談話的な手がかりを利用した日本語の節の受動化

飯田 龍　　徳永 健伸
東京工業大学 大学院情報理工学研究科
{ryu-i,take}@cl.cs.titech.ac.jp

1 はじめに

可読性の高い文章を書くためには，文章中に出現する名詞句などの談話要素のつながりを考慮して適切に配置し，文章の結束性を高める必要がある．例えば，例 (1) と (2) では，その内容は命題レベルでは等しいが，2 文目で「太郎」と「大きな犬」のどちらが主語となるかが異なっている．

(1) a. 太郎は公園に行った．
　　b. 彼はそこで大きな犬に追いかけられた．

(2) a. 太郎は公園に行った．
　　b. 大きな犬がそこで彼を追いかけた．

ここで，この 2 文を生成する問題を考え，入力として「行く (ガ:太郎 $_1$, ニ:公園 $_2$)」「追う (ガ:大きな犬, ヲ:太郎 $_1$, デ:公園 $_2$)」[1]の 2 つの述語項構造を想定した場合，結束性の高い（つまり，1 文目で主題化されている「太郎」が 2 文目でも継続して主題化される）生成結果である例 (1) のような出力を得ることは，複数文書要約の出力生成などの応用で結束性の高い文章を生成するために必要不可欠である [6]．また，我々はこのような談話的な特徴を捉えた文章の記述・推敲支援を目指しており，それを実現するために結束性が高くなるように文章を生成する処理は必須な部分処理となる．

述語項構造などの表現形式から実際の文章を生成する研究は，古くから言語生成の研究，特に談話のプランニングに関する研究分野で研究が進められてきた [5]．ただし，既存研究の中で利用されている生成規則の多くは作り込まれた生成システムの中に存在し，明確にどのように談話の特徴を利用して生成しているかについての議論は比較的少ない．そこで，本研究では，上述のような共参照関係も含む述語項構造の集合が与えられた際の文章生成の課題を考える．

ここで考える文章生成の課題はおおきく 2 つの問題に分割できる．一つは，述語項構造に横断的に現れる共参照関係にある談話要素に関して，どの部分を名詞句として記述し，どの部分を代名詞に置きかえる，もしくは省略するかという参照表現生成の課題である．例えば，例 (1) の 1 文目で「太郎」がそれぞれ固有名で記述されているのに対し，1 文目で「太郎」が主題化されているため，2 文目ではそれらを生成する場合には結束性を高めるために代名詞「彼」として生成する必要がある．

もう一つの問題では，述語項構造の各項を前方の文脈に応じて適切な格で生成する必要がある．例えば，例 (1) の 2 文目では，主題の遷移に関してつながりを良くするために,「彼」を主語とし，述語「追いかける」は受動化して「追いかけられる」と生成する必要がある．一方，受動化の処理は，上述のような述語項構造を横断した談話要素の共有によって引き起こされ場合に加え，言語に応じた述語の利用法を加味する必要がある．例えば，動詞によっては受動態で利用されやすい，日本語では有生物と無生物の両方が同じ述語の項となる場合に有生物を主語として記述するなどの特徴を捉えた生成のモデルを考える必要がある．

そこで，本研究では，特に文脈中での述語の受動化に着目し，どのような文脈でどのような表現が受動化するのかを，談話的な特徴や述語そのものの選好など，3 種類の特徴に基づいてモデル化する．まず，2 節で本研究に関連する研究について述べ，3 節で受動化を説明する特徴について考える．次に，それらの特徴を学習するための素性について 4 節で紹介し，各特徴がどのくらい受動化に影響するかを調査した結果を 5 節で述べ，6 節でまとめと今後の課題について述べる．

2 関連研究

節の受動化の問題は言語生成の分野で主に研究が進められてきた．例えば，Abb ら [1] の研究では，受動化の処理は言語外の情報と言語的な情報の 2 つによって説明されている．言語外の情報とは，生成する文における動作主の背景化のような認知的な機能をもとに説明されるもので，例えば，容易に推測可能な動作主や，そもそも動作主が未知の場合など，動作主が不特定である場合に生成器が文を受動化することを許す．一方で，言語的な情報としては，発話を逐次的に生成する過程で，短期記憶の中で顕在化している対象を意味役割を決定する前に主語の位置に配置することによって，結果的に受動化した文が生成されるというものである．Abb らはこのような 2 つの観点から受動化の仕組みについて言及しているが，具体的な評価などは行っていないため，ここで述べられた受動化に関係する情報のうち実装可能なものは実装し，実際の受動化にどの程度影響するかを調べる必要

