3P1-7in2

# 語の共起情報による概念・単語選択の改善 ―統合物語生成システムにおける利用―

# Revising Concepts and Word Selection based on Word Co-occurrence Information: The Use in an Integrated Narrative Generation System

照井和舎[*1] Kazuya Terui　小野淳平[*2] Jumpei Ono　小方孝[*1] Takashi Ogata

[*1] 岩手県立大学 Iwate Prefectural University

[*2] 岩手県立大学大学院 Graduate School of Iwate Prefectural University

An automatic narrative generation system called INGS that we have been developing selects linguistic concepts using the conceptual dictionaries for generating events in the story generation phase. For controlling the comprehensibility of concepts and the corresponding words, we have presented a method that uses the co-occurrence information of words. In this paper, we revise the method by using many linguistic data more than previous experimentation. In a result, we confirm that we can use the co-occurrence information for selecting the words having the same degree of comprehensibility.

## 1. はじめに

筆者らが開発を進めている統合語物語生成システム（Integrated Narrative Generation system: 以下, INGS）（総合的構想は[小方 10]，現状での実装の詳細は[Akimoto 14]や[Ogata 16]）において，物語の基本単位は事象と呼ばれる，基本的に一つの動詞概念と複数の名詞概念を含む格構造（現在形容詞概念等の利用は実質的に行っていない）であるが，その中の動詞概念や名詞概念を選択する機構の改善を行って来た．概念選択は基本的に INGS におけるストーリー生成機構が担当する．概略，二種の方式があり，一つはストーリー技法及びストーリーコンテンツ知識を用い，予めかなり限定された概念候補の中から一つの概念に絞り込むという方式であり，もう一つはこの方式により最低限決められた以外の概念選択を，動詞概念辞書や名詞概念辞書を用いて行うという方式である．動詞概念辞書及び名詞概念辞書は共にカテゴリーを中間節点とし個々の概念を終端節点とする階層構造であるが，このカテゴリーは現状ではかなり大きく精緻化されていない．また，これらの概念辞書は特定の物語の主題やジャンルに沿って構築された特殊なものではなく，どのような主題やジャンルの物語にも対応可能な一般的なものを企図しているため，様々な性質を持った概念（例えば一般の受け手にとって理解容易な概念や困難な概念）が一つのカテゴリーの中に混在している．上記二番目の方式による概念選択は一つもしくは複数のカテゴリーの中から無作為に一つを選択するが，上記のような理由により，一つの生成ストーリーの中に現れる概念に統一性がなくなる場合がある．

なお，INGS における概念辞書の項目は表層的な語彙にかなり対応する詳細なレベルで組織化・定義されている．例えば，「食べる 1」のような動詞概念が存在するが，これは「仕事をする（教師で食べる）」という意味（概念）を表現しており，「1」という末尾の数次は「食べ物を食べる」という別の動詞概念と区別するために設けられる．この「食べる 1」という動詞概念は文に変換される時，基本的にはそのまま「食べる」とされる．このように，概念はそのまま語彙表現に直結しており，概念の選択方式は語彙の生成時にも直接的な影響を与える．但し，表層的な文の生成に際しては言語表記辞書を用いて表記を変更することができるようになっている．例えば，「食べる」は「たべる」「タベル」等とも表記可能である．

さて以上のような問題を解決するためには，概念辞書のカテゴリー分類を再検討して改訂することが考えられるが，特定の観点から精緻化された体系は逆にその一般的・汎用的な使用を阻害する恐れもあるので，一つのアプローチとして概念辞書の構成自体は現状のままとして概念選択の方式自体に工夫を加えることを考えた．その一つは他のテクストのおける語彙の出現頻度情報を調べてその結果を概念辞書における各概念に割り当てその情報を概念選択の制御情報として利用するという方式であり，これについては[小野 14, Ogata 16, 吉田 16]で報告している．この場合は，基本的には，出現頻度情報が高い語彙（概念）は一般の受け手（読者）による理解容易性も高めるという調査結果が出ており，ここから一般の受け手による理解容易性を高めるためには出現頻度が高い概念（語彙）を用い，逆の場合は出現頻度が低い概念（語彙）を故意に用いるという戦術的知識（規則）が得られる．

