言語統計解析に基づく文生成の計算モデル構築
Construction of Computational Models of Sentence Generation based on a Statistical Language Analysis

堀田崇史†, 木村玲菜‡, 寺井あすか†, 中川正宣†
Takafumi Hotta, Reina Kimura, Asuka Terai, Masanori Nakagawa

† 東京工業大学, ‡ 株式会社ゆうちょ銀行
Tokyo Institute of Technology, Japan Post Bank Co., Ltd
hotta@nm.hum.titech.ac.jp, asuka@nm.hum.titech.ac.jp, nakagawa@nm.hum.titech.ac.jp

Abstract

The purpose of this study is to construct a computational model which generates sentences, in the form of "subject (S) verb (V) object (O)" in Japanese. In this research, when a subject (S) and a verb (V) are input to the model, the model outputs the adequate nouns for the object (O) in the sentence. At first, we extracted modification relations "noun(S) - GA - noun(O) - WO - verb(V)". And, knowledge structure was estimated based on the language statistical analysis (Naive Bayes Clustering, Kameya & Sato 2005) using the modification relation data. In this research, we constructed to two types of models of sentence generation for a certain verb and three types of models for some verbs. The results of the models for some verbs show that two types of connectionist models are better than a Bayes model and it is difficult to simulate using Bayes' theory.

Keywords — Sentence Generation, Connectionist Model, Statistical Language Analysis

1. はじめに

本研究では、「主語(S)が目的語(O)を動詞(V)」という形式の文生成を対象とし、主語(S)、動詞(V)を入力することで目的語としてふさわしいと考えられる候補を出力する計算モデルを構築する。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ることで「SがOを襲う」という文における意味フレームを確率的に推定できることが示された（永山 2007)。そこで、本研究では「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、言語統計解析に基づき「SがOをV」における意味フレームを確率的に推定し、主語（S)、動詞（V）を入力することで 目的語（O）としてふさわしい名詞を出力する計算モデルを構築する。本研究では、1つの特定の動詞に関する文生成モデルとともに、類似した意味を持つ複数の動詞に関する文生成モデル構築を行う。

2. １つの動詞に関する文生成モデル

１つの特定の動詞vに関する文生成モデル構築を行う。

2.1 言語統計解析に基づく意味フレームの推定

はじめに特定の動詞vに対し、「主語（S）が目的語（O）を動詞v」文における主語（S）と目的語（O）の共起関係を抽出する。すなわち、動詞「守る」の場合、「SがOを守る」という文における主語（S）と目的語（O）の共起頻度データを、日本語作文支援システム「なつめ」(Bor et al. 2011)で用いられている抽出ルールに従い、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、青空文庫、辞書・辞典データ等から抽出した。次に、抽出された係り受け頻度データに対しNave Bayes Clustering (NBC, Kameya & Sato 2005)を用いて、動詞vにおける主語（S）と目的語（O）に関する潜在クラスの推定を行う。この手法ではある特定の動詞vに対する主語s_i、目的語o_jの共起確率を潜在クラスc_k^vを用いて以下の式(1)によって決定されると仮定し、$P(c_k^v)$,$P(s_i|c_k^v)$,$P(o_j|c_k^v)$を推

定する。

$$P(s_i, o_j) = \sum_k P(c_k^v)P(s_i|c_k^v)P(o_j|c_k^v) \quad (1)$$

潜在クラスc_k^vとs_i, o_jとの関連はBayesの定理を用いて計算される$P(c_k^v|s_i)$,$P(c_k^v|o_j)$により表される。動詞「守る」では、潜在クラスの数を3として分析を行い、「組織による利権の維持（例：中日が首位を守る）」「社会的ルールの遵守（例：国家が人権を守る）」「生活環境の保護（例：父親が家庭を守る）」を表現する潜在クラスが意味フレームとして推定された。推定された各潜在クラスと関連の強い主語、目的語上位5単語を表1に示す。

