

日本語における帰納的推論の計算モデルの拡張と 検索システムへの応用

The Extension of Computational Model for Inductive Reasoning in Japanese and the Application to a Search Engine

張 寓杰¹, 寺井 あすか², 上西 秀和³, 菊地 賢一¹, 中川 正宣⁴
Yujie Zhang, Asuka Terai, Hidekazu Kaminishi, Kenichi Kikuchi, Masanori Nakagawa

¹東邦大学理学部情報科学科, ²公立ほこだて未来大学システム情報科学部

³獨協医科大学基本医学情報教育部門, ⁴大妻女子大学人間生活文化研究所

Toho University, Future University Hakodate, Dokkyo Medical University, Otsuma Women's University
zhang@sci.toho-u.ac.jp

Abstract

In the previous study, we constructed a computational model of inductive reasoning based on the probabilistic concept structure estimated by the statistical analysis of large scale Japanese language data. However, in this model, only the relationships 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, the relationships not only between nouns and verbs, but also nouns and adjectives were used to construct a new computational model of inductive reasoning based on the statistical analysis of large scale Japanese language data. Then, the validity of the model was verified using the psychological experiment.

Furthermore, a search engine of inductive reasoning was constructed by PHP language, which simulate human's inductive reasoning process and can be applied in other many fields.

Keywords — Inductive Reasoning, Computational Model, Search Engine

1. 研究背景と研究目的

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。本研究では心理学や認知科学の分野で、広く一般的に用いられている以下の帰納的推論の形式を取り扱う(例えば Rips, 1975; Osherson, Smith, Wikie, Lopez, and Shafir, 1990; Sakamoto & Nakagawa, 2007, 2008, 2010 など)。

前提:

Aさんはステーキが好きである。(正事例)

Aさんはうどんが好きではない。(負事例)

結論: Aさんはハンバーグが好きである。

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。この形式の帰納的推論に関しては、従来さまざまな理論が提唱されている。たとえば 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、Sloman(1993)は類似性に基づく理論を提案している。これらの帰納的推論の各仮説では前提事例からカテゴリへの一般化や共有属性からの単語間の類似性の推定といった何らかの内的な心理学的メカニズムを想定している。しかし、そのような内的なメカニズムを直接心理学実験だけで実証することは困難である。帰納的推論の内的メカニズムを説明するために、今までに様々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、心理実験の評定に基づき構成されており、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。

心理実験の評定だけで、人間の持つ膨大な知識を構成する数えきれない概念間における類似度や特徴の関連強度を推定することは極めて困難である。坂本(Sakamoto & Nakagawa 2007, 2008, 2010)は、以上の問題点に対し、日本語の大規模言語データの統計解析に基づいて確率言語知識構造を構成し、広い知識領域での帰納的推論の計算モデルを構築した。さらに、張ほか(2013)は中国語の帰納的推論の計算モデルを構築して、このモデルの日本語以外での有用性を明らかにし、日本語と中国語における帰納的推論の計算モデルのシミュレーション結果を比較した。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。

一方、張ほか(2015)は、上記の中国語の計算モデルにおいて名詞と動詞の関係に名詞と形容詞の関係を加

え、モデルを拡張し、心理学実験によりモデルの妥当性を検証した。

本研究の目的は、張ほか(2013)の研究で構築した日本語の計算モデルにおける、日本語の名詞と動詞の関係に名詞と形容詞の関係を加え、計算モデルを拡張し、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、今までの研究成果の発展的応用性として、帰納的推論の検索システムをプログラミング言語 PHP により構成する。

2. 研究方法

2.1 モデルの構成

帰納的推論の計算モデルを構成するため、まず、日本語の大規模言語データの統計解析を用いて、名詞と動詞、名詞と形容詞の共起頻度から潜在意味クラスを抽出し、名詞、動詞、形容詞、各々との確率的関係を構成する(今回抽出した潜在クラス数は 800)。ここで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。次に、このように構成された確率的言語知識構造における潜在意味クラスを特徴次元と考え、各名詞の特徴強度を、その名詞が与えられた時の各潜在意味クラスの条件付き確率で定義し、以下のようなカーネル関数に基づく計算モデルを構成する。