[1] 下付きの数字が同じ談話要素は共参照関係にあることを表す．つまり，入力として述語項構造の情報に加え，共参照関係の情報もわかっているという前提で処理を考える．

がある．

また，Abu Sheikha ら [7] は文語体や口語体を区別して文を生成するために，それぞれの典型的な特徴をまとめており，受動化に関しては口語体のほうが能動態で記述する傾向にあり，一方文語体の場合は口語体と比較して無生物主語で受動態で記述するという特徴について述べている．彼らは提案した生成のモデルを評価するために定量的な評価実験を行っているが，この評価ではあるユーザが指定した形式（文語体もしくは口語体）で文を生成し，それを別の被験者が文語体であるか口語体であるかの程度を評価するというもので，受動化そのものの評価とはなっていない．

3 受動化に関連する言語的特徴

述語の受動化については 1 節で述べたように，談話要素の共参照関係だけではなく，それ以外の言語的な特徴が影響する．本研究では特に (1) 受動化をともなう動詞そのものがそもそもどのくらい受動化しやすいのかといった語彙的な特徴，(2) 述語項構造間の共参照関係，(3) 動作主となる項の不特定性に関する情報の 3 種類の特徴がどう受動化に影響するかを説明し，それぞれの特徴をどう捉えるのかを説明する．

動詞の受動化の選好 動詞が受動化して生成される要因の一つとして，動詞そのものがどのくらい受動化して使用されるかが影響すると考えられる．例えば，動詞「立たす」は「(動作主をある状況に) 立たす」という用法については「(ある状況に) 立たされる」という表現が好んで使用されたり,「(ある対象に) 注目する」という用法は「(ある対象) に注目される」という言い回しが利用される傾向にある．これらの例は，新聞記事に固有の言い回しであるが，生成されるドメインごとに大規模に文章集合を収集することができれば，その選好を見積ることができると考えられる．

この選好を数値化するために，本研究では形態素解析済みの大規模コーパスから下記の式に基づいて選好のスコアを求める．

$$score_{pas}(v_i) = \frac{freq_{pas}(v_i)}{freq_{all}(v_i)} \log freq_{all}(v_i) \quad (1)$$

ここで，v_i は対象となる動詞，$freq_{all}(v_i)$ はコーパス中の v_i の出現頻度，$freq_{pas}(v_i)$ は v_i が「(ら) れる」をともなって出現した頻度を表す．つまり，動詞が受動化している割合をその出現頻度で重み付けた結果をその動詞の選好のスコアとする．動詞の語義ごとにこの選好が異なることが考えられるが，ここではそれを考慮せずに表記の一致でスコアを求めることとする．

項の意味的なカテゴリの情報 述語の項を適切な位置（日本語の場合は格）に生成するためには，その項の有生性が重要となると考えられる．つまり，述語が行為をともなう事態である場合，その動作主をできるだけガ格に配置し，相対的に無生物をガ格以外の位置に配置することで，間接的にその有生物の顕現性をさらに高め，以降の文脈における省略をより許容する方向へ生成を行う．この特徴を捉えるために，述語の項の意味カテゴリを利用する．具体的には，項が固有名である場合はその固有名のクラス（人名，組織名など），それ以外の普通名詞の場合はある概念体系に基づいた意味カテゴリの情報を用いることで，間接的に項として現れた名詞句の有生性の情報を導入する．

前方文脈の表現との共参照関係 生成対象となっている述語の項の中で前方文脈の表現と共参照関係となる場合，その項は旧情報となるため，主題化して文頭の位置に生成される傾向にある．このため，ガ格以外の項が旧情報に相当する場合であっても，その項は係助詞「は」で主題化され，文頭に配置されることになる．この結果，ガ格，ニ格，ヲ格という典型的な格の順序でこの文を生成しようとする場合，実際に埋まるべき述語の格とは整合しなくなるため，これを許容するために述語を受動化して生成することになる．例えば，1 節で示した例 (1) では,「太郎」が 2 文目で旧情報に相当するため，この表現を 2 文目で生成するために主題化して生成され，結果として動詞「追いかける」は「追いかけられる」と受動化されることとなる．そこで，述語のそれぞれの格に対し，前方に共参照関係にある表現があるかという情報を生成時に利用する．