もう一つがここで取り扱う語彙どうしの共起情報を利用する方式であり，これについてのこれまでの作業は[小野 15a, Ogata 16]で報告した．基本的に，共起情報の強い語彙どうしは同程度の理解容易度（困難度）を持つのではないかと考えた．この仮説に基づけば，ある生成物語における概念（語彙）における理解容易度を，上述の出現頻度の利用とは別の方式で制御することが可能となる．また上で述べた頻度情報について，頻度情報を得られなった概念に対し，共起情報を用いた推測を試みた[小野 15b, Ono 16]．本稿では，語彙どうしの共起情報を利用した概念選択に関してまず従来の成果をまとめるが，そこでの問題は，理解容易度の値は人間へのアンケート調査で行ったので，それを割り当てられた概念の個数が少なく，網羅的な調査を行えなかったということである．これに対して本稿では，単語難易度を各種の手法で調べた既存研究を利用して，より多くの概念（語彙）をカバーする調査を試みる．

連絡先：小方孝，岩手県立大学ソフトウェア情報学部，岩手県滝沢市巣子 152-52，t-ogata@iwate-pu.ac.jp

## 2. ストーリー生成方式及び共起情報を用いた概念選択

まず本研究の前提となる INGS のストーリー生成機構におけるストーリー生成の概略を紹介し，共起情報を用いた概念選択方式を説明する．

### 2.1 INGS におけるストーリー生成機構

INGS における物語生成のための三つの大きなモジュールは，ストーリー生成機構・物語言説機構・物語表現機構であり，事象生成における概念選択は主に最初のストーリー生成機構によって担われる．ここで選択された概念は上述のように生成される文を構成する語彙記述に受け継がれるが，その処理は物語表現機構の下位モジュールとしての自然言語生成機構で行われる．

ストーリー生成機構は，ストーリー技法及びストーリーコンテンツ知識ベース中のストーリーコンテンツ知識を用いてストーリーの構造を段階的に生成するが，その基本単位の一つの事象である．もう一つの基本単位に状態があるがこれについては[福田 14]を参照されたい（状態の中でも概念選択が行われる場合があるが，その方式は本稿で説明しているものと同じである）．ストーリー技法とは，一つの事象または複数の事象を含むストーリーの部分もしくは全体構造を入力として，複数の事象を含むストーリーの部分もしくは全体構造を出力する機構であり，これは事象どうしを結合する意味論的関係に基づいて幾つかのグループに分類されている．さらに，ある意味論的関係に基づいて特定の事象を拡張するための具体的な知識を保持し，これをストーリーコンテンツ知識と呼ぶ．事象概念生成では，あるストーリー技法で指定された動詞概念の格構造を元に，その格構造中の各名詞概念に付加された値の制約条件に基づき名詞概念が一つ選択される．この値の制約条件は名詞概念辞書における一つもしくは複数のカテゴリーの指定であり，これらのカテゴリー中に含まれた名詞概念の中から無作為に一つの概念が選択される．ストーリー生成機構はさらに，選ばれた名詞概念を元に，各名詞概念の性質や特徴を記録する属性フレームを参照し，ストーリーを構成する具体的な人・物・場所（インスタンス）を生成する．出現頻度情報や共起情報を利用した方式は，上記の「無作為に」に取って代わられる．

### 2.2 共起情報を利用した動詞・名詞概念の選択方式

INGS のストーリー生成において，ある事象が存在し，それに続く事象の動詞概念を決定する際，動詞概念どうしの共起情報を利用して，典型的にはより共起関係が強い動詞概念を選択し，それに基づいて事象を生成する．一方名詞概念どうしの共起情報は，生成される一つの事象の格構造における格の内容を決定する際に利用される．すなわち，ある格における名詞概念が決まると，例えばそれとより強い共起関係を持つ名詞概念が次に選択される．さらにまた，動詞概念と名詞概念の間の共起情報も利用される．すなわち，事象内の動詞概念と例えばより強い共起情報を持つ名詞概念が用いられる．このように，これまでの研究[小野 15a, Ogata 16]において，動詞概念どうし・名詞概念どうし・動詞概念と名詞概念との間の共起情報が，単独の事象及び事象の連鎖を生成する際に利用可能な仕組みを構築した．

