Vol. 45 No. 4 情報処理学会論文誌 Apr. 2004

自動生成された言い換え文における不適格な動詞格構造の検出

藤田 篤[†] 乾 健太郎[†] 松本 裕治[†]

本論文では，語彙・構文的言い換えにおいて頻繁に生じる動詞格構造の不整合を自動的に検出する方法を提案する．我々は，コーパスから獲得した大規模な正例に基づいて格構造の適格さを定量化する確率的言語モデルと，人手で収集した小規模な負例に基づいて格構造の不適格さを定量化するモデルを構築し，これら 2 つを混合し，正例のみに基づく言語モデルと比較して精度の高い誤り検出器を実現した．また，誤り検出に対して貢献度が高い負例を効率良く収集するために能動学習を試行した．

Detection of Incorrect Case Assignments in Automatically Generated Paraphrases

Atsushi Fujita ,[†] Kentaro Inui [†] and Yuji Matsumoto[†]

This paper addresses the issue of detecting transfer errors in paraphrasing. Our previous investigation revealed that case assignment of verb tends to be incorrect, irrespective of the types of lexical and structural paraphrasing of Japanese sentences. Motivated by this observation, we propose an empirical method to detect incorrect case assignment. Our error detection model combines two error detection models. They are separately trained on a large collection of positive examples and a small collection of manually labeled negative examples. Experimental results show that our combined model significantly enhances the baseline model which is trained only on positive examples. We also propose a selective sampling scheme to reduce the cost of collecting negative examples, and confirm the effectiveness for the error detection task.

1. はじめに

言い換えの自動生成は，自然言語処理のさまざまなアプリケーションに利用可能な要素技術として注目されている[1),19)]．たとえば，機械翻訳や手話生成では，対象テキストを機械で処理しやすい（翻訳しやすい，あるいは手話で表現可能な）表現に言い換えることで，訳質や翻訳の被覆率を向上させたり，変換のための知識を簡素化したりできる[23),25),28)]．また，音声合成の前処理として曖昧性が少なくなるようにテキストを言い換えておけば，人間の聞き取りが容易になるし，語彙・構文的に簡単な表現に言い換えることによって読解支援を図ることも考えられる[3),11)]．我々は，これらのアプリケーションを想定して，言い換えを自動生成する機構の実現を目指している．

上にあげたアプリケーションのいずれにおいても，出力としては適格な自然言語テキストが求められる．ところが，適格な言い換えを生成するために言い換えパターンの適用条件を書き尽くすことは困難であるため，さまざまな変換誤りが生じる可能性がある．我々は文献 6) において，さまざまな種類の語彙・構文的言い換えパターンを用いて言い換え事例を自動生成し，事例の分析に基づいて変換誤りの種類と各誤りが生じる傾向を調査した．そして，言い換えパターンの種類にかかわらず頻繁に生じる，最も優先して解消すべき変換誤りの一つとして，**動詞格構造における動詞と格要素の名詞の不整合**をあげた．

動詞格構造における不整合は次の 2 つのレベルの制約違反からなる．1 つ目は，動詞の下位範疇化構造に関する誤り，すなわち統語レベルの制約違反である．例文☆(1t) は，動詞「貫く」がニ格を取ることができないため不適格である．2 つ目は，格要素の選択制限の違反，すなわち意味レベルの制約違反である．動詞「上回る」のヲ格要素は,『記録』，あるいは『精度』のような概念の名詞句でなくてはならない．しかし，例文 (2t) は，ヲ格の名詞「国境」が上記の制約を満た

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

☆ 本論文の例文中，R.，s.，t.，r. は各々，言い換えパターン，言い換え前の文，言い換え後の文，修正後の言い換え文を指す．

していないため不適格である．

(1) s. チームプレーに**徹する**．
t. *チームプレーに**貫く**．
r. チームプレー**を**貫く．

(2) R. N1 が N2 を超える ⇒ N1 が N2 を上回る
（N1，N2 は名詞を表す変数とする）
s. ネットワークが**国境を超えた**．
t. *ネットワークが**国境を上回った**．

例文 (1t)，(2t) のような動詞格構造における不整合は，言い換えに関する先行研究において指摘されてきた[5),12),15)] が，解決に取り組んだ例はない．そこで，本論文では，動詞格構造における名詞と動詞の不整合の自動検出について議論する．

動詞格構造における不整合は，動詞結合価辞書に照らすことで検出できるように見えるかもしれない．たしかに，(1t) のような下位範疇化構造に関する誤りは，格助詞の種類が有限であるため，動詞ごとにとりうる格助詞を記述することで検出できる可能性がある．しかし，(2t) のような格要素の選択制限の誤りは，オープンクラスの語（動詞や名詞など）の組合せの問題である．ゆえに，選択制限を満たす単語集合を辞書に記述するアプローチは現実的ではない．

一方，既存の名詞シソーラスを用いてクラスベースで選択制限を与える統計的なモデルもいくつかある．宇津呂ら[27)]，宮田ら[18)] は，動詞と格要素の共起用例に基づく確率モデルを提案している．また，河原ら[13)] は，動詞の直前の格要素をキーとして格フレームを自動的に獲得する手法を提案している．しかしながら，名詞シソーラスを用いると，同じ意味クラスに分類される単語を区別できないことが問題となる．

たとえば，「基盤」，「土台」，「根底」の 3 語は，EDR 日本語単語辞書[4)] において『3cf93c:「基礎」物事や行動の基礎』という概念（意味クラス）に属する**同概念語**である．EDR 日本語単語辞書の意味クラスは他の日本語シソーラスと比較して粒度が細かい☆ため，同じ意味クラスに属する同概念語は原則として同義語と見なすことができ，次のような言い換えを生成できる．

