言語統計解析に基づく計算モデルを用いた文章の適切性判断に関する検討

An examination of aptness in a sentence decision using computational models based on a statistical language analysis

白水優太朗[1]，寺井あすか[2]，王婉瑩[3]，中川正宣[4]，
Yutaro Shiramizu, Asuka Terai, Wanying Wang, Masanori Nakagawa

[1] 東京工業大学社会理工学研究科，[2] 公立はこだて未来大学システム情報科学部，[3] 清華大学人文学院，
[4] 大妻女子大学人間生活文化研究所

[1]Tokyo Institute of Technology, [2]Future University Hakodate, [3]Tsinghua University,
[4]Otsuma Women's University

aterai@fun.ac.jp

Abstract

Selection restrictions are co-occurrence constraints or possibilities which are observed between given lexical items. Although a sentence is grammatically correct, it is hardly interpreted (e.g. “The typhoon attacks happiness.”) with violations of selection restriction. In the present study, three types of computational models are constructed based on corpora that estimate aptness of a noun as the Object or the Subject in the sentence “Subject-Object-Verb”. One model is a Bayes model and the two others are network models. Psychological experiments are conducted to elucidate which model estimate the aptness most successfully. Comparison among the results of psychological experiment and model simulation endorses that the Bayes model brings about best performance.

Keywords — Sentence generation, Language statistical analysis

1. はじめに

本研究では、「主語 (S) が目的語 (O) を動詞 (V)」という形式の文章を対象とし、文章の適切性がどのように判断されているかについて、言語統計解析に基づく計算モデルの構築・シミュレーションにより検討する。上記のような文章は、主語、目的語、動詞にそれぞれ単語を当てはめることで「台風が町を襲う」といった、文を生成することが可能である。しかし、「台風が喜びを襲う」のように、文法的には正しいが意味的に解釈が不可能なものが存在する。これは「台風は一般的に場所や建物を襲うことはあるが、感情は襲わない」という、知識に基づいた判断が行われているためであり、このような動詞が主語と目的語の組み合わせに要求するルールを、動詞の選択制限と呼ぶ。 中本・黒田（2005）は、動詞の選択制限の成立過程を、意味フレームを用いて表現し、「S が O を襲う」という文における主語（S）と目的語（O）の選択制限における意味フレームの存在を心理実験により明らかにした。さらに、言語統計解析を用いて意味フレームを「S が O を襲う」という文における意味フレームの推定が可能であることが示された（永山 2007）。さらに、「襲う」以外の様々な動詞に対応可能な文生成モデルとして、主語と動詞を与えることで目的語としてふさわしい名詞を推定可能な計算モデルが構築されている（堀田他 2012, 2013）。しかし、先行モデルでは「主語・動詞が与えられた際に目的語としてどのような名詞がふさわしいか（目的語選択）」、という点にのみ着目されており、「目的語・動詞が与えられた際に主語としてどの名詞がふさわしいか（主語選択）」、に関する検討は行われていない。そこで、本研究では目的語、動詞を与えることで主語としてふさわしい名詞を推定可能な計算モデルを構築し、心理実験結果とシミュレーション結果の比較を通じて文章の適切性判断の認知メカニズムを検討した。

2. 文生成モデル

はじめに、モデル化の対象となる動詞 v_l における主語（S）、目的語（O）の共起関係を抽出することで、確率的知識構造の構築を行った。すなわち、「S が O を v_l」という文における主語（S）、目的語（O）、動詞 v_l の共起頻度データを、毎日新聞 18 年分（1991 年-2008 年）、小学校国語教科書、ブログデータ等から抽出した。抽出された係り受け頻度データに対し Naive Bayes Clustering (NBC, Kameya & Sato 2005)

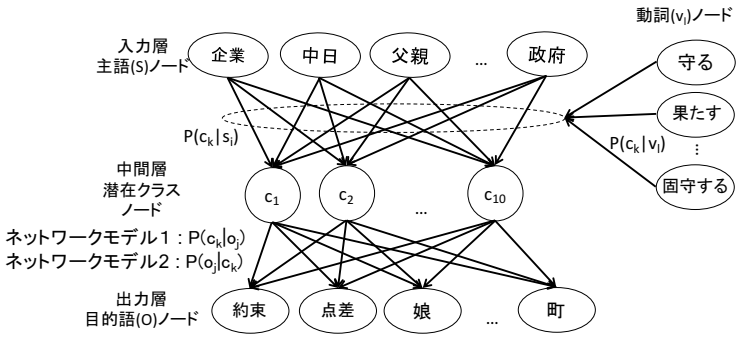

図 1　ネットワークモデル 1・2（目的語選択課題）（例：「父が O を守る」）

を用いて潜在クラスの推定をした。この手法では主語 s_i、目的語 o_j, 動詞 v_l の共起確率を潜在クラス c_k を用いて以下の式 (1) によって決定されると仮定し、EM アルゴリズムを用いて $P(c_k)$, $P(s_i|c_k)$, $P(o_j|c_k)$, $P(v_l|c_k)$ を推定する。本研究では、潜在クラスの数を 10 として推定を行った。

