中国語の言語統計解析に基づく帰納的推論の計算モデルとその実験的検証

A Computational Model of Inductive Reasoning based on Statistical Analysis of Chinese Language Data and its Experimental Verification

張寓杰[1] 董媛[1] 王月[1] 寺井あすか[2] 中川正宣[1]

Zhang Yujie, Dong Yuan, Wang Yue, Asuka Terai, Masanori Nakagawa

[1]東京工業大学大学院社会理工学研究科 [2]東京工業大学グローバルエッジ研究院

Tokyo Institute of Technology

zhang.y.ag@m.titech.ac.jp, asuka@nm.hum.titech.ac.jp, nakagawa@nm.hum.titech.ac.jp

Abstract

The purpose of the present research is to propose a model of human inductive reasoning using the statistical analysis of Chinese linguistic data. The current research deals with induction involving predicates about which subjects have few prior beliefs. An example is "*Prof. N loves wine, therefore he also loves champagne,"* when you have little idea what Prof. N loves. Sloman (1993) proposed an alternative model in which the inductive reasoning is based on features of the arguments, (the feature-based model). This model is constructed using the result of psychological evaluation for the relationship between arguments and their attributes chosen ahead. However, it is difficult to objectively choose all features which cover the general knowledge of human-beings necessarily in order to construct the general model simulating the inductive reasoning process of human-beings.

In order to avoid those problems for construction of the general model of inductive reasoning, we have already developed a computational model of inductive reasoning using kernel functions based on statistical analysis of large scale Japanese language data (Sakamoto & Nakagawa, 2008). In this study, we constructed the same computational model for Chinese language based on statistical analysis of large scale Chinese language data. Furthermore, the validity of the model is verified from the comparison between the result of model's simulation and that of the psychological experiment.

Keywords: Inductive Reasoning, Statistical Language Analysis, Chinese Language

1．研究目的

坂本は日本語の言語統計解析を用いて推定された日本語の確率的言語知識構造に基づき Rips タイプの帰納的推論の計算モデルを構成している(Sakamoto & Nakagawa 2007, 2008, 2010)。この研究では、同じタイプの課題を用いて帰納的推論の心理学実験を実施し、実験結果から計算モデルのパラメータ推定を行い、モデルのシミュレーション結果と実験結果の比較を通じて、計算モデルの妥当性を実証している。しかし、これらの一連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。

本研究の目的は、中国語の言語統計解析を用いて推定した確率的言語知識構造に基づき、中国語帰納的推論の計算モデルを構成し、日本語帰納的推論の計算モデルでの結果との比較を行うことで、このモデルの多言語への拡張可能性を検討することである。

2．研究方法

1. 中国語言語データに対して、中国語の文法解析ツールを用いて係り受け解析を実施する。
   中国語言語データとしては以下のようなコーパスを用いる。
   ① ChineseTreebank4.0(2010) (40 万語)
   ②人民日報タグ付きコーパス (730 万語)
   ③新京報電子版 (26 万文)
   ④文学作品の電子テキスト (830 万文)
   中国語の文法解析ツールとしては CNP パーサーを利用する。CNP パーサーは、情報通信研究機構が、新たに開発した高精度の中国語文法解析システムである（2012 年 8 月公開。 正確度:93. 45%)。
2. 係り受け解析の結果得られた、「名詞(目的語)と動詞」、「名詞(主語)と動詞(述語)」について、全言語データ中の共起頻度を計算する。
3. 各共起頻度データに基づき、以下の関係を仮定し EM 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。

$$P(n_i, v_j) = \sum_k P(n_i \mid c_k) P(v_j \mid c_k) P(c_k)$$

表 1-1 名詞(主語)と動詞(述語)

競争

名词	名詞(日本語)	P(c|n)
马拉松	マラソン	0.794703
选手	選手	0.753383
男单	男子	0.7144
歌咏	歌	0.694411
跳台	プラットホーム	0.67102
自由泳	クロール	0.627971
双打	ダブル	0.623174
田径	陸上競技	0.589775

动词	動詞(日本語)	P(c|v)
决赛	決選する	0.911376
大赛	試合する	0.886439
比赛	試合する	0.855792
辈出	輩出する	0.786033
开赛	開始する	0.781426
出线	アウトする	0.75201
速滑	スピードスケートする	0.745346
演出	ショーをする	0.739362

表 1-2 名詞(目的語)と動詞

食べ物

名词	名詞(日本語)	P(c|n)
早饭	朝ごはん	0.961824
晚饭	晩ごはん	0.961413
午饭	昼ごはん	0.951058
饭	ご飯	0.905891
便饭	ファーストフード	0.85376
中饭	昼ごはん	0.845695
白食	無料のご飯	0.825327
茶	お茶	0.822573