語彙間の共起情報の算出には KH Coder[Higuchi 04]の機能を用いている．対象テクストに含まれる語彙（動詞及び名詞）どうしの共起情報を抽出し，それぞれの語彙を INGS の動詞概念辞書及び名詞概念辞書に含まれる概念と重ね合わせ，その語彙が持つ共起情報を概念の共起情報とした．なお動詞概念は複数の意味を持つ場合，数字によってその意味の区別を行っている．例えば，「食べる 1」は“任意の職業で生計を立てる”という意味であり，「食べる 2」は“物を食べる”という意味である．しかし，語彙だけからその意味を判断することはできないため，ここではそのような意味の違いは考慮していない．これについては今後の検討課題とする．今回，対象テクストは『青空文庫』に収録された 1872 年から 1963 年までの新字新仮名の全作品 4980 作品（2014 年 9 月時点．小説が中心．他，評論など含む．本文部分のみを利用）を利用した．結果として，動詞概念については全 11951 個中 4866 個に共起情報を設定し[Ogata 16]，名詞概念については全 115765 個中 76029 個に共起情報を設定した[小野 15a]．また，31125 個の名詞概念について動詞概念との共起情報を獲得・設定した[小野 15a]．この結果，小規模なデータを用いた実験により，共起関係が強い概念どうしは同程度の理解容易度（困難度）になり，弱い場合はそれとは逆転する傾向があることを確認することができた．本稿ではより多くのデータを用いてこの傾向を検証する．

なお 1 節で述べた頻度情報を用いた概念選択ついて，頻度情報が 0 である場合に共起情報を利用することを試みた[小野 15b, Ono 16]．これはまだ横光利一の小説 30 作品のみを使った小規模な実験であるが，0 頻度の概念と共起関係にある概念の頻度情報の平均値をその 0 頻度の概念の頻度情報とした．

## 3. 既存の単語難易度判定研究を利用した概念間共起関係の調査

### 3.1 方法

まず概念辞書に含まれる INGS の各概念辞書における名詞概念及び動詞概念について，チュウ太のレベルチェッカー[Kawamura 13]を利用して，難易度を判定する．この調査は[吉田 16]と共通であるが，表 1 に各判定規準とそれによって難易度判定が可能となった概念数を整理する．一方表 2 は，表 1 にある判定可能なすべての動詞概念及び名詞概念について，上のような意味で判定可能な概念を対象に，共起関係を持つすべての概念間の組み合わせを割り出した数字である．

表 1 判定基準と難易度判定可能な語（概念）の数

| 判定基準 | 動詞概念 | 名詞概念 |
|---|---|---|
| 朝日新聞記事における頻度情報に基づく分類（10 段階表示） | 2845 | 2155 |
| 朝日新聞記事における頻度情報に基づく分類（6 段階表示） | 2581 | 3421 |
| 旧日本語能力試験出題基準に基づく分類（6 段階表示） | 3066 | 6529 |
| 経済用語のレベルに基づく分類（7 段階表示） | 1914 | 2866 |
| 筑波大学との共同研究による新基準に基づく分類（6 段階表示） | 3940 | 8040 |
| 介護用語のレベルに基づく分類（4(5)段階表示） | 2733 | 5105 |
| 単語親密度に基づく分類（4 段階表示） | 4337 | 17509 |

各判定基準を用いて，二つの概念間での難易度の差と共起情報の強さをチェックする．強い共起関係にある概念どうしの難易度の差が小さく，弱い共起関係にある概念どうしではその逆である，というのが筆者らの仮定であり，従来の調査でもこのことを確認した[小野 15a, Ogata 16]．以下，カバーできる概念の数が最も多い「単語親密度に基づく分類」を中心に以下に結果と考察を示す．