$$P(s_i, o_j, v_l) = \sum_k P(c_k)P(s_i|c_k)P(o_j|c_k)P(v_l|c_k) \tag{1}$$

さらに推定された確率値に対し Bayes の定理を用いて $P(c_k|s_i)$, $P(c_k|o_j)$, $P(c_k|v_l)$ を求める。

これらの確率値を用いて目的語選択課題・主語選択課題の 2 種類の課題に対応する計算モデルを各課題に対しネットワークモデル 2 種とベイズモデルを構築した。目的語選択課題に対する 2 種類のネットワークモデルは 3 層構造からなっており（図 1）、入力層に主語（S）と動詞（V）を入力することで、各名詞（o_j）が目的語として用いられた際の文章「S が o_j を V」の適切性を出力する。ネットワークモデル 1 の出力値 $x_c1(o_j)$ は式（2）、ネットワークモデル 2 の出力値 $x_c2(o_j)$ は式（3）により表される。

$$x_c1(o_j) = \sum_k P(c_k|s_i)P(c_k|o_j)P(c_k|v_l) \tag{2}$$

$$x_c2(o_j) = \sum_k P(c_k|s_i)P(o_j|c_k)P(c_k|v_l) \tag{3}$$

ベイズモデルの出力 $x_b(o_j)$ は式（4）により推定される。

$$x_b(o_j) = P(o_j|s_i, v_l) \propto \sum_k P(s_i|c_k)P(o_j|c_k)P(c_k|v_l)P(c_k) \tag{4}$$

ベイズモデルによる推定値は「主語 (S) が目的語 (O) を動詞 (V)」という形式での主語 s_i、目的語 o_j、動詞 v_l の共起頻度を反映した値となっているが、Naive Bayes Clustering（Kameya&Sato 2005) において式（1）を仮定し潜在クラス c_k の推定を行う際、平滑化を行うことで 0 頻度問題の解消を行う。そのため、言語統計解析結果を用いて条件付き確率を推定することで、「主語 (S) が目的語 (O) を動詞 (V)」における 3 単語共起頻度をそのまま用いることで生じる 0 頻度問題を解消した推定を可能としている。

同様に、主語選択課題に対するモデルとして入力層に目的語（O）と動詞（V）を入力することで、各名詞（s_i）が主語として用いられた際の文章「s_i が O を V」の適切性を出力する 3 種類のモデルを構築した。出力値は式（1）〜（3）の o_j を s_i に、s_i を o_j に変換することで、主語選択課題における文章の適切性が推定される。本研究では、日本語能力試験に含まれる動詞を選択し、選択された動詞と似た意味を持つ動詞を分類語彙表 (2004) に従ってグループ化を行うことで、モデル化の対象となる動詞グループを作成し、各動詞グループを対象として計算モデルを構築した。モデル構築に用いた動詞グループは全 39 グループ、総計 861 語の動詞が含まれている。

目的語選択課題に対するモデルのシミュレーション結果を表 1、主語選択課題に対するモデルのシミュレーション結果を表 2 に示す。

表 1　目的語選択課題に対するモデルのシミュレーション結果 (() 内の数値は、出力値 $x_{c1}(o_j)$、$x_{c2}(o_j)$ または $x_b(o_j)$).