(3) s. 遺伝子治療の**基盤**が崩れる．
t1. 遺伝子治療の**土台**が崩れる．
t2. 遺伝子治療の**根底**が崩れる．

しかし，同様に例文 (4s) を言い換えた場合，選択制限の違反が生じる．

(4) s. 政策責任者が党の**基盤**を固める．
t1. 政策責任者が党の**土台**を固める．
t2. *政策責任者が党の**根底**を固める．

「基盤」，「土台」，「根底」は，動詞結合価辞書の 1 つである日本語語彙大系においても『2446：基・源』という同じ意味クラスに分類されているため，こうしたシソーラスを使うだけでは，「固める」のガ格として「土台」は適格，「根底」は不適格，とは判定できない．

以上より，動詞格構造が適格であるか否かを判定するためには，既存のシソーラスにおける単語の意味クラスよりも細かい粒度の分類に基づくモデルが必要であると考える．そこで，個々の単語を区別した統計的な誤り検出モデルを提案する．

適格さを統計的に見積もるには，コーパスから大規模に獲得できる共起用例（正例）を用いることが考えられる[14),16)]．さらに，誤り検出においては負例を用いることが有効であると考えられるが，負例のリソースは現状では存在しないため人手で収集する必要がある．したがって，次の 2 つが課題となる．

課題 1. 大規模な正例と人手で収集できる非常に小規模な負例をいかにうまく利用するか．

課題 2. 誤り検出に必要な負例をまんべんなく収集することは不可能である．有効な負例のみをいかに効率良く収集するか．

本論文では，この 2 つの課題を考慮して構築した誤り検出モデルと，その評価実験について述べる．以下，まず，2 章では，動詞格構造の誤り検出タスクとアルゴリズムについて述べる．3 章では，それぞれの課題に対して本論文で提案する誤り検出モデルについて述べる．4 章でモデルの訓練および誤り検出の評価実験の結果を示し，5 章でまとめる．

2. 不適格な動詞格構造

2.1 問題の深刻さ

我々は，文献 6) において，日本語の自動言い換えにおける変換誤りを 2 つの点から分析した．1 つ目は言い換えにおいて生じる変換誤りの種類の分析であり，2 つ目は各変換誤りが生じる頻度に関する分析である．ここでは，約 28,000 個の言い換え規則を実装し，自動的に生成した言い換え事例 630 件を対象とした分析の結果，162 件（25.7%）が動詞格構造の不整合を含んでいた．これは，動詞や形容詞の活用形の誤り（303 件）に次いで頻出した深刻な問題である．

2.2 タスク設定

本論文で適格か不適格かを評価する対象は，言い換えによって新出した単語，あるいは依存構造を変更さ

☆ 日本語語彙大系[10)] や角川類語新辞典[21)] では，「油絵」と「線画」，あるいは「林檎」と「桃」といった類義語は同じ意味クラスに収録されているが，EDR 日本語単語辞書はこれらを別の意味クラスに収録している．

れた単語を含む動詞格構造である．誤り検出モデルは，以下の理由から依存構造を考慮して設計する．

- n-gram に基づく統計量が表現の適格さを判定するタスクに有効であるという報告がある[14),16)]．しかし，これは英語を対象とした場合であり，日本語に対しても同様であるとは限らない．日本語は比較的語順が自由な言語であるため，我々は，依存構造を扱う方がより正確に適格さを見積もることができると考える．
- 既存の日本語言い換えシステムの多くは依存構造をデータ構造として言い換えを実現している[12),15),24)]．これらのシステムでは，言い換え後の文がたとえ言語的に適格でなくても，その依存構造を参照することが可能である．

我々は，1 章で述べたように，個々の単語を区別した誤り検出モデルを構築する．以下，動詞格構造中，名詞 n が格助詞 c を介して動詞 v に係っている（依存している）関係を 3 つ組 $\langle v, c, n\rangle$ で表す．

例文 (1t)，(2t) は，⟨ 貫く, に, チームプレー ⟩，⟨ 上回る, を, 国境 ⟩ という，動詞格構造中の 1 つの $\langle v, c, n\rangle$ が不適格な例である．一方，例文 (5t) は，⟨ ある, が, 言葉 ⟩，⟨ ある, に, 各地 ⟩ がともに適格であるにもかかわらず，これらの共起が不適格な例である．

(5) s. 文語体，しかも難解な言葉が**随所**にある．
t. *文語体，しかも難解な言葉が**各地**にある．

したがって，ある動詞格構造が適格であるか否かは，次のような決定木に照らして判定できる．

決定木 1： 動詞格構造の適格性判定（人手）
- $\langle v, c, n\rangle$ が 1 つでも不適格ならば不適格．
- すべての $\langle v, c, n\rangle$ が適格
  - 兄弟格要素の共起が不適格ならば不適格．
  - 兄弟格要素の共起も適格ならば適格．

兄弟格要素との共起の適格性は，文献 26) のように，$\langle v, c_1, n_1, c_2, n_2\rangle$ の同時分布をモデル化すればとらえることができるが，共起用例のデータスパースネスが深刻になる．また，文献 6) における調査では，例文 (5t) のように個々の $\langle v, c, n\rangle$ が適格であるにもかかわらず兄弟格要素の共起が適格でない事例は，162 件中 8 件（誤りのうち 4.9%）と稀であった．

これらをふまえ，本論文では，解くべき問題を，個々の $\langle v, c, n\rangle$ を適格/不適格の 2 つのクラスに分類する問題に単純化する．すなわち，次の決定木によって適格性を判定する．

決定木 2： 動詞格構造の適格性判定（モデル）
- $\langle v, c, n\rangle$ が 1 つでも不適格ならば不適格．
- すべての $\langle v, c, n\rangle$ が適格ならば適格．

3. 提案モデル

3.1 課　　題

不適格な動詞格構造を検出するには，統計的機械翻訳における翻訳候補選択（デコーディング）[2),7)] が参考になる．候補生成の後処理として，個々の候補の尤もらしさを言語モデルに照らして評価するためである．ただし，デコーディングや他の自然言語処理タスクにおいて，言語モデルは，可能な出力候補の中から**相対的な尤もらしさ**に基づいて 1 つの候補を選択するために用いられている．これに対し，誤り検出では，次に示す理由から，個々の候補の**絶対的な尤もらしさ**を見積もらなければならない．