动词	動詞(日本語)	P(c|v)
吃	食べる	0.971722
呷	飲む	0.929668
饱餐	食べ終わる	0.830018
煎	フライーする	0.799878
冰镇	凍える	0.688354
享用	食べる	0.581294
接风	御馳走する	0.55522
畅饮	飲む	0.538926

ここで、$P(n_i,v_j)$は名詞 n_i と動詞 v_j の共起確率、$P(n_i|c_k)$と $P(v_j|c_k)$は各々潜在クラス c_k が与えられた時の名詞 n_i と動詞 v_j の条件付確率、$P(c_k)$は潜在クラス c_k の出現確率である。このようにして推定された「名詞(目的語)と動詞」、「名詞(主語)と動詞(述語)」の各対についての条件付き確率と潜在クラスの確率の全体を確率的言語知識構造と呼ぶ。

4. 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、中国語の帰納的推論の計算モデルを構成する。計算モデルは以下のようなカーネル関数に基づき構成される。

$$v(\text{結論}) = a\,\text{SIM}_+(\text{結論}) - b\,\text{SIM}_-(\text{結論})$$
$$\text{SIM}_+(\text{結論}) = e^{-\beta d(\text{結論,正事例})}$$
$$\text{SIM}_-(\text{結論}) = e^{-\beta d(\text{結論,負事例})}$$
$$d(\text{結論,正事例}) = \sum_k^m \left(P(c_k \mid \text{結論}) - P(c_k \mid \text{正事例})\right)^2$$

ここで、v(結論)は「結論」のもっともらしさの値、SIM_+(結論)は「結論」と「正事例」の類似性、SIM_-(結論)は「結論」と「負事例」の類似性。a, b はバランスパラメータ。$P(c_k \mid x)$は上記 3 で計算される、x が与えられた時の潜在クラス c_k の条件付確率。

5. 中国人の被験者に対して中国語の帰納的推論の心

理学実験を実施する。

6. 心理学実験の結果から計算モデルのパラメータを推定し、心理学実験で用いた帰納的推論と同じ課題を用いてコンピューターシミュレーションを実施する。
7. 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する。

3. 結果

3.1 言語統計解析の結果

まず中国語言語データに対してパーサーを用いて係り受け解析を行い、計算した「名詞目的語と動詞」、「名詞主語と動詞述語」の共起頻度に基づき、各々について、潜在クラス数 100 で、確率的言語知識構造を推定した。表 1 は推定された中国語の確率的言語知識構造の例である。

3.2 モデルのシミュレーション結果

さらに、推定した中国語の確率的言語知識構造を用いて試験的に中国語の帰納的推論の計算モデルを構成した。表 2 は構成された計算モデルを用いたシミュレーション結果の例である。

4. 心理学実験:に基づくモデルの妥当性の検証

以下では、中国人の被験者に対して中国語の帰納的推論の心理学実験を実施し、心理学実験で用いた帰納的推と同じ課題を用いてコンピューターシミュレーションを行い、シミュレーション結果と実験結果を定量的に比較して、計算モデルの妥当性を実証する。

4.1 実験方法

実験参加者　中国人留学生　10 名

実験手続き　中国語のタイプの帰納的推論課題を用いた。実験参加者は前提条件の正事例 2 例、負事例 2 例に対して結論のもっともらしさを 5 点法で評定した。前提条件の課題は二種類(表 1.1 と表 1.2 の正負事例参照)で、各々結論として前提条件の対象も含めて 30 個の名詞を用いた。事前に正事例 2 例だけを用いて上記のモデルに基づきシミュレーションを行い出力された上位 15 個の名詞と、負事例 2 例を正事例として用いてシミュレーションを行い出力された上位 15 個の名詞、計 30 個の名詞を選んだ(表 1.1 と表 1.2 の中国語(日本語)の列参照)。

表 2.1 シミュレーション結果の例 1

名詞	もっともらしさ
妈妈(ママ)	1.0
爸爸(パパ)	1.0
叔叔(おじさん)	0.966910388838
大嫂(おばさん)	0.963034050626
姑姑(おばさん)	0.933177446243
阿姨(おばさん)	0.922247158229
⋮	
弟(弟)	0.24522162711
孩(子ども)	0.244590519477
老人(年寄り)	0.233567936442
仆人(使い物)	0.232179862338

表 2.2 シミュレーション結果の例 2

名詞	もっともらしさ
主席(議長)	1.0
部长(部長)	1.0
秘书长(秘書長)	0.916555157592
国防部长(防衛大臣)	0.916503085377
总理(首相)	0.894662461933
外交部长(外務大臣)	0.851249820774
⋮	
会长(会長)	0.318788086147
村干部(村の幹部)	0.31493375489
研究员(研究員)	0.31493375489
元首(首長)	0.304489644294