表 2　判定基準と難易度判定可能な語(概念)の数

| 判定基準 | 動詞概念の組み合わせ数 | 名詞概念の組み合わせ数 |
|---|---|---|
| 朝日新聞記事における頻度情報に基づく分類(10 段階表示) | 228329 | 11080 |
| 朝日新聞記事における頻度情報に基づく分類(6 段階表示) | 202526 | 3866 |
| 旧日本語能力試験出題基準に基づく分類(6 段階表示) | 274768 | 3165 |
| 経済用語のレベルに基づく分類(7 段階表示) | 157518 | 1252 |
| 筑波大学との共同研究による新基準に基づく分類(6 段階表示) | 354172 | 9881 |
| 介護用語のレベルに基づく分類(4(5) 段階表示) | 225688 | 2718 |
| 単語親密度に基づく分類(4 段階表示) | 48969 | 34517 |

## 3.2　結果と考察

図 1 は上記「単語親密度に基づく分類」による理解難易度の判定結果について，難易度の差ごとに動詞概念どうしの組をまとめ，その共起情報の平均を取ったものである．左縦軸は概念どうしの共起関係の強さの平均を示しており，値が大きい程強い共起関係を持つことを示す．右縦軸は共起関係を持つ概念どうしの理解困難度の差の大きさを示す．「単語親密度に基づく分類」における理解困難度の段階は 4 つであるため，差の幅は 0 から 3 までである．図 2 は同じ手続きによる名詞概念の共起情報に関してのグラフである．これらの図より，共起関係が強い概念どうし程，難易度の差が小さい傾向にあることが分かり，この全般的傾向は従来のものと同様であった．

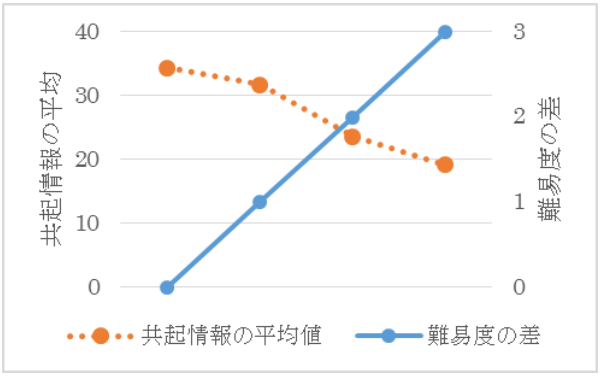


図 1　動詞概念における難易度の差と共起情報の平均値の関係

しかし，上記の傾向と一致しない事例，すなわち共起関係が強い概念どうしだが難易度の差が大きい事例や，その逆の事例も存在した．表 3 に，「単語親密度に基づく分類」におけるその事例を，動詞概念と名詞概念につきそれぞれ 10 ずつ示す．このような場合，具体的な語彙(概念)と共起関係を抽出して個別に対処する等の通常とは別の処理が必要となる．

他の判定基準に関しても，概ね上記と同様に，共起情報の平均と難易度の差が交差する形を取り，共起関係が強い概念どうし程難易度の差が小さい傾向にあった．INGS における利用に際しては，特定の難易度を持つ概念(語彙)に全体を統一する，

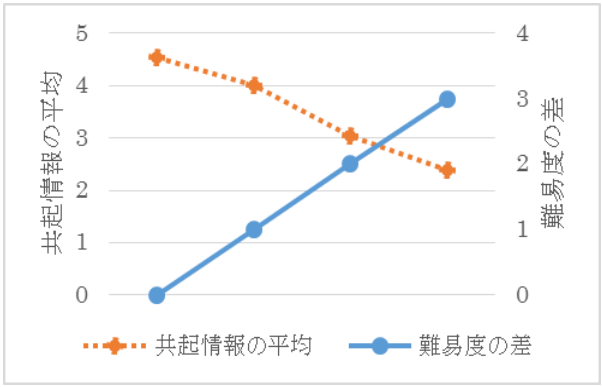


図 2　名詞概念における難易度の差と共起情報の平均値の関係

表 3　仮定にそぐわない事例(「単語親密度に基づく分類」)