2.2 文生成モデル

これらの確率値を用いて３層構造のコネクショニストモデル、ベイズモデル（図1）を構築した。入力層の各ノードは主語となる名詞、中間層のノードは言語統計解析により推定される潜在クラス（意味フレーム）、出力層の各ノードは目的語となる名詞を表す。モデルは主語（s_{IN}）を入力することで、各名詞の目的語（O）としてのふさわしさを出力する。すなわち、動詞「守る」に関するモデルは、「【入力されたs_{IN}】がOを守る」という文におけるOとしての名詞o_jのふさわしさを出力する。入力層から中間層への結合過重値は$P(c_k^v|s_i)$により推定される。また、中間層から出力層への結合過重値として$P(c_k^v|o_j)$を用いたコネクショニストモデル、中間層から出力層への結合過重値として$P(o_j|c_k^v)$を用いることで条件付き確率$P(o_j|s_{IN})$を出力するベイズモデルの2種類を構築した。両モデル共に、入力された主語s_{IN}を表すノードに1をそれ以外を表す入力層のノードに0を入力する事で、出力層の各ノードはそれぞれが表す名詞o_jの目的語としてのふさわしさを出力する。コネクショニストモデルの出力値$x_c(oj)$は式（2）により,

$$x_c(o_j) = \sum_k P(c_k^v|s_{IN})P(c_k^v|o_j) \quad (2)$$

ベイズモデルの出力値$x_b(oj)$は式（3）により推定される。

$$\begin{aligned} x_b(o_j) &= \sum_k P(c_k^v|s_{IN})P(o_j|c_k^v) \\ &= P(o_j|s_{IN}) \end{aligned} \quad (3)$$

シミュレーション結果を、表2に示す。コネクショニストモデルはベイズモデルと比較し、より推定された潜在クラス（意味フレーム）を反映した結果を出力している。しかし、両者の間に大きな違いは見られなかった。

3. 複数の動詞に対する文生成モデル

１つのモデルでより多くの動詞を扱うことを可能にすることを目的とし、類似した意味を持つ複数の動詞を対象とした文生成モデル構築を行った。本研究では、分類語意表 (2004) に基づき、類似した意味を持つ動詞を抽出した。すなわち、動詞「守る」と最下層の分類において同じカテゴリに分類されている「果たす」「固守する」「厳守する」など計11種類の動詞v_lを対象とした。

3.1 言語統計解析に基づく意味フレームの推定

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）と目的語（O）の共起関係を抽出する。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、日本語作文支援システム「なつめ」(Bor et al. 2011)で用いられている抽出ルールに従い、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、青空文庫、辞書・辞典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NBC (Kameya & Sato 2005) を用いて主語（S）、目的語（O）、動詞v_lに関わる潜在クラスの推定を行う。この手法ではある主語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(4)によって決定されると仮定し、$P(c_k)$,$P(s_i|c_k)$,$P(o_j|c_k)$,$P(v_l|c_k)$を推定する。

$$P(s_i, o_j, v_l) = \sum_k P(c_k)P(s_i|c_k)P(o_j|c_k)P(v_l|c_k) \quad (4)$$

本研究では、潜在クラスの個数を10として推定を行った。推定された潜在クラスの中には、動詞「守る」に関連するクラスとして、動詞を「守る」のみに固定し主語（S）と目的語（O）に関わる潜在クラスの推定を行った際に抽出された「組織による利権の維持」「社会的ルールの遵守」「生活環境の保護」と同様の解釈が可能な3クラスが含まれていた。

3.2 文生成モデル

これらの確率値を用いて2種類の3層構造コネクショニストモデル（図2）を構築した。入力層は主語となる名詞を表すノードからなっており、中間層のノードは言語統計解析により推定される潜在クラス、出力層の各ノードは目的語となる名詞を表す。また、入力層から中間層への結合過重値に対し、動詞ノードから重み付けが行われる。このモデルは主語（s_{IN}）と動詞（v_{IN}）を入力するこ

表 1 言語統計解析結果（守る）((()内の数値は、$P(c_k^v|s_i)$または$P(c_k^v|o_j)$).