	動詞：守る					
	ネットワークモデル 1		ネットワークモデル 2		ベイズモデル	
	父親が o_j を守る	中日が o_j を守る	父親が o_j を守る	中日が o_j を守る	父親が o_j を守る	中日が o_j を守る
1	安全 (0.234)	リード (0.167)	身 (0.012)	首位 (0.035)	身 (0.051)	首位 (0.203)
2	環境 (0.233)	首位 (0.165)	安全 (0.008)	リード (0.032)	安全 (0.033)	リード (0.185)
3	沈黙 (0.232)	得点 (0.165)	ルール (0.006)	トップ (0.008)	ルール (0.026)	トップ (0.049)
4	人権 (0.232)	点差 (0.163)	環境 (0.006)	得点 (0.006)	環境 (0.026)	得点 (0.038)
5	日本 (0.232)	大量点 (0.159)	沈黙 (0.006)	座 (0.006)	沈黙 (0.023)	座 (0.035)

表 2　主語選択課題に対するモデルのシミュレーション結果 (() 内の数値は、出力値 $x_{c1}(s_i)$、$x_{c2}(s_i)$ または $x_b(s_i)$).

	動詞：運ぶ					
	ネットワークモデル 1		ネットワークモデル 2		ベイズモデル	
	s_i が資材を運ぶ	s_i が幸運を運ぶ	s_i が資材を運ぶ	s_i が幸運を運ぶ	s_i が資材を運ぶ	s_i が幸運を運ぶ
1	鳥 (0.067)	鳥 (0.801)	人 (0.002)	人 (0.023)	容疑者 (0.017)	人 (0.027)
2	女中 (0.066)	風 (0.772)	風 (0.002)	風 (0.020)	業者 (0.014)	風 (0.024)
3	風 (0.065)	渡り鳥 (0.766)	容疑者 (0.001)	私 (0.015)	自衛隊 (0.011)	私 (0.017)
4	渡り鳥 (0.064)	親鳥 (0.766)	私 (0.001)	方 (0.013)	男 (0.011)	方 (0.015)
5	親鳥 (0.064)	コウノトリ (0.757)	方 (0.001)	ファン (0.009)	輸送機 (0.010)	ファン (0.011)

3. 実験

モデルの妥当性を検証するため目的語選択課題（実験 1)、主語選択課題（実験 2）を用いた 2 種類の実験を実施した。

実験 1 では、実験参加者は大学生・大学院生 86 名、動詞 8 種類、主語 8 種類を用い、各動詞・主語の 16 ペアに対し 24 種類の目的語となる名詞を用いて評定実験を実施した。実験参加者は「名詞が【　】を動詞」(例：「父が【　】を守る」）という文の目的語として提示された名詞（例：身, 得点）がどの程度ふさわしいかを 7 件法で評定してもらった。

実験 2 では、実験参加者は大学生・大学院生 103 名、動詞 8 種類、目的語 16 種類を用い、各動詞・目的語の 16 ペアに対し 24 種類の名詞を用いて評定実験を実施した。実験参加者は、「【　】が名詞を動詞」(例：「【　】が資材を運ぶ」）という文の主語として提示された名詞（例：業者、鳥）のふさわしさを 7 件法で評定してもらった。

計算モデルによる出力値と評定平均値の順位相関係数を表 3 に示す。目的語選択課題、主語選択課題の課題の違いに関わらずベイズモデルが最も高い相関 (0.49**, 0.41**; $^{**}p < .01$) を示している。

表 3　計算モデルの出力値と評定平均値の順位相関係数 (N=384)

	目的語選択課題	主語選択課題
ネットワークモデル 1	0.28	0.24
ネットワークモデル 2	0.48	0.36
ベイズモデル	0.49	0.41

4. 考察

本研究では、目的語選択課題、主語選択課題の 2 種類の課題を用いて「主語 (S) が目的語 (O) を動詞 (V)」という形式の文章の適切性がどのように判断されているかについて、計算モデルシミュレーションと心理実験を用いて検証した。その結果、課題に関わらずネットワークモデルと比較し、ベイズモデルがより人間の行う文生成を表現しており、主語・目的語・動詞の共起頻度に基づき文章の適切性が判断されている可能性が示唆された。

参考文献

[1] 中本敬子・黒田航, (2005) “意味フレームに基づく選択制限の表現：動詞“襲う”を例にした心理実験による検討”, 言語科学会第 7 回大会ハンドブック.
[2] 永山遼, (2007) “言語統計解析に基づく文生成メカニズムの計算モデル”, 東京工業大学社会理工学研究科人間行動システム専攻修士論文.
[3] 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, pp.65-68.
[4] 堀田崇史、木村玲奈、寺井あすか、中川正宣, (2012) “言語統計解析に基づく文生成の計算モデル構築”, 第 29 回日本認知科学会発表.
[5] 堀田崇史、寺井あすか、中川正宣, (2013) “言語統計解析に基づく文生成の計算モデルの実験による検討”, 第 30 回日本認知科学会発表.