また，既存研究 [1] でも言及されているように，述語の動作主が不特定の場合，動作主に対応する格（一般的にはガ格）への焦点化を取り消すために，述語を受動化することで他の格要素を焦点化する．つまり，述語の格パタンを想起した場合に，その格は埋まっているべきなのに，対応する表現が文章内に出現しない外界照応の場合には受動化して生成する必要が生じる．文脈や動詞の用法によってはそのまま原形で生成すべき場合もあるため，必ずしも受動化するべきではないが，ガ格が外界照応の関係にあるか否かはその述語を受動化して生成すべきかを捉える重要な手がかりとなる．

4 節の受動化モデル

3 節で示した手がかりを利用した受動化を実現するため，述語項構造と前方文脈を入力とし，その述語を受動化するか否かの 2 値分類問題を考える．3 種類の手がかりそれぞれを分類に利用するために，表 1 に示す素性集合を学習・分類に利用する．

3 節の式 (1) 述語の受動化の選好スコアの計算には，毎日新聞 91 年から 94 年，96 年から 2003 年の合計 12 年分を対象に CaboCha[1] を用いて形態素・係り受け解析を行い，その結果を利用して選好のスコアを計算した．また，有生性の情報を導入するために項の固有名情報と意味カテゴリを推定する必要があるが，固有名の情報としては CaboCha が出力する IREX[2] の 8 種類の固有名ラ

[1] http://code.google.com/p/cabocha/
[2] http://nlp.cs.nyu.edu/irex/index-j.html

表 1: 素性

素性タイプ	素性名	説明
pred	score_{pas}	式 (1) に示した動詞の受動化に関する選好のスコア
	lexical	述語の基本形の語彙項目
	func	述語を含む文節中の「(ら) れる」以外の機能語
	adnom	述語を含む文節が連体節に出現しているか否か
	first_sent	述語が文章の最初の文に出現しているか否か
	last_sent	述語が文章の最後の文に出現しているか否か
	sent_end	述語を含む文節が文末に出現しているか否か
arg	{ga,o,ni}_ne	ガ格（ヲ格，ニ格）が固有名である場合，その固有名の種類（例: 人名，組織名）
	{ga,o,ni}_noun	ガ格（ヲ格，ニ格）の名詞句が日本語語彙大系 [8] の名詞意味体系に登録されている場合，その意味体系に基づく名詞の意味クラス
	{ga,o,ni}_embedded	ガ格（ヲ格，ニ格）が述語と連体修飾の関係にあるか否か
coref	{ga,o,ni}_exo	ガ格（ヲ格，ニ格）が省略されており，外界照応の関係にある
	{ga,o,ni}_srl_order	ガ格（ヲ格，ニ格）の格要素が SRL のどのスロットに入っているか
	{ga,o,ni}_srl_rank	ガ格（ヲ格，ニ格）の格要素の SRL 内の順位
	{ga,o,ni}_coref_num	ガ格（ヲ格，ニ格）の格要素が前方文脈に共参照関係になる表現を持つ場合，その談話要素の個数

素性タイプ pred，arg，coref はそれぞれ 3 節で導入した動詞の受動化の選好，項の意味的なカテゴリの情報，前方文脈の表現との共参照関係に対応する素性を表す．

ベルを利用し，また意味カテゴリの情報として，項の主辞となる形態素が日本語語彙大系 [8] の名詞意味体系のどの意味クラスに属しているかを素性として利用する．

また，共参照関係を扱う素性のうち，センタリング理論に基づく情報として Nariyama[4] の提案する Salience Reference List（SRL）を利用する．SRL はセンタリング理論 [2] の foward looking center の拡張に相当するもので，センタリング理論では前文の談話要素しか考慮しないのに対し，SRL では文章の最初から読み進めていき，顕現性の高い談話要素を保持，同一レベルの顕現性の談話要素が出現した場合にはその談話要素でリストを上書きしながらリストを更新する．このように，談話要素の情報を保持することで，センタリング理論では扱うことができなかった前文以外の顕現性の高い情報を加味したリストが作成できる．このリストに保持された談話要素のうち，各項がリストに存在するか，存在した場合はどの格として出現しているのか，また，リストに保持された談話要素のうち何位に相当するかという情報を素性として利用することで，項の顕現性の情報を受動化の分類に反映する[3]．