$$v(N_i^c) = a\text{SIM}_+(N_i^c) + b\text{SIM}_-(N_i^c) - h \quad (1)$$

$$\text{SIM}_+(N_i^c) = \sum_j^{n^+} e^{-\beta d_{ij}^+} \quad (2)$$

$$\text{SIM}_-(N_i^c) = \sum_j^{n^-} e^{-\beta d_{ij}^-} \quad (3)$$

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

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

ここで $v(N_i^c)$ は結論(N_i^c)の尤もらしさの値、 $P(c_k|N_i^c)$ 、 $P(c_k|N_j^+)$ 、 $P(c_k|N_j^-)$ は各々結論(N_i^c)、正事例(N_j^+)、負事例(N_j^-)が与えられたときの潜在クラス(c_k)の条件付確率を示している。また、 d_{ij}^+ は、正事例の潜在意味クラスの条件付確率 $P(c_k|N_i^c)$ 、 $P(c_k|N_j^+)$ に基づいて計算される結論 N_i^c と正事例 N_j^+ の距離であり、 d_{ij}^- は、負事例の潜在意味クラスの条件付確率 $P(c_k|N_i^c)$ 、 $P(c_k|N_j^-)$ に基づいて計算される結論 N_i^c と正事例 N_j^- の距離である。 m は距離の計算に用いる潜在意味クラスの数で、対象が潜在意味クラスへのメンバーシップ値からなる m 次元の特徴空間において距離が計算される。 $\text{SIM}_+(N_i^c)$ と $\text{SIM}_-(N_i^c)$ は各々、結論と正事例、結論と負事例の類似性の大きさを表す。 a と b は、正事例と負事例の重みづけパラメータで、 β は d_{ij}^+ と d_{ij}^- が類似性の変化を反映する程度を表す相対的感度を示している。

本研究ではこのモデルの形式を用いて、日本語の帰納的推論のシミュレーションを行った。構築した計算モデルの妥当性を検証するため、8 課題のシミュレーションの結果に基づき、日本人の被験者 17 名に対して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検証する。

2.2 実験方法

構築した計算モデルの妥当性を検証するために、帰納的推論の心理学実験を実施した。実験材料は計算モデルのシミュレーションに基づき、張ほか(2013)の研究で使用した 8 組の課題の正事例と負事例を用い、結論には課題ごとにシミュレーション結果から上位 10 個、中位 10 個、下位 10 個の単語を抽出し、合計 30 個の単語を用いた。

この 8 課題の実験材料を使い、日本語を母語とする大学生と大学院生 17 名に対して、インターネットでのアンケート調査を実施した。調査での質問内容は課題ごとに異なっており、現実の様々な場面に対して、帰納的推論を行うように設定した。たとえば「趣味」課題の質問は、「ある人はバスケットボールとサッカーが好き。社会学と政治学が好きではない。この人が以下のものが好きである可能性に対して、どう思いますか？」である。結論の評定には「かなりあり得る～まったくありえない」の 5 段階評定を用いた。

3. 結果と考察

以下の表 1、表 2 に構築した確率的言語知識構造の

具体例を示す。表1は「形容詞-名詞」、表2は「名詞（目的語）-を-動詞（述語）」の組み合わせに対応した確率的言語知識構造の一例である。

表1 「形容詞-名詞」の確率的言語知識構造の一例：

人物関連クラス				
	形容詞	$P(c a)$	名詞	$P(c n)$
1	高名な	0.9323	首領	0.7739
2	優秀な	0.9051	領袖	0.6490
3	著名な	0.8913	同志	0.5610
4	偉大な	0.8292	マンネリズム	0.5524
5	有能な	0.8087	エコノミスト	0.5478
6	有為な	0.7431	外科医	0.5352
7	寡作な	0.6844	学者	0.5087
8	希有な	0.6220	人材	0.4932
9	凡庸な	0.4189	参謀	0.4532
10	けうな	0.3823	芸術家	0.4416

表2 「名詞（目的語）-を-動詞（述語）」の確率的言語知識構造の一例：会合関連クラス

	名詞	$P(c n)$	動詞	$P(c v)$
1	個展	0.9093	開く	0.9956
2	総会	0.8961	招集する	0.7119
3	胸襟	0.8840	欠席する	0.6759
4	リサイタル	0.8621	傍聴する	0.6286
5	例会	0.8450	中座する	0.6201
6	集会	0.8277	開会する	0.6016
7	会合	0.8256	受講する	0.5173
8	集い	0.7719	退席する	0.4816
9	会議	0.7666	開催する	0.4477
10	閣議	0.7424	催す	0.4332

表1にある $P(c|a)$ は形容詞が与えられた時の潜在クラスの条件付き確率であり、 $P(c|n)$ は名詞が与えられた時の潜在クラスの条件付き確率である。表2にある $P(c|v)$ は動詞が与えられた時の潜在クラスの条件付き確率である。これらの表では、特定の一つの意味クラスについて、名詞、動詞、形容詞ともに意味クラスの中に条件付確率の高い順に上位10個を示した。表1は「人物」に関連したクラス、表2は「会合」に関連したクラスであると推測できる。