言い換えシステムは，多くの場合，テキストの簡単化や制限言語への変換など，特定のタスクへの応用を目的として構築される．このようなシステムで用いられる言い換えパターンは,「難解な語から平易な語へ」など，目的に応じた制約を受けるため，任意の入力文に対して必ずしも適格な言い換えを生成できるとは限らない．したがって，**絶対的な尤もらしさ**に従って個々の言い換え候補を評価し，適切な言い換えを生成できない場合には「言い換え不可能である」と出力しなくてはならない．

個々の言い換え候補を適格/不適格の 2 つのクラスに分類する問題ととらえるのであれば，正例と負例の両者を用いて機械学習に基づく分類器を構築するアプローチが有効であると考えられる．しかし，コーパスに既存の解析技術を適用することで大規模な正例が獲得できるのに対して，負例のリソースは存在しない．さらに，単語そのものからなる 3 つ組 $\langle v, c, n\rangle$ の共起空間は非常に広大であるため，人手で収集できるどのような負例集合も，共起空間に対してきわめてスパースである．したがって，サポート・ベクタ・マシン（SVMs; Support Vector Machines）など優れた性能が報告されている 2 値分類アルゴリズムを用いて誤り検出モデルを構築したとしても，有効に働くとは期待できない．

これらをふまえ，本論文では，次の 2 つの解決策を提案する．

提案 1. 大規模な正例と小規模な負例をいかにうまく利用するか，という課題に対し，正例，負例のそれぞれのみを用いた誤り検出モデル *Pos*，*Neg* を別々に構築し（**図 1**），これら 2 つを混合して誤り検出モデル *Com* を構築する．また，

提案 2. 誤り検出に有効な負例をいかに効率良く収集するかという課題に対し，上記 *Pos*，*Neg* を用い

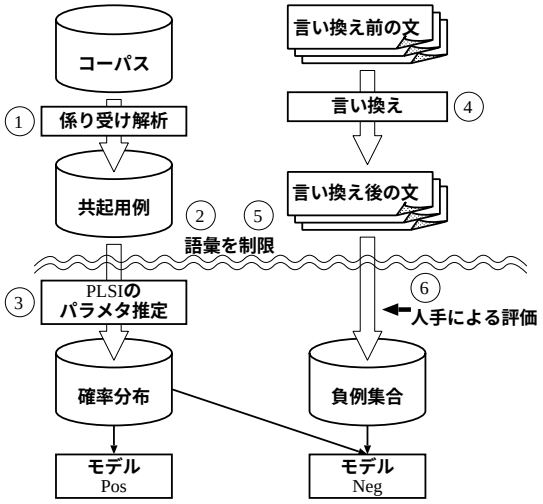

図 1 誤り検出モデルの構築手順
Fig. 1 Model construction scheme.

て負例を選択的に収集する手法を提案する．

3.2 正負例各々に基づくモデルの混合 ——課題 1 に対するアプローチ

正例は，既存のコーパスから大規模に獲得することができるため，まず，これを用いて統計的言語モデル *Pos* を構築する．本論文の目的は，負例を用いることによって，*Pos* よりも優れたモデルを実現することである．ただし，統計的言語モデルの学習に負例を直接用いることはできないし，きわめて小規模である．そこで，*Pos* とは独立に，負例のみに基づくモデル *Neg* を構築し，最終的に *Pos* と *Neg* を混合した誤り検出器 *Com* を構築する．以下，各モデルについて詳述する．

3.2.1 正例に基づく誤り検出モデル *Pos*

コーパスから獲得できる正例はきわめて大規模であるため，高い頻度で出現する 3 つ組 $\langle v,c,n\rangle$ については，その生起確率 $P(\langle v,c,n\rangle)$ を正確に見積もることができると考えられる．そこでまず，正例のみを用いて統計的言語モデル *Pos* を構築する．

$\langle v,c,n\rangle$ の共起空間は広大であるため，$P(\langle v,c,n\rangle)$ を推定する際にデータスパースネスの影響が生じると予想される．ただし，v や n になんらかのクラスを考慮することで軽減できる可能性がある．そこで，$P(\langle v,c,n\rangle)$ を推定するさまざまな手法の中から，分布クラスタリング[22] に基づく Probabilistic Latent Semantic Indexing（PLSI）[9] を採用した．

$\langle v,c,n\rangle$ を $\langle v,c\rangle$ と n の共起と見なす☆と，PLSI における共起確率 $P(\langle v,c,n\rangle)$ は次式で与えられる．

$$P(\langle v,c,n\rangle)=\sum_{z\in Z}P(\langle v,c\rangle|z)P(n|z)P(z),$$

Z は共起に関する潜在的な意味クラス（隠れクラス）を指す確率変数であり，この式が示すとおり，分布クラスタリングはこの Z に基づく確率的ソフトクラスタリングである．式中の確率的パラメータ $P(\langle v,c\rangle|z)$，$P(n|z)$，$P(z)$ は，文献 9) にならい，EM アルゴリズムを適用して推定する．

$P(\langle v,c,n\rangle)$ が与えられれば，相互情報量などのさまざまな共起尺度を用いて，$\langle v,c,n\rangle$ の適格さを推定することができる．今回は，共起確率そのもの（*Prob*）とともに，$\langle v,c\rangle$ と n の相互情報量（*MI*），$\langle v,c\rangle$ と n の Dice 係数（*Dice*）を適格さの尺度として，その性能を調査した．モデル *Pos* は，入力の $\langle v,c,n\rangle$ に対して，*Prob*，*MI*，もしくは *Dice* をスコアとして出力する．

$$MI(\langle v,c\rangle,n)=\log\frac{P(\langle v,c,n\rangle)}{P(\langle v,c\rangle)P(n)},$$
$$Dice(\langle v,c\rangle,n)=\frac{2\times P(\langle v,c,n\rangle)}{P(\langle v,c\rangle)+P(n)}.$$