	組織による利権の維持		社会的ルールの遵守		生活環境の保護	
	主語	目的語	主語	目的語	主語	目的語
1	広島 (0.9871)	首位 (0.9985)	イラク (0.9904)	約束 (0.9968)	女性 (0.9908)	身 (0.9967)
2	中日 (0.9854)	リード (0.9985)	政府 (0.9902)	ルール (0.9945)	親 (0.9760)	家庭 (0.9850)
3	ダイエー (0.9854)	得点 (0.9925)	北朝鮮 (0.9882)	合意 (0.9936)	父親 (0.9719)	命 (0.9840)
4	日本ハム (0.9830)	日本 (0.9900)	自民党 (0.9862)	利益 (0.9921)	さんら (0.9705)	家 (0.9832)
5	オリックス (0.9830)	点差 (0.9891)	政権 (0.9837)	人権 (0.9910)	大蔵省 (0.9701)	体 (0.9813)

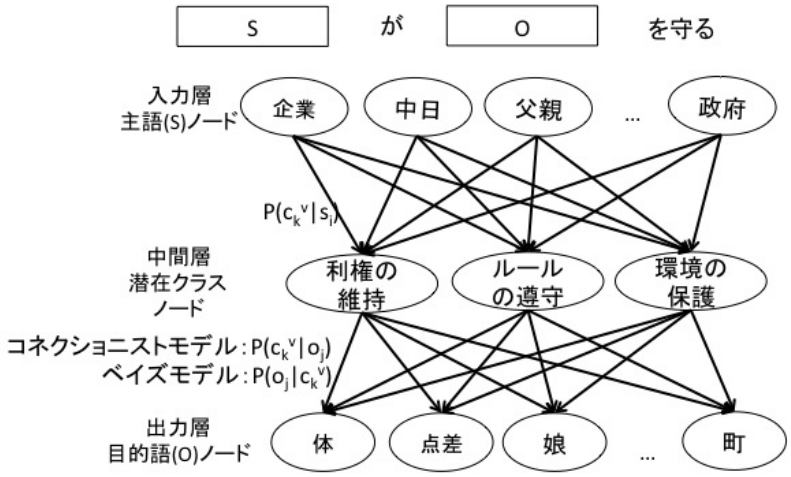

図 1 1つの動詞に関する文生成モデルのアーキテクチャ（動詞：守る）

表 2 文生成モデルのシミュレーション結果((()内の数値は、出力値$x_c(o_j)$または$x_b(o_j)$).

	コネクショニストモデル		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)
1	身 (0.9688)	首位 (0.9840)	身 (0.0769)	首位 (0.1633)
2	家庭 (0.9576)	リード (0.9840)	命 (0.0402)	リード (0.1521)
3	命 (0.9567)	得点 (0.9781)	沈黙 (0.02760)	安全 (0.0484)
4	家 (0.9558)	日本 (0.9757)	トップ (0.0257)	日本 (0.0316)
5	体 (0.9540)	点差 (0.9747)	子ども (0.0218)	座 (0.0296)

とで、各名詞の目的語（O）としてのふさわしさを出力する。すなわち、モデルは、「【入力されたs_{IN}】がOを【入力されたv_{IN}】」という文におけるOとしての名詞o_jのふさわしさを出力する。主語となる名詞を表すノードから中間層への結合過重値は$P(c_k|s_i)$により推定されるとともに、それらに動詞による重み付け$P(c_k|v_{IN})$が加味される。また、中間層から出力層への結合過重値を$P(c_k|o_j)$を用いて推定した コネクショニストモデル1、中間層から出力層への結合過重値を$P(o_j|c_k)$を用いて推定したコネクショニストモデル2を構築した。両モデル共に、入力された主語s_{IN}を表すノードに1をそれ以外を表す入力層のノードに0を入力する、また入力された動詞v_{IN}を表すノードに1をそれ以外の動詞ノードに0を入力する事で、出力層の各ノードはそれぞれが表す名詞o_jの目的語としてのふさわしさを出力する。