4.2 実験結果とシミュレーション結果の比較

表 3.1 と表 3.2 の「実験結果」の列はそれぞれ 2 種類の帰納的推論課題に対する各結論のもっともらしさの評定値の被験者平均値である。一方、各表の「モデル」の列は各結論のもっともらしさのシミュレーション出力値である。評定平均値とモデルのシミュレーション出力値の相関係数は各々、0.8826327 (p<0.001) と 0.896913 (p<0.001)である。

5. 考察

上記の二例とも実験結果とモデルのシミュレーション結果の相関は非常に高く検定結果も有意で、今回、中国語の言語統計解析に基づき構成された中国語の帰納的推論の計算モデルの心理学的妥当性が実証されたといえる。

表 3.1

正事例　股票(株)、国债(国債)

負事例　篮球(バスケット)、足球(サッカー)

	中国語	日本語	モデル	実験結果
13	股票	株	0.8183276	4.8
2	国债	国債	0.8146168	4.5
22	证券	証券	0.517546	4.5
25	外汇	外国為替	0.7096766	4.5
29	债券	債券	0.980222	4.4
4	原油	原油	0.7830529	3.7
23	商品	商品	0.7238232	3.7
9	原料	原料	0.647182	3.6
20	药品	薬品	0.5519925	3.6
6	产品	製品	0.7617016	3.5
16	石油	石油	0.7505095	3.5
7	木材	木材	0.7179974	3.4
12	军火	兵器弾薬	0.6753893	3.4
17	天然气	天然ガス	0.6190384	3.4
27	食品	食品	0.6099854	3.4
18	川剧	四川オペラ	−0.490158	3
5	棋类	将棋類	−0.474397	2.9
11	杂技团	雑技団	−0.493367	2.9
19	花剑	フェンシング	−0.448911	2.7
1	排球	バレーボール	−0.59284	2.6
8	羽毛球	バドミントン	−0.649435	2.6
10	垒球	ソフトボール	−0.501156	2.5
21	围棋	碁盤	−0.528429	2.5
26	马球	ポロ	−0.594769	2.5
15	乒乓球	卓球	−0.481206	2.4
24	网球	テニス	−0.614507	2.3
30	曲棍球	ホッケー	−0.550458	2.2
14	橄榄球	ラグビー	−0.556186	2.1
28	篮球	バスケット	−0.886222	1.7
3	足球	サッカー	−0.792885	1.4
			相関	0.8826327

表 3.2

正事例　制服(制服)、汗衫(シャツ)

負事例　帽子(帽子)、手套(手袋)

	中国語	日本語	モデル	実験結果
30	制服	制服	0.8912055	4.8
13	汗衫	シャツ	0.7488062	4.6
1	衬衫	ブラウス	0.6720483	4.6
5	套装	スーツ	0.59487	4.5
18	工作服	作業服	0.6305069	4.3
20	单衣	裏を付けない服	0.6505309	4
17	风衣	コート	0.6061598	4
2	礼服	礼服	0.6945908	4
29	长袍	長い中国服	0.72165	3.9
27	洋服	洋服	0.6793778	3.9
8	夹克	ジャケット	0.6081554	3.8
23	连衣裙	ワンピース	0.6291586	3.7
14	法衣	法衣	0.6572005	3.6
26	短裤	ショート	0.7119687	3.3
15	内裤	パンツ	0.6528138	2.9
25	袖章	バッジ	−0.64288	2.6
12	花冠	花冠	−0.63831	2.6
21	假发	ヅラ	−0.692374	2.5
22	皇冠	王冠	−0.594625	2.4
9	王冠	王冠	−0.723054	2.4
24	耳环	イアリング・ピアス	−0.646601	2.3
7	护目镜	ゴーグル	−0.627146	2.3
3	安全帽	ヘルメット	−0.6802	2.3
4	面具	マスク	−0.627073	2.2
11	口罩	マスク	−0.661991	2.1
10	手套	手袋	−0.858649	2
28	风帽	フード	−0.60989	1.8
16	帽子	帽子	−0.798284	1.6
6	草帽	麦わら帽子	−0.697012	1.3
19	黄帽	黄色い帽子	−0.463267	1.2
			相関	0.8969913

今後は、同じ中国言語データを用いて「形容詞と名詞」の対に関する共起頻度を計算し既に推定した中国語の確率的言語知識構造を拡張する。次に、拡張した中国語の確率的言語知識構造に基づき、新たな中国語の帰納的推論の計算モデルを構成する。

最終的には、これらの中国語での結果を、日本語での結果と比較し、このモデルの多言語への拡張可能性を検討する。

参考文献

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.

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

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.