3.2.2 負例に基づく誤り検出モデル *Neg*

$\langle v,c,n\rangle$ の共起空間は広大であるため，$P(\langle v,c\rangle)$，もしくは $P(n)$ が低い場合，相対的に *Pos* が出力するスコアも信頼できなくなる．この欠点を補うために，負例のみに基づく誤り検出モデル *Neg* を構築する．利用できる負例の数がごく少数であること，共起空間に対してスパースであることを考慮し，k-最近隣法（k-Nearest Neighbor）を採用した．

入力の $\langle v,c,n\rangle$ と任意の学習済事例（負例）$\langle v',c',n'\rangle$ の距離は，n と n'，および $\langle v,c\rangle$ と $\langle v',c'\rangle$ の距離に基づいていると仮定する．n どうし，$\langle v,c\rangle$ どうしの距離は，*Pos* の構築の際に得られた隠れクラス Z への帰属確率分布を用いて算出できる．これらをふまえ，$\langle v,c,n\rangle$ と $\langle v',c',n'\rangle$ の距離を次の関数で見積もる．

$$\begin{aligned}&Dist(\langle v,c,n\rangle,\langle v',c',n'\rangle)\\&\quad=DS\Big(P(Z|n),P(Z|n')\Big)\\&\qquad+DS\Big(P(Z|\langle v,c\rangle),P(Z|\langle v',c'\rangle)\Big),\end{aligned}$$

関数 *DS* は，確率分布どうしの分布類似度☆☆を表す．

分布類似度の尺度としては，文献 17) で示されて

☆ $P(\langle v,c,n\rangle)$ は，下位範疇化構造の尤度 $P(\langle v,c\rangle)$ と選択制限の充足度 $P(n|\langle v,c\rangle)$ の積と解釈できる．この解釈も含めて，文献 8) は，英語を対象として $\langle v,c,n\rangle$ を表現するためのさまざまな確率モデルを検討している．日本語を対象とした追試も必要である．

☆☆ 『類似度』と呼ばれるが，実際の計算値は 2 つの確率分布が類似しているほど 0 に近く，『距離』を示している．

いる分布類似度の中から，優れた実験結果が得られている Jensen-Shannon divergence（本論文では以下，DS_{JS} と表記する）を採用した☆．2 つの確率分布 q, r に対して，DS_{JS} は次の式で与えられる．

$$DS_{JS}(q,r) = \frac{1}{2}\left[D\left(q \,\middle\|\, \frac{q+r}{2}\right) + D\left(r \,\middle\|\, \frac{q+r}{2}\right)\right],$$

ここで，D は Kullback-Leibler（KL）divergence であり，DS_{JS} は，KL divergence に確率 0 への耐性と対称性を持たせたものである．

$$D\Big(P_1(X) \,\Big\|\, P_2(X)\Big) = \sum_{x\in X} P_1(x)\log\frac{P_1(x)}{P_2(x)}.$$

モデル *Neg* は，入力の $\langle v,c,n\rangle$ に対して，$\langle v,c,n\rangle$ とその k 個の最近傍事例との分布類似度 d の重みつき平均を適格さのスコア $Score_{Neg}$ として出力する．

$$Score_{Neg} = \frac{1}{k}\sum_{i=1}^{k} \lambda_i\, Dist\left(\langle v,c,n\rangle, \langle v',c',n'\rangle_i\right),$$

ここで，λ_i は i 番目の最近傍事例，$\langle v',c',n'\rangle_i$ に対する重みである．

3.2.3 *Pos* と *Neg* の混合モデル *Com*

Pos は正例のみを用いた確率的言語モデル，*Neg* は既知の負例との距離ををを計算するモデルである．これら学習データ，出力の尺度が異なる 2 つのモデルの混合モデル *Com* を構築する．

Pos は確率（$0 \le Prob \le 1$），あるいは共起尺度（$-\infty \le MI, Dice \le \infty$），*Neg* は分布類似度の平均（$0 \le Score_{Neg} \le \infty$）というように，各モデルが出力するスコアの尺度は異なる．そこでまず，モデルが出力するスコア s を信頼度 C（$0 \le C \le 1$）に写像する．C は，「$\langle v,c,n\rangle$ のスコアが s であるならば $\langle v,c,n\rangle$ は適格である」ということをどれだけ信頼できるかを表す．s から C への写像関数は，学習データと関数を推定するためのデヴェロップメントデータを用いて次の手順で導出する．

(Step 1) 学習データを用いてモデルを構築し，デヴェロップメントデータ中の $\langle v,c,n\rangle$ に対してスコアを付与する．

(Step 2) スコアを付与された $\langle v,c,n\rangle$ を，スコアの部分区間ごとに割り当て，各区間において適格である $\langle v,c,n\rangle$ の割合を，部分区間の中央値 s が示す $\langle v,c,n\rangle$ の適格さの信頼度 C とする．

(Step 3) $\langle s,C\rangle$ の点を線形補間することで写像関数とする．

なお，4 章で述べる評価実験においては，5 分割交差検定を行う際に，5 分の 4 にあたる訓練データの 4 分割交差検定によって写像関数を導出した．

Com は，入力の $\langle v,c,n\rangle$ に対して *Pos*，*Neg* が出力するスコアを信頼度を各々C_{Pos} と C_{Neg} に写像し，その重み付き平均を自身のスコアとして出力する．

$$Score_{Com}(\langle v,c,n\rangle) = \beta\, C_{Pos} + (1-\beta)\, C_{Neg},$$

ここで，β（$0 \le \beta \le 1$）はモデルの重みであり，$Score_{Com}$ は *Pos*，*Neg* のスコアと同様に，小さいほど不適格，大きいほど適格であるという意味を持つ．

3.3 能動学習——課題 2 に対するアプローチ

負例を収集するコストをいかにして削減するか．この課題に対して，我々は能動学習を導入した．能動学習は，未知の事例に対するモデルの出力をもとに次に学習（人手でラベル付け）すべき事例を決定し，モデルの学習を効率良く進める手法である．ラベル付けする事例が少数であっても，有益な事例を選択的に収集できれば，高精度なモデルを構築できる．