5 評価実験

3 節に示した 3 種類の特徴がそれぞれ受動化にどう影響するかを調査するために評価実験を行った．

5.1 評価用データ

提案する受動化のモデルの性能を調査するために，述語項構造が人手でアノテーションされた NAIST テキストコーパス [9] を利用する．このコーパスでは，述語の原形に対して項構造がアノテーションされているため，その原形に関する述語項構造と前方文脈の情報を入力として分類対象となる述語が受動化をともなうか否かを分類する．NAIST テキストコーパスでは，述語項構造に加え共参照関係も人手で付与されているため，この関係を利用して表 1 に示した共参照関係に関する素性を抽出する．

表 2: 訓練・評価用データ

	記事数	述語数	受動化された述語数
訓練用データ	1,753	65,592	4,974 (7.6%)
評価用データ	696	24,884	1,891 (7.6%)

また，NAIST テキストコーパスでは訓練データと評価データの明確な区別が存在しないため，データの分割については既存研究 [10] にしたがってコーパスを訓練データと評価データに分割した．データの内訳を表 2 に示す．

5.2 実験設定

4 節で導入した素性を利用し学習・分類を行うために，最大エントロピーモデル[4]を利用した．表 1 の各素性タイプの有効性を調査するために，その素性を利用する場合としない場合の分類性能を評価する．また，表 1 で導入した素性のタイプは，例えば，有生性の情報が共参照関係に影響するなど，依存関係が存在すると考えられる．そこで，素性の組み合わせの有効性も調査するために，SVM[5]を用いた学習・分類を行い，線形・多項カーネルを利用した場合の精度の変化も調査する．

表 1 に示した SRL に関する情報を正しく計算するためには，前方文脈の情報が正しく生成されている必要がある．そこで，評価実験では分類対象となる述語を含む述語項構造より前方では，正しく文章が生成できていると仮定し，コーパスに出現している表現をそのまま生成できたと仮定して評価を行う．評価の際は受動化した場合を正解とした再現率，精度，F 値で性能を評価する．

提案する受動化モデルの性能を評価するために，2 種類のベースラインモデルと比較を行う．1 つ目のモデルは 3 節に示した受動化に必要と考えられる情報のうち外界照応の情報のみを参照するモデルである．このモデ

[3] SRL の詳細については文献 [4] を参照されたい．

[4] http://www.cs.utah.edu/~hal/megam/

[5] http://svmlight.joachims.org/

表 3: 評価結果: 動詞の受動化の選好スコアのみを利用

	再現率	精度	F 値
$\theta = 0.1$	0.768	0.269	0.399
$\theta = 0.2$	0.573	0.357	**0.440**
$\theta = 0.3$	0.403	0.450	0.425
$\theta = 0.4$	0.293	0.512	0.373
$\theta = 0.5$	0.161	0.591	0.253
$\theta = 0.6$	0.091	0.692	0.162
$\theta = 0.7$	0.060	0.717	0.111
$\theta = 0.8$	0.030	0.851	0.058
$\theta = 0.9$	0.014	1.000	0.027

ルでは，ガ格が外界照応の場合に受動化し，それ以外は受動化しないという判断を行う．もう一つのモデルは，動詞の受動化に関する選好に基づくモデルで，式 (1) の score_{pas} が閾値 θ 以上の場合は受動化し，それ以外の場合は受動化しないという分類を行う．

5.3 実験結果

まず，2 つ目のベースラインモデルの閾値 θ を決定するために，θ の値は 0.1 から 0.9 まで 0.1 刻みで変動させた結果を表 3 に示す．この結果より，θ が 0.2 の場合に最も F 値が高くなっているのがわかる．そこで，この結果を他のモデルとの比較対象とする．

次に，ベースラインモデルと提案モデルを比較した結果を表 4 に示す．この結果より，単一の素性タイプ（ME: pred, arg, coref）のみを利用した提案モデルはベースラインモデルよりも精度が低くなっていることがわかる．一方，述語に関する素性タイプとそれ以外を組み合わせた結果（ME: pred+arg, pred+coref）はベースラインを上回る結果を得ていることがわかる．この結果より，述語の受動化はベースラインで採用した個別の特徴を組み合わせることで精度が向上することがわかる．さらに，3 種類すべての特徴を組み合わせることで 2 種類の特徴を組み合わせた場合と比較してさらに精度が向上しており，F 値で 0.604 という結果を得た．さらに，素性の組み合わせを考慮するために多項カーネルを用いた SVM で学習・分類した場合にはさらに性能が向上しており，特に多項 2 次カーネルを用いた場合，最大エントロピーモデルと比較して F 値で 0.15 以上の性能の改善が見られた．この結果より，3 節で導入した述語受動化の特徴には依存関係があり，それらを明示的に組み合わせて問題を解くことが重要であることがわかった．