コネクショニストモデル1の出力値$x_{c1}(oj)$は式(5）により,

$$x_{c1}(o_j) = \sum_k P(c_k|s_{IN})P(c_k|v_{IN})P(c_k|o_j) \qquad (5)$$

コネクショニストモデル2の出力値$x_{c2}(oj)$は式(6）により推定される。

$$x_{c2}(o_j) = \sum_k P(c_k|s_{IN})P(c_k|v_{IN})P(o_j|c_k) \qquad (6)$$

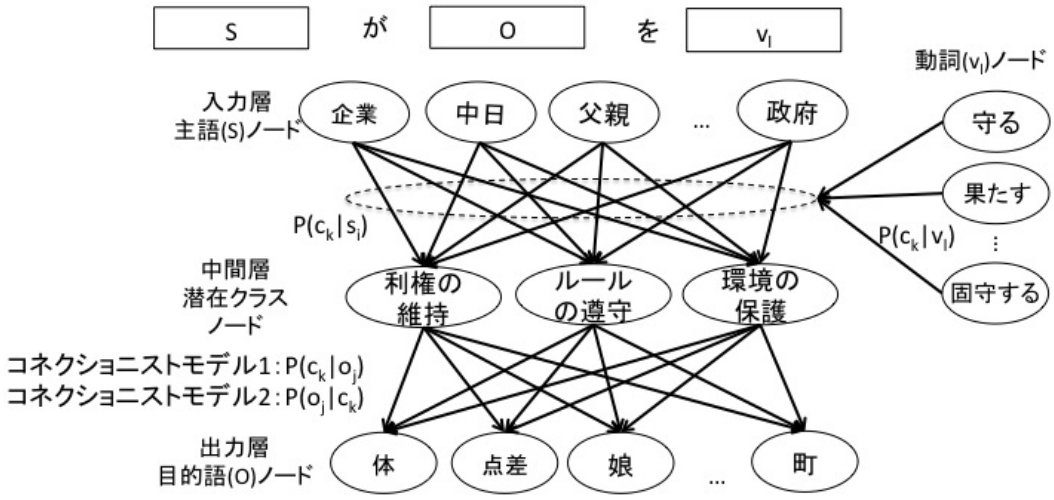

図 2 複数の動詞に関する文生成モデルのアーキテクチャ

表 3 文生成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					
	コネクショニストモデル1		コネクショニストモデル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	命 (0.3121)	リード (0.2005)	身 (0.0188)	首位 (0.0412)	役割(0.8484)	首位 (0.1763)
2	環境 (0.3121)	得点(0.2003)	安全 (0.0109)	リード (0.0377)	責任 (0.0140)	リード (0.1613)
3	伝統 (0.3119)	点差(0.2001)	命(0.0099)	約束(0.01367)	仲介役 (0.0111)	約束 (0.0613)
4	日本 (0.3119)	大量点 (0.1997)	環境 (0.0095)	トップ (0.0100)	貢献 (0.0075)	トップ (0.0427)
5	子ども(0.3119)	雇用 (0.1993)	ルール(0.0085)	それい(0.0083)	機能 (0.0067)	それ (0.0356)
	動詞：果たす					
	コネクショニストモデル1		コネクショニストモデル2		ベイズモデル	
	父親がo_jを果たす	中日がo_jを果たす	父親がo_jを果たす	中日がo_jを果たす	父親がo_jを果たす	中日がo_jを果たす
1	役割 (0.0446)	優勝 (0.0161)	役割 (0.0415)	優勝 (0.0063)	責任(0.1301)	首位 (0.2055)
2	仲介役 (0.0446)	連覇(0.0161)	責任 (0.0108)	連覇 (0.0018)	機能 (0.0623)	リード (0.1881)
3	けん引役(0.0445)	出場(0.0161)	機能(0.0052)	出場(0.0014)	役割 (0.0394)	約束(0.0681)
4	回復 (0.0444)	進出 (0.0161)	使命 (0.0008)	進出(0.0012)	身 (0.0314)	トップ(0.0498)
5	国入り(0.0444)	入賞 (0.0161)	仲間入り(0.0008)	再選(0.0009)	首位 (0.0285)	それ (0.0414)

