較した．しかし，この研究で構築された帰納的推論の計算モデルは名詞と動詞の関係しか用いていない．

本研究の目的は，中国語における名詞と動詞の関係に名詞と形容詞（形容詞の役割を担っている「名詞修飾語」と定義する）の関係を加え，大規模言語データの統計解析に基づき，確率的言語知識構造を再構築して，中国語における帰納的推論の計算モデルを構成し，心理学実験によりモデルの妥当性を検証することである．さらに，本研究で構築した新しいモデルと先行研究のモデルを比較し，シミュレーションの結果が改善されていることを検証する．

2. 研究方法

中国語の言語データとして以下の表 1 に示したコーパスを用いる．これらのコーパスはすべて一般公開されており，新聞記事や文学作品を含んでいて，政治，経済，社会，スポーツ，犯罪，あるいは文学，芸術等，中国語のさまざまな言語知識領域をカバーするこができる．

表 1. 本研究で用いた中国語コーパス
(サイズ: 651.44MB)

コーパスの種類	サイズ (MB)
ChineseTreebank4.0(2010 取得)	2.34
人民日報タグ付きコーパス(1998)	23
新京報電子版(2010 取得)	21.1
文学作品の電子テキスト(2010 取得)	605

本研究ではまず上記のデータの係り受け解析の結果得られた,「形容詞と名詞」,「名詞(目的語)と動詞」,「名詞(主語)と動詞(述語)」の各対について，全言語データ中の共起頻度を計算する．次に各対の共起頻度に基づき，Kameya & Sato (2005)の方法を用いて各対の共起確率と各条件付き確率，潜在クラスの確率の最尤値を推定する．ここで「形容詞と名詞」,「名詞(目的語)と動詞」,「名詞(主語)と動詞(述語)」の各対について推定された条件付き確率と潜在クラスの確率の総体を確率的言語知識構造と呼ぶ.

最後に推定された各確率，すなわち中国語の確率的言語知識構造に基づき，以下のような計算モデルを構成する．

$$v\left(N_i^c\right)=a\,SIM_+\left(N_i^c\right)+b\,SIM_-\left(N_i^c\right)-h \quad (1)$$

$$SIM_+\left(N_i^c\right)=\sum_j^{n^+} e^{-\beta d_{ij}^+} \quad (2)$$

$$SIM_-\left(N_i^c\right)=\sum_j^{n^-} e^{-\beta d_{ij}^-} \quad (3)$$

$$d_{ij}^+=\sqrt{\sum_k^m\left(P\left(C_k|N_i^c\right)-P\left(C_k|N_j^+\right)\right)^2} \quad (4)$$

$$d_{ij}^-=\sqrt{\sum_k^m\left(P\left(C_k|N_i^c\right)-P\left(C_k|N_j^-\right)\right)^2} \quad (5)$$

ただし，このモデルでは正事例 N_j^+と負事例 N_j^-が与えられた時の帰納的推論の結論 N_i^c のもっともらしさを $v(N_i^c)$として上記の数式に従い計算する．

ここで，d_{ij}^+と d_{ij}^-は，それぞれ結論の対象 N_i^cと正事例の対象 N_j^+，結論の対象と負事例の対象 N_j^-の距離で，各名詞を与えた時の各潜在意味クラスの条件付確率 $P(C|N)$ (名詞の潜在意味クラスへのメンバーシップ値)に基づいて計算される．m はその距離の計算に用いた潜在意味クラスの数で，ここにおける距離とは，各潜在意味クラスへのメンバーシップ値から構成される m 次元の特徴空間において計算されたユークリッド距離である．また，$SIM_+(N_i^c)$と $SIM_-(N_i^c)$は，各々正事例と結論，負事例と結論の対象間の距離 d_{ij}^+と d_{ij}^-を変数とするカーネル関数 $e^{-\beta d_{ij}^+}$, $e^{-\beta d_{ij}^-}$の和で定義され，各々，結論と正事例，結論と負事例の類似性の大きさを表している．a，b は各々，正事例と負事例の重みづけパラメータである．また β は距離，d_{ij}^+と d_{ij}^-の変化が類似性の大きさの変化にどの程度反映するかを示す相対的感度と考えることができる．

本研究ではこのモデルの形式を用いて，中国語の帰納的推論の 8 課題を用いてシミュレーションを行った．さらに，構築された計算モデルの妥当性を検証するため，この 8 課題の正事例，負事例と結論の単語を使い，各課題に対して具体的な質問を設定し,「かなりあり得る～まったくありえない」の 5 段階評定を用い，中国人の被験者 38 名に対してアンケート調査の心理学実験を実施し，シミュレーション結果と実験結果を定量的に比較し，計算モデルの妥当性を実証した．

3. 結果

本研究で構築したモデルと先行研究のモデルのシミュレーション結果の比較として，以下の 2 つの例を挙げる．例 1 は表 2 の正事例と負事例を入力し，出力結果の尤もらしさの値の上位 10 個の単語を抽出した比較結果である．