| 種類 | 共起情報が大きいが難易度の差が大きい | 共起情報が小さいが難易度の差が小さい |
|---|---|---|
| 動詞概念 | 当たる-踏み外す，当たる-困り果てる，焼く-名付ける，焼く-呼び込む，役-償う，焼く-仕向ける，当たる-余す，焼く-出払う，焼く-絡まる，当たる-こす | 降下する-降りる，降下する-曇る，降下する-救う，降下する-考え込む，行進する-降りる，行進する-曇る，行進する-救う，行進する-考え込む， |
| 名詞概念 | ずっこけ-結局，ずっこけ-満足，ずっこけ-努力,ずっこけ-計画，ずっこけ-男子，アンパイア-顔，アンパイア-昨日，アンパイア-試合，イースター-会社，ガソリンスタンド-武者 | 夢中-スロー，女子-洗剤，課長-スペシャリスト，皮肉-堅実，奴等-再任，当人-他動詞，彼方-再任，警部-連打，警部-当て外れ，司令-原油 |

あるいは様々な難易度を持つ概念(語彙)を全体に散布させるといった制御方法に利用することができるだろう．

しかし，動詞概念の共起情報における「朝日新聞記事における頻度情報に基づく分類」及び「経済用語のレベルに基づく分類」については異なる結果が見られた(図 3，図 4，図 5)．これらでは，上記図 1 や図 2，あるいは省略した他の結果と比較して，難易度の差の大きさと共起情報の平均が反比例しない傾向が見られた．図 3 から図 5 の何れのグラフでも，中間よりやや低い難易度の差の箇所で，共起情報の平均のピークが見られた．これは，上述のように本研究において共起情報は『青空文庫』すなわち小説を中心とする文学作品から取得しているのに対して，上記二つの分類は新聞記事のデータに基づいている違いに起因するのではないかと推測される．すなわち，新聞記事において共起しやすい語彙関係において，難易度のレベルが異なるものが比較的多く含まれているからではないかと推測される．

## 4.　おわりに

筆者らが開発を進めている INGS では，事象を構成する動詞及び名詞概念及び対応するそれぞれの語彙の選択のために，テクスト資料における語彙どうしの共起関係を利用し，受け手における理解難易度と対応付ける方法を用いている．すなわち，共起情報の強い語彙どうしは同程度の理解難易度を持つと仮定し，これを INGS の事象を構成する概念選択（及びそれと対応する語彙選択）に利用することを方針とする．しかしこれまでの筆者らの研究ではこの対応付けのための理解難易度に関するデータが少なかったた

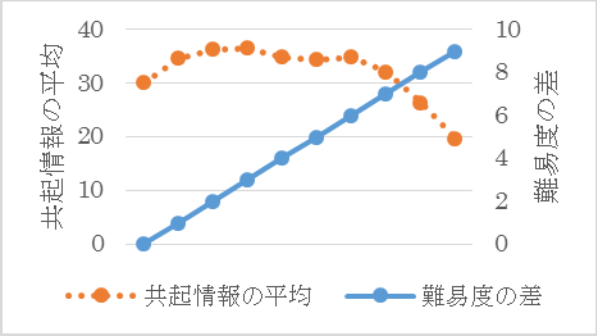


図 3　朝日新聞記事における頻度情報に基づく分類(10 段階表示)

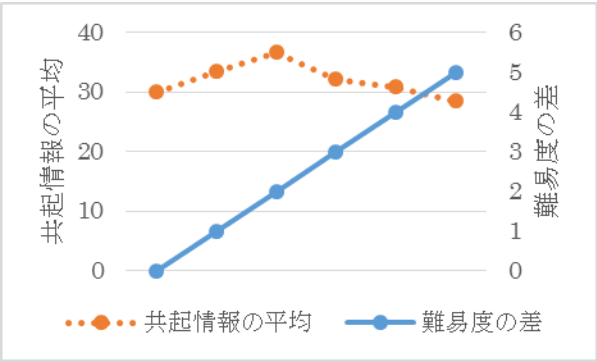


図 4　朝日新聞記事における頻度情報に基づく分類(6 段階表示)