表3は計算モデルのシミュレーション結果の一例である。「趣味」課題には、正事例の単語「バスケットボール」と「サッカー」を入力し、負事例の単語「社会

学」と「政治学」を入力し、出力結果のもっともらしさにより、「最上位十個」、「中位十個」、「最下位十個」合計30個の単語を抽出した。

表3 趣味課題のシミュレーション結果

	日本語	もっともらしさの値
最上位十個	バスケットボール	0.8972
	サッカー	0.8262
	野球	0.5245
	卓球	0.5195
	テニス	0.4969
	ラグビー	0.4852
	ゲーム	0.4844
	読書	0.4671
	スポーツ	0.4508
	ソフトボール	0.4365
中位十個	水泳	0.2635
	稽古	0.2576
	洗濯	0.2560
	整理	0.2330
	俳句	0.2323
	恋愛	0.2304
	雑談	0.2243
	歌舞伎	0.2048
	園芸	0.2038
	相撲	0.1423
最下位十個	経営学	-0.3840
	経済学	-0.3872
	工学	-0.3876
	病理学	-0.3998
	教育学	-0.4050
	地質学	-0.4091
	建築学	-0.4178
	心理学	-0.4299
	社会学	-0.7576
	政治学	-0.8704

一方、表4に示すように、本研究で構築した計算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、心理学実験で用いた8課題ともに高い値を示しており、すべての検定結果も1%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 評定の平均値とシミュレーション結果の
相関係数(**:p<.01)

課題	相関係数
課題1 (身分)	0.7356**
課題2 (衣料)	0.8462**
課題3 (商品)	0.8798**
課題4 (交通)	0.7263**
課題5 (会合)	0.6952**
課題6 (趣味)	0.8926**
課題7 (業界)	0.8613**
課題8 (場所)	0.9108**

4. 検索システムの構成

帰納的推論の検索システムを構成するため、ソフトウェア XAMPP を使い、PHP のローカル環境で検索システムを構成した。

The screenshot shows a web browser window at localhost/mysearch/research.php. The page title is '検索システム'. It has two sections: '正事例を入力してください' (Enter positive examples) with input fields for 'バスケットボール' and 'サッカー', and '負事例を入力してください' (Enter negative examples) with input fields for '社会学' and '政治学'. Below these is a field for '表示する単語数を入力してください (0: すべて表示)' (Enter the number of words to display) with '30' entered and a '実行' (Execute) button. The results section shows '正事例' (Basketball, Soccer) and '負事例' (Sociology, Political Science). A '計算結果は以下に表示されます' (Results are displayed below) section lists 14 items with their similarity scores, such as '1 バスケットボール 0.8972365' and '14 バレーボール 0.3878951'.

図1 検索システムの画面例 (単語入力と結果出力)

ローカルサーバで検索システムの画面を開くと、以上の図1に示した画面で正事例、負事例、表示する単語数の部分が空欄で入力待ち状態となる。次に任意の正事例の単語と負事例の単語を入力する。たとえば、

本研究で使った「趣味」課題に、課題の内容「ある人はバスケットボールとサッカーが好き。社会学と政治学が好きではない。この人はどんなものが好きか？」という事態を想定し、図1に示したように、正事例の単語「バスケットボール」と「サッカー」を入力し、負事例の単語「社会学」と「政治学」を入力する。さらに出力する単語数、例として30を入力し、「実行」ボタンをクリックすると、出力結果が画面に表示される。

検索システムに用意された二万以上の単語から任意の単語を選んで、正事例、負事例として入力すると、それに対応して人間の帰納的推論を模擬する結果が出力される。どんな組み合わせで、どんな結果が出力されるのかは簡単に予測できず、応用上、大変興味深いシステムと考えられる。この検索システムはまだ試作段階であり、現在はローカル環境に限定されている。

5. 今後の予定

今後はシミュレーションでの課題の種類や実験参加者数を増やし、システムの妥当性と検索システムや教育システムへの応用を視野に入れて、システムの有用性も検証していきたい。さらに、検索システムの実用化を目指して、ローカル環境だけではなく、一般のインターネット上に接続できるようにする予定である。

参考文献

- [1] Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 411-422.
- [2] 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.
- [3] 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.
- [4] 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
- [5] 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.

- [6] Osherson, D. N., Smith, E. E., Wilkie, O., López, A., and Shafir, E. (1990). Category based induction. *Psychological Review*, 97, 185-200.
- [7] Rips, L. J. (1975). Inductive judgment about natural categories. *Journal of Verbal Learning and Verbal Behavior*, 14, 665-681.
- [8] Sloman, S., A. (1993). Feature based Induction. *Cognition*, 49, 67-96.
- [9] Sanjana, N., & Tenenbaum, J. B. (2002). Bayesian models of inductive generalization. In: Becker, S., Thrun, S., Obermayer, K (Eds.) *Advances in Neural Information Processing Systems*, 15. Cambridge, MA: MIT Press.
- [10] 張寓杰, 寺井あすか, 董媛, 王月, 中川正宣. 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. 2013.
- [11] 張寓杰, 孫星越, 菊地賢一, 中川正宣. 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を用いて—. *日本認知科学会第 32 回大会発表論文集*, 472-476, 2015.