我々の誤り検出モデルにおける能動学習の対象は *Neg* であり，学習事例となる 3 つ組 $\langle v,c,n\rangle$ は，(i) 正例ではない，(ii) すでに収集済の負例のいずれとも似ていない，という性質を満たしているほど有益であると考えられる．以下，3 つ組をサンプルと呼ぶ．

サンプルがいかに正例らしくないかは，*Pos* によって推定できる．また，サンプルとすでにラベル付けされている負例との距離は，*Neg* によって算出できる．あるサンプル x に対して *Pos* が出力する尤度を p_x，*Neg* が出力する最近傍事例（負例）との距離を s_x として，次のような選好関数を作成した．

$$Preference(x) = -s_x \log(p_x),$$

この式では，値が大きいほどラベル付けの優先度が高いとする．p_x が低いほど好ましいので符号を反転させている．また，広大な共起空間の広い範囲をカバーするようなサンプルの方が，学習に大きく寄与するであろうと仮定し，学習済の負例との距離 s_x を重視するため，p_x の対数を用いて調整している．

能動学習アルゴリズムは以下のとおりである．

(Step 1) 対象とするドメインの文書から取り出した文に言い換え規則を適用し，言い換え事例集合を生成する．

(Step 2) 言い換え事例集合から，$\langle v,c,n\rangle$ の集合を

☆ 文献 17) は，α-skew divergence という尺度が，Jensen-Shannon divergence よりも優れていると報告している．我々のタスクにおいても，α-skew divergence はパラメータ α の値によっては優れた精度を示したが，Jensen-Shannon divergence に対して有意ではなかった．また，最適なパラメータ α の推定も困難であるため，本論文ではこの尺度については言及しない．

取り出す．以下，$\langle v, c, n\rangle$ をサンプル，取り出した集合を，サンプルプールと呼ぶ．

(Step 3) サンプルプールから少数のサンプルをランダムに取り出し，人手で正負のラベルを付与する．このうち，負例のみを *Neg* の初期学習事例とする．

(Step 4) サンプルプールの各サンプルに対して，上に示した選好関数によって優先度を付与する．

(Step 5) 最も優先度が高いサンプルを取り出し，人手で正負のラベルを付与する．負例であれば *Neg* の学習事例に追加し，正例であれば次に優先度が高いサンプルを取り出す．

(Step 6) 停止条件を満たすまで Step 4, 5 を繰り返す．停止条件とは，たとえば，*Neg* あるいは *Com* の精度が収束する，などである．

4. 誤り検出実験

4.1 モデルの訓練および評価事例の作成

評価事例集合の作成およびモデルの訓練の手順を以下に示す．図 1 も参照されたい．

(Step 1) 新聞記事 19 年分☆1の係り受け解析結果☆2から，のべ 53,157,450 組，異なり 7,993,331 組の $\langle v, c, n\rangle$ を収集した☆3．名詞は 65,384 語，動詞は 33,884 語であった．

(Step 2) 2 つの語彙でモデルを構築した．1 つ目は，のべ 2,000 回以上出現した名詞 3,365 語，動詞 2,516 語，および頻度が高い 7 つの格助詞 “が”，“を”，“に”，“で”，“へ”，“から”，“より” のみからなる $\langle v, c, n\rangle$ のを用いたモデル *LexS* である．異なりで 3,628,345 組がこの語彙で表現されており，$\langle v, c\rangle$ の異なりは 16,899 種，$\langle v, c\rangle$ と n の共起行列要素充填率は 6.38%であった．2 つ目は，2 回以上出現しており，かつ *LexS* と同じ格助詞のみを含む $\langle v, c, n\rangle$ を用いた *LexL* である．こちらは，異なりで 3,110,546 組，$\langle v, c\rangle$，n の異なりはそれぞれ 66,484 種，38,512 種，共起行列要素充填率は 0.12%であった．

(Step 3) Step 2 で得た 3 つ組の集合を PLSI 学習パッケージ☆4 に入力し確率的パラメタを推定した．隠れクラスの数 $|Z|$ は，*LexS* に対しては 2〜3,000 の中のいくつか，*LexL* については計算機資源の制約上，2〜1,500 のいくつかを試行した．

(Step 4) 評価事例集合およびモデル *Neg* の訓練に用いるための負例を収集した．Step 1 で用いた新聞記事の一部（日本経済新聞，2,000 年）からランダムに取り出した 90,000 文を言い換えエンジン☆5に入力し，文献 6) で用いたさまざまな種類の言い換えパターンを網羅的に適用して 7,167 事例を生成した．

(Step 5) Step 4 で生成した言い換え事例から，(i) 格要素も動詞も言い換えられていない事例，(ii) 言い換えられた動詞格構造中の n，c，v のいずれかが Step 2 で定義した *LexS* に含まれない事例を除き，3,166 事例を取り出した．

(Step 6) Step 5 で得た 3,166 事例，および事例に含まれる異なりで 3,704 組の $\langle v, c, n\rangle$ を人手で適格/不適格に分類した．結果，事例としては正例 2,358，負例 808（25.5%）☆6を，$\langle v, c, n\rangle$ としては正例 2,853，負例 851 を得た．

4.2 評価方法

モデルが出力するスコア（確信度）に対して閾値を設け，モデルを 2 値分類器として扱う．すなわち，閾値以下の値を持つ事例を不適格，閾値を超える値を持つ事例を適格と判定する．ある閾値のもとでのモデルの誤り検出の能力を，以下に示す再現率 R，および精度 P で評価する．

$$R = \frac{\text{モデルが検出に成功した変換誤り事例の数}}{\text{変換誤り事例の数}},$$

$$P = \frac{\text{モデルが検出に成功した変換誤り事例の数}}{\text{モデルが変換誤りとラベル付けした事例の数}}.$$

閾値を変化させることで不適格と見なす事例の数を制御し，再現率–精度曲線（*R-P* 曲線）を描画した．また，個々の *R-P* 曲線は，$R = 0.0, 0.1, \ldots, 1.0$ の 11 点平均精度で評価し，2 つの *R-P* 曲線間の有意差の有無は，再現率の 11 点をサンプルとした Wilcoxon 符号順位和検定，有意水準 5%で検定した．