6 おわりに

本稿では，動詞の受動化の選好，述語の項の意味カテゴリ情報，前方文脈に出現する表現の談話的な関係の 3 種類の手がかりを利用した節の受動化モデルを提案した．日本語新聞記事コーパスに述語項構造と共参照関係が人手でアノテーションされた評価用データを利用し，提案モデルの評価を行い，F 値で 0.758 という結果を得た．

今後の課題としては，受動化のプロセスと参照表現生成の統合が考えられる．この論文で提案した受動化のモデルは前方文脈の生成結果を仮定しており，前方文脈の談話要素の生成結果が必要であるが，既存研究 [10] で提案した参照表現生成モデルと組み合せることで，参照表現と受動化の生成を同時に達成することが可能になると考えられる．このため，それぞれのモデルを組み合せる際に，どのようにして最適な結果を得るかを検討することが今後必要となる．

また，自動推敲など，すでに記述された文章に対して文章を再生成する問題を考えた場合，述語項構造や共参照関係は自動解析する必要がある．日本語の場合，ゼロ照応関係の解析は非常に困難であるため [3]，文章全体の述語項構造・共参照解析を行い，その結果から生成という流れでは，多くの誤りを含んだ状態で生成を行う危険性も含まれる．このため，このような問題設定の場合，どの解析が信頼度が高く，その結果を利用してよいのか，またどの部分は解析してはならないのかという生成に関する取捨選択を行う必要がある．この問題についても，受動化と参照表現の生成の統合後に取り組む予定である．

表 4: 評価結果: 全体

	再現率	精度	F 値
baseline1: exophora	0.493	0.329	0.395
baseline2: $\text{score}_{pas} \geq \theta$ ($\theta = 0.2$)	0.573	0.357	0.440
ME: pred	0.264	0.608	0.369
ME: arg	0.130	0.555	0.211
ME: coref	0.092	0.574	0.159
ME: pred+arg	0.397	0.656	0.494
ME: pred+coref	0.482	0.761	0.590
ME: arg+coref	0.270	0.651	0.381
ME: all	0.507	0.747	0.604
SVM(linear): all	0.456	0.792	0.579
SVM(poly-2d): all	0.679	0.858	**0.758**
SVM(poly-3d): all	0.625	0.878	0.730

baseline1 が外界照応の情報のみを参照するモデルであり，baseline2 が動詞の受動化に関する選好に基づくモデルを表す．

謝辞

本研究は科研費若手研究（A）「談話解析技術に基づいた文章推敲支援」（課題番号: 23680014）の支援を受けた．記して謝意を表する．

参考文献

[1] B. Abb, M. Herweg, and K. Lebeth. The incremental generation of passive sentences. In *Proceedings of the 6th EACL*, pp. 3–11, 1993.
[2] B. J. Grosz, A. K. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, Vol. 21, No. 2, pp. 203–226, 1995.
[3] R. Iida and M. Poesio. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of ACL-HLT 2011*, pp. 804–813, 2011.
[4] S. Nariyama. Grammar for ellipsis resolution in japanese. In *In Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 135–145, 2002.
[5] R. Power. Planning texts by constraint satisfaction. In *Proceedings of COLING 2000*, pp. 642–648, 2000.
[6] D. R. Radev and K. R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, Vol. 24, No. 3, pp. 469–500, 1998.
[7] F. Abu Sheikha and D. Inkpen. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pp. 187–193, 2011.
[8] 池原悟, 宮崎正弘, 白井論, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系. 岩波書店, 1997.
[9] 飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治. 述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から. 自然言語処理, Vol. 17, No. 2, pp. 25–50, 2010.
[10] 飯田龍, 徳永健伸. 日本語書き言葉を対象とした参照表現の自動省略-人間と機械処理の省略傾向の比較-. 情報処理学会自然言語処理研究会予稿集, NL-206-15, 2012.