表 2. 本研究と先行研究のシミュレーション結果の比較（例 1）

正事例	日本語訳
芭蕾	バレー
绘画	絵画
負事例	日本語訳
物理学	物理学
科学	科学

上位 10 個（本研究）	上位 10 個（日本語訳）	上位 10 個（先行研究）	上位 10 個（日本語訳）
芭蕾	バレー	绘画	絵画
绘画	絵画	芭蕾	バレー
诗词	詩	防风林	防風林
交谊舞	社交ダンス	感叹号	感嘆符
音乐	音楽	连江	台湾にある町
作曲家	作曲家	本体主义	存在論
舞蹈	ダンス	古体诗	古体詩
油画	油絵	国医	中国医学
歌坛	音楽業界	练习曲	練習曲
乐曲	楽曲	古兰经	クルアーン

本研究の結果は全体的に「芸術」潜在クラスに属する単語が出力されている一方，先行研究の結果では，「防風林」，「感嘆符」，「台湾にある町」など，「芸術」潜在クラスに全く関係ない単語が含まれている．

例 2 は表 3 の正事例と負事例を入力し，出力結果の尤もらしさの値の上位 10 個の単語を抽出して比較する．

表 3. 本研究と先行研究のシミュレーション結果の比較（例 2）

正事例	日本語訳
电影	映画
动画片	アニメ
負事例	日本語訳
篮球	バスケットボール
足球	サッカー

上位 10 個（本研究）	上位 10 個（日本語訳）	上位 10 個（先行研究）	上位 10 個（日本語訳）
电影	映画	动画片	アニメ
动画片	アニメ	电影	映画
电视剧	ドラマ	武侠小说	武侠小説
中篇小说	中篇小説	反光镜	リフレクター
中青年	中青年	心窍	認識や思考の能力
全集	全集	频道	チャンネル
熟路	Eudora Welty の代表作	顿河	ドン川
日记	日記	活剧	アクションドラマ
影片	映画	下款	支払い
巨制	大作	舌苔	舌苔

本研究の結果は全体的に「マスコミュニケーション」潜在クラスに属する単語が出力されているが，先行研究の結果では，「リフレクター」，「認識や思考する能力」，「舌苔」など，「マスコミュニケーション」潜在クラスに全く関係ない単語が出力されている．少なくとも，これら二つの例から，本研究のモデルは先行研究よりモデルの精度が良くなったと言える．

一方，心理学実験結果のデータから，「かなりあり得る」を５，「あり得る」を４，「わからない」を３，「ありえない」を２，「全くありえない」を１として各結論ごとに被験者の平均を算出した．さらに各課題ごとに評定の平均値とシミュレーション結果(a＝1，b＝-1，h=0，β＝1 の場合)の相関係数を算出した．

表 4 に示すように，当該モデルのシミュレーション結果と心理学実験における評定平均値との相関係数は実験で用いた 8 課題ともに高い値を示しており，すべて検定結果も 1 パーセント水準で有意である．この結果から，本研究のモデルの心理学的妥当性が実証されたと言える．

表 4. モデルのシミュレーション結果と心理学実験における評定平均値との相関係数（:p<.01）**

課題	相関係数
課題 1（身分）	0.807523233**
課題 2（衣料）	0.487119396**
課題 3（交通）	0.751105613**
課題 4（趣味）	0.694709752**
課題 5（会議）	0.772780203**
課題 6（業界）	0.841989676**
課題 7（商品）	0.860115335**
課題 8（場所）	0.809506622**
平均	0.666141158**

4. 今後の予定

今後はコーパスの量と種類を拡張し，中国語の帰納的推論の計算モデルを再構築する．また，本研究で構築された中国語のモデルと日本語のモデルを比較する．

参考文献

[1] Kameya, Y., & Sato, T. (2005) “Computation of probabilistic relationship between concepts and their attributes using a statistical analysis of Japanese corpora”, Proceedings of Symposium on Large-scale Knowledge Resources, 65-68.

[2] Kayo Sakamoto, Asuka Terai, Masanori Nakagawa, (2007) "Computational models of inductive reasoning using a statistical analysis of a Japanese corpus", Cognitive Systems Research, 8, 282-299.

[3] Kayo Sakamoto, Masanori Nakagawa, (2008) “A Computational Model of Risk-Context-Dependent Inductive Reasoning Based on a Support Vector Machine”, T. Tokunaga and A. Ortega (Eds.): LKR2008, LNAI 4938, Springer-Verlag Berlin Heidelberg, pp.295-309.

[4] Kayo Sakamoto, Fang Xie, Masanori Nakagawa, (2010) “Syntactic Dependency Analysis Reveals Semantic Concept Structure Underlying Inductive Reasoning: Towards a Domain-Inclusive Structure that Enables Context-Dependent Knowledge Selection”, Cognitive Studies, Vol.17, No.1, 143-168.

[5] Osherson, D. N., Smith, E.. E.., Wilkie, O., López, A.., and Shafir, E., (1990) “Category based induction”, Psychological Review, 97, 185-200.

[6] Rips, L. J., (1975) “Inductive judgment about natural categories”, Journal of Verbal Learning and Verbal Behavior, 14, 665-681.

[7] Sloman, S., A., (1993) “Feature based Induction”, Cognition, 49, 67-96.

[8] 張寓杰, 寺井あすか, 董媛, 王月, 中川正宣, (2013) “日本語と中国語における帰納的推論の比較研究―言語統計解析に基づく計算モデルを用いて―”, 認知科学, No. 20 Vol. 4, 439-469.

[9] 楠見孝, (1998) “世界大百科事典(第２版)”, 平凡社.