4.3 実験結果

4.1 節で作成した 3,166 事例を用い，5 分割交差検定によって動詞格構造の誤り検出の実験を行った．すなわち，評価事例集合の 5 分の 4 に含まれる $\langle v, c, n\rangle$ の負例を *Neg* の学習に用い，残りの 5 分の 1 をスコ

☆1 毎日新聞 9 年分，日本経済新聞 10 年分，のべ 25,061,504 文．

☆2 係り受け解析には CaboCha を用いた．
http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/

☆3 “れる/られる”，“せる/させる”，などの格交替に関わる動詞性接尾辞が後続する場合は，これを含めて動詞 1 語とし，格の交替現象を直接扱っている．

☆4 http://cl.aist-nara.ac.jp/~taku-ku/software/plsi/

☆5 言い換えの自動生成には KURA を用いた．
http://cl.aist-nara.ac.jp/lab/kura/doc/

☆6 評価は 2.2 節の決定木 1 に基づいている．すべての $\langle v, c, n\rangle$ が適格であるにもかかわらず兄弟格要素の共起が適格でない事例は 41 件（誤りのうち 5.1%）であった．

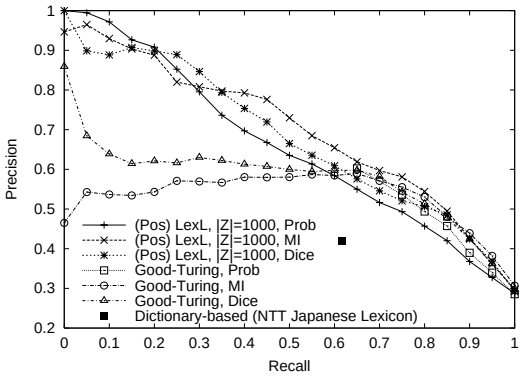

図 2 ベースラインモデルの *R-P* 曲線
Fig. 2 *R-P* curves of baselien models.

ア付けする．以下，各誤り検出モデルごとに *R-P* 曲線，および 11 点平均精度を示し考察する．

4.3.1 ベースライン

まず，ベースラインモデルによる誤り検出実験の結果を示す．ここでは，動詞結合価辞書に基づくモデル，統計量の頻度ディスカウント手法に基づくモデルを取り上げる．さらに，本論文では，負例を用いることによって *Pos* よりも高い精度の誤り検出器を構築することを目的としているため，*Pos* もベースラインと見なす．動詞結合価辞書としては日本語語彙大系[10]（*Dic*）を，統計量の頻度ディスカウント手法としては Good-Turing 法（*GT*）を用いた．上の各モデルの誤り検出における *R-P* 曲線を**図 2** に示す．

Dic は，下位範疇化構造および選択制限を満たすか否かで個々の $\langle v, c, n\rangle$ の適格/不適格を判定し，2.2 節の決定木 2 に基づいて事例の適格/不適格を判定するモデルである．ただし，3,166 事例中 338 事例（10.7%）は，動詞結合価辞書にエントリを持たない動詞が評価の対象となっていたため，*Dic* は残りの事例に対してのみ，適格/不適格を判定した（被覆率 89.3%）．この被覆率のもとで，*Dic* の誤り検出の再現率は 61.6%，精度は 41.9%であった．

GT は，共起用例の頻度ディスカウンティングにより PLSI と比較して低コストで $P(\langle v, c, n\rangle)$ 推定できるが，コーパス中に同じ回数出現する $\langle v, c, n\rangle$ 間の優劣をつけることはできない☆．このため，*Prob* では，低い再現率の範囲（今回の評価事例集合に対しては $R \leq 0.656$）では *R-P* 曲線を描くことができなかった．高い再現率の範囲（$R \geq 0.656$）でも，*Pos* の *Prob* と比べるとわずかに高い精度を得ているものの，*MI* と比べると低い精度しか得られなかった．また，共起尺度を用い，頻度 0 の $\langle v, c, n\rangle$ 間の尤度に優劣をつけた場合でも，*MI* が 51.9%，*Dice* が 58.0%と *Pos* の同じ尺度に比べて低く，有意な差があった．

LexL，$|Z| = 1,000$ における *Pos* の各尺度の 11 点平均精度は，*Prob* が 65.6%，*MI* が 69.2%，*Dice* が 67.5%であった．また，*LexS*，$|Z| = 2,000$ においては，*Prob* が 63.3%，*MI* が 66.1%，*Dice* が 65.5% であった．*R-P* 曲線を比較した場合でも，*Pos* は他のモデルに比べて顕著に高い精度を示しており，提案する誤り検出モデル *Com* を構築・評価するうえで十分高い精度のベースラインであるといえる．

モデル *Pos* の各尺度について，分布クラスタリングにおける隠れクラス数 $|Z|$ と 11 点平均精度の関係を**図 3** に示す．全体的に共起尺度（*MI* および *Dice*）が *Prob* よりも優れた 11 点平均精度を示したが，図 2 が示すように，*Prob* と共起尺度（*MI*，*Dice*）では異なる特徴を示しているため，どの尺度が最適であるかは一概には決定できない．具体的には，再現率が低い区間，すなわち最も自信を持って負例と判定している区間では *Prob* が，それ以外の区間では共起尺度の方が優れた精度を示している．検定の結果，*R-P* 曲線を比較した場合，各尺度の間に有意な差はなかった．