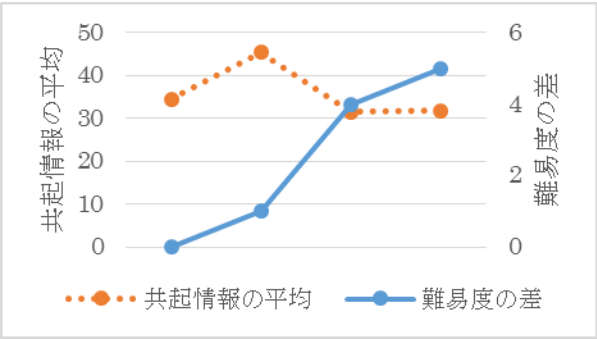


図 5　経済用語のレベルに基づく分類

め，本稿では既存の理解難易度判定ツールを用いてより多くのデータによる検証を行った．その結果，動詞概念どうしの場合も名詞概念どうしの場合も，概ね，語彙どうしの共起関係の大きさと理解難易度の平均値は比例関係にあるとの結果を得た．しかし一部両者が比例しない場合もあり，それは語彙難易度判定ツールで用いられた言語資料と本研究で共起関係のために用いられた言語資料との性質の違いと推測した．この推測の検証，動詞概念と名詞概念との間の共起関係の調査は今後の課題である．今後 INGS の中にこの方法を本格的に組み込んで行くことにするが，その際，上記のような言語資料の性格（ジャンル，時代等）等の要因も考慮する．資料とするテクストの変更によって異なる結果を得ることができれば，これを生成戦略の中に組み込むことも可能だろう．

## 参考文献

[Akimoto 14] Akimoto, T. and Ogata, T.: An Information Design of Narratology: The Use of Three Literary Theories in a Narrative Generation System, The International Journal of Visual Design, 7(3), 31-61 (2014)

[福田 14] 福田至，小方孝：統合物語生成システムにおける状態 - 事象変換知識ベースの現状と課題，人工知能学会全国大会（第 28 回）論文集，2F4-OS-01a-8in (2014)

[Higuchi 04] Higuchi, K.: Quantitative Analysis of Textual Data: Differentiation and Coordination of Two Approaches, Sociological Theory and Methods 19(1), 101-115 (2004)

[Kawamura 13] Kawamura Y.: The basic concept for multilingualization of the Reading Tutorial Dictionary Tool, The 8th Symposium on Japanese Language Education in Europe (2013)

[小方 10] 小方孝，金井明人：物語論の情報学序説―物語生成の思想と技術を巡って―，学文社 (2010)

[Ogata 16] Ogata, T. and Ono, J.:A Way for using the Verb Conceptual Dictionary in an Integrated Narrative Generation System: Focusing on the use of Co-occurrence Information on the Verb Concepts, Proc. of The 2016International Conference on Artificial Life and Robotics, 437-440 (2016)

[小野 15a] 小野淳平，小方孝：動詞概念と名詞概念の共起関係に基づく事象における名詞概念の選択―統合物語生成システムにおけるストーリー生成のための機構―，第 14 回情報科学技術フォーラム講演論文集 第 2 分冊，239-242 (2015)

[小野 15b] 小野淳平，小方孝：統合物語生成システムにおける概念選択／語彙表記選択及びその制御，第 29 回人工知能学会全国大会論文集，3G4-OS-05a-3 (2015)

[Ono 16] Ono, J. and Ogata, T.: A Way in Verb Concept Selection using Co-occurrence Information of Verb Concepts: A Mechanism in an Integrated Narrative Generation, Proc. of The 4th IIAE International Conference on Industrial Application Engineering 2016 (2006) (印刷中)

[吉田 16] 吉田和樹，小野淳平，小方孝：語の頻度情報による概念･単語選択の改善―統合物語生成システムにおける利用―，2016 年度人工知能学会全国大会（第 30 回）予稿集 (2016) (印刷中)