また、比較としてベイズモデルとして主語s_{IN}、動詞v_{IN}が与えられた時の条件付き確率$P(o_j|v_{IN}, s_{IN})$を用いて、名詞o_jの目的語としてのふさわしさ$x_b(o_j)$を推定した（式(7)）。

$$x_b(o_j) = P(o_j|v_{IN}, s_{IN}) = \frac{\sum_k P(c_k)P(s_{IN}|c_k)P(v_{IN}|c_k)P(o_j|c_k)}{\sum_k P(c_k)P(s_{IN}|c_k)P(v_{IN}|c_k)} \quad (7)$$

それぞれの結果を、表3に示す。特定の動詞に対する文生成モデルと同様、コネクショニストモデル1が最も推定された潜在クラス（意味フレーム）を反映した結果を出力している。また、コネクショニストモデル1とコネクショニストモデル2の間に大きな差異はみられなかった。しかし、条件付き確率$P(o_j|s_{IN}, v_{IN})$を用いて推定を行ったベイズモデルでは、動詞の違いをきちんと反映した結果が得られなかった。

4. 考察

本研究では、言語統計解析に基づき意味フレームの推定を行うことで文生成モデルの構築を行った。1つの動詞に関する文生成モデルについては、コネクショニストモデルとベイズモデルのシミュレーション結果に大きな違いは見られなかった。動詞「襲う」に固定した場合のモデルでは、心理実験結果との比較を通じ、ベイズモデルに対しコネクショニストモデルがよりよく心理実験結果の推定を行うことが既に示されている（永山 2007)。また、1つのモデルでより多くの動詞を扱う事を目的とし、類似した意味を持つ複数の動詞に対応したモデル構築を行った。モデル構築では、主語、目的語、動詞の3種類の共起頻度データを用い、これらに関わる潜在クラスを抽出することで、類似した意味を持つ複数の動詞に関する意味フレームを確率的に推定した。さらに、3種類のモデルを用いてシミュレーションを行った結果、2種類のコネクショニストモデルと比較し、ベイズモデルでは動詞の違いを反映した結果が得られなかった。すなわち、ベイズ推定のみによるモデル構築が困難であることが示された。しかし、シミュレーション結果のみから2種類のコネクショニストモデルのどちらがより妥当であるかは不明である。今後は、心理実験を実施し、類似した意味を持つ複数の動詞に対応した文生成モデルとしていずれのモデルがもっとも妥当なものであるかを明らかにしたいと考えている。

謝辞

本研究はJSPS科研費若手研究（B）(23700160)の援助を受けて行われた。

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] Kameya, Y., Sato, T. (2005) Computation of probabilistic relationship between concepts and their attributes using a statistical analysis of Japanese corpora. Proc. of Symposium on Large-scale Knowledge Resources: LKR2005. 65-68.
[3] 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..
[4] Bor Hodoscek, 阿辺川武, Andrej Bekes, 仁科喜久子 (2011) レポート作成のための共起表現算出支援ー作文支援ツール「なつめ」の使用効果ー. 専門日本語教育研究. 13. 33-40.
[5] 国立国語研究所 編(2004) 分類語彙表-増補改訂版. 大日本図書刊