PLSI は隠れクラス z に基づく一種のスムージングであるため，たとえ真の $P(\langle v, c, n\rangle)$ が低い場合でも，周辺分布 $P(\langle v, c\rangle|z)$，$P(n|z)$，$P(z)$ が高ければ，高く見積もられる．これに対して，$\langle v, c\rangle$ と n の共起尺度を用いることによって，このスムージングによる悪影響を回避することできたと考えられる．ただし逆に，周辺分布が低い場合では，高い場合に比べて尤度推定の信頼性が相対的に低下する．*Prob* と共起尺度の *R-P* 曲線が異なる特徴を示したのは，尤度推定における上記の特徴によるものと考えられる．

図 3 では，$|Z|$ を大きくするにつれて 11 点平均精度は向上するが，試行した $|Z|$ の範囲で収束の傾向を見せている．*LexS* は $|Z| = 1,000$ と $1,500$ で，*LexL* でも $|Z| = 100$ と 200 で精度に有意な差が見られたが，それ以上の区間では精度に有意な差はなかった．

☆ Good-Turing 法では，出現回数 r の補正値として，次の式で示される r^* を用いる．

$$r^* = (r+1)\frac{N_{r+1}}{N_r},$$

ここで，N_r は，学習データ中に r 回出現した $\langle v, c, n\rangle$ の総数である．そして，任意の $\langle v, c, n\rangle$ の出現確率は，次の式で与えられる．

$$P(\langle v, c, n\rangle) = \frac{r^*}{N},$$

ここで，N はあらゆる $\langle v, c, n\rangle$ の総出現回数を表す．

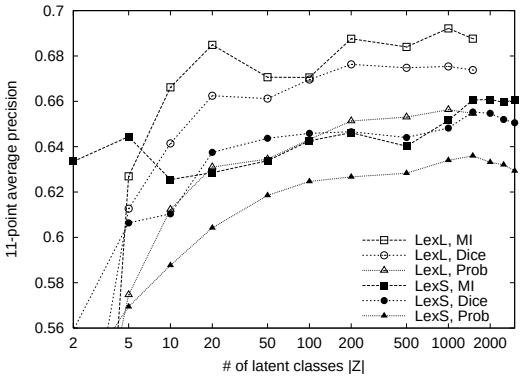

図 3 $|Z|$ に対するモデル Pos の 11 点平均精度
Fig. 3 11-point average precision of Pos over $|Z|$.

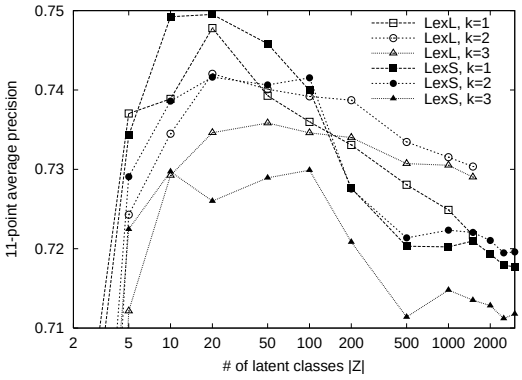

図 4 $|Z|$ に対するモデル Neg の 11 点平均精度
Fig. 4 11-point average precision of Neg over $|Z|$.

すなわち，最も高い精度を得る $|Z|$ の付近で精度は比較的安定しているといえる．$|Z|$ の最適値は，上に示したように対象とする語彙に依存して変化するため，さまざまな $|Z|$ の値について分布クラスタリングを適用し，デヴェロップメントデータに対する誤り検出精度から経験的に推定するしかない．しかし，対象とする語彙を変更することはあまりなく，また，最も高い精度を得る $|Z|$ 付近で精度は安定していることから，人的負担という点では推定のコストは低い．

4.3.2 *Neg* の誤り検出能力と特徴

分布類似度の重みを，最も近い（順位が低い）事例ほどその距離を信頼するという意味で $\lambda_i = 1/i$ とし，5 分割交差検定によってモデル Neg を評価した．

分布クラスタリングにおける隠れクラス数 $|Z|$ と 11 点平均精度の関係を**図 4** に示す．この図が示すように，$LexS$，$LexL$ ともに，11 点平均精度のピークは $|Z| = 20$ 付近にあった．これは，異なる語彙，学習データに対しても容易に Neg を構築できるかということについて，良い見通しを与える結果である．$|Z|$ が小さければ，確率的パラメータの推定に要する時間も少ないため，Neg が最も高い精度を得る $|Z|$ が Pos に比べて容易に推定できる．また，Neg において確率分布 $P(Z|n)$，$P(Z|\langle v, c\rangle)$ の変数が少なくなるため，距離の計算に要するコストも少なくてすむ．

参照する最近傍事例数 k については，$|Z|$ の値によっては $k = 2$ で最も高い 11 点平均精度が得られたが，$|Z|$ を先に決定した場合，すなわち，$|Z| = 20$ では，$k = 1$ の方が高い精度を得ていた（$k = 2$ に対して有意差あり）．共起空間が広大，かつ学習に利用可能な負例がスパースであるため，より多くの最近傍事例を参照すると精度が低下したと考えられる．すなわち，よほど大規模な負例を収集しない限りは $k = 1$ を選択して差し支えない．

図 3 と図 4 を比較すると，Neg の 11 点平均精度は Pos のそれと比較して高すぎるように見えるかもしれないが，これは予測できた結果である．なぜならば，我々が，評価事例集合および負例の収集に用いた言い換え規則集合[6] は，特定の目的で構築されており，生成される表現のバリエーションが限られているためである．すなわち，生成される言い換えに含まれる $\langle v, c, n\rangle$ の種類は $\langle v, c\rangle$ と n のあらゆる組合せを考えるよりは比較的少なく，少数の負例であっても誤りの傾向をとらえるのに十分であったと考えられる．3.1 節で述べたように，言い換えを用いた他のアプリケーションにおいても言い換えの生成能力は限られているため，少数の負例で誤りの傾向をとらえようとする我々のアプローチは現実的であるといえる．

4.3.3 負例の選択的収集と *Com* による誤り検出

前項では，4.1 節で収集した $\langle v, c, n\rangle$ の負例 851 組すべてを用いてモデル Neg を構築し，特徴を分析した．しかし，現実にはあらゆる誤りの傾向をとらえるのに十分な負例を収集するコストを無視するわけにはいかない．そこで，すでに正負例のラベルを付与した 3,704 組の $\langle v, c, n\rangle$ を用い，負例の能動学習の模擬実験を行う．この実験を通じて次の 2 点を確認する．

(i) 少数の負例を用いて構築した Com がどれだけ Pos の精度を向上させることができるか，
(ii) 3.3 節で示した能動学習アルゴリズムがどれだけ効果的か．

4.3.3.1 セッティング

能動学習は 2 度試行した．まず，3,704 組の正負例のラベルを付与した 3 つ組集合をサンプルプールとし，ランダムに 100 組を取り出した．ここで，初期学習事例 S_1 は負例 16 組，S_2 は負例 22 組であった．S_1，S_2 の各々について，残りの 3,604 組から選択的に事例（負例）を収集した．

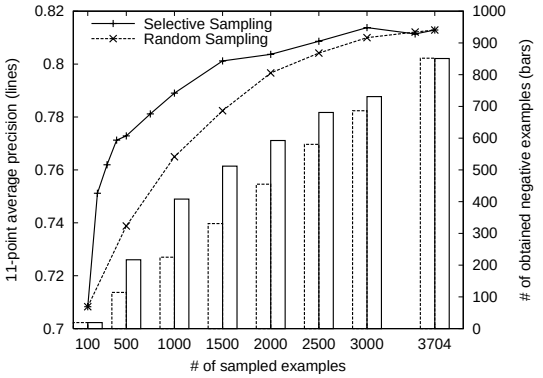

図 5 *Com* の学習曲線．棒グラフ：収集した負例の数，折れ線グラフ：11 点平均精度
Fig. 5 Learning curves of *Com*. Lines: 11-point average precision, bars: # of obtained negative triplets.

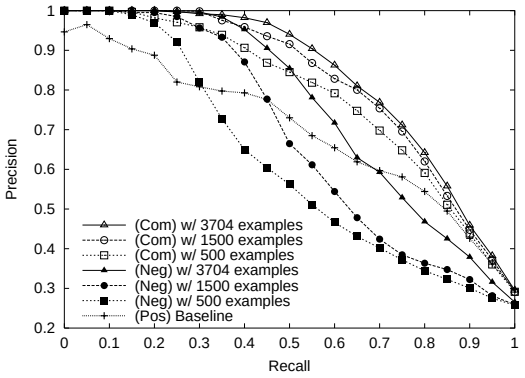

図 6 各学習時点での *R-P* 曲線
Fig. 6 *R-P* curves of our models.

能動学習に用いたパラメータは以下のとおりである．まず，より現実に近い設定として大規模語彙 *LexL* を用いた．*Pos* の尺度としては，*Prob* を選択した．*Prob* は，4.3.1 項において，再現率が低い区間（最も確信を持って負例と判定している区間）で最も高い精度を示したためである．隠れクラス数 $|Z|$ は，すでに考察したとおり，扱う語彙にのみに依存する．したがって，学習の各時点で同じ値を用いる．具体的には，*Pos* には $|Z| = 1{,}000$，*Neg* には $|Z| = 20$ を用いた．これらは，前項までの予備実験において最も高い精度を得たパラメータである．さらに，学習の各時点で利用できる負例が 4.3.2 項よりも少ないため，参照する最近傍事例数は $k = 1$ とした．

学習の各時点で *Com* を構築し，*Neg* と同様に 5 分割交差検定によって誤り検出の性能を評価した．ここでも前項までの予備実験において最も高い精度を得たパラメータを用いた．すなわち，*Pos* については $|Z| = 1{,}000$，*MI*，*Neg* については $|Z| = 20$，$k = 1$ を採用した．モデルの重みは $\beta = 0.5$ とした．

4.3.3.2 実験結果と考察

選択的サンプリング，および比較対象として取り上げたランダムサンプリングによる実験結果を**図 5** に示す．横軸は，サンプリングして人手で正負例のラベルを付与した 3 つ組の数を表す．図 5 の棒グラフと折れ線グラフはそれぞれ，収集した負例の数，*Com* の 11 点平均精度を表す．また，実線と点線はそれぞれ，選択的サンプリングとランダムサンプリングを表す．なお，ここでは初期学習事例 S_1，S_2 を用いた 2 度の試行の平均を示している．

棒グラフが示すとおり，選択的サンプリングでは，ランダムサンプリングと比較して優先的に負例を収集できている．また，折れ線グラフからは，とくに学習初期における顕著な精度の向上を見ることができる．すなわち，収集された負例が，誤り検出においても効果的であったといえる．

図 6 は，*Pos* および学習の各時点での *Neg*，*Com* の *R-P* 曲線を示している．この図からは，まず，低い再現率の範囲では *Neg* の方が *Pos* よりも優れており，高い再現率の範囲では *Pos* の方が *Neg* よりも優れていたことが分かる．これに対して *Com* は，ベースラインである *Pos* の欠点を *Neg* によって補い，あらゆる再現率の範囲で両モデルを上回る精度を得ている．この際，収集した負例の数が少なく，*Neg* 自身が十分な精度を持っていない場合でも，*Pos* と *Neg* を組み合わせることで，つねに両モデルを上回る精度を得ていることから，語彙や事例が変化しても同様にモデルの改良が可能であるといえる．

今回の実験では，利用可能な負例をすべて（851 組）用いて構成した *Com* は 81.3%の 11 点平均精度を示した．同じパラメータの *Pos*，*Neg* の精度をそれぞれ約 12.1 ポイント，約 6.5 ポイント上回っていた．

4.4 パラメータに関する考察

本論文で提案した誤り検出モデル *Com* には次の 5 つのパラメータがある．

1. 隠れクラス数 $|Z|$
2. *Pos* の尺度（*Prob*，*MI*，*Dice*）
3. 参照する最近傍事例数 k
4. *Pos* と *Neg* の線形結合の重み β
5. 各モデルが「誤り」と出力するスコアの閾値

モデルの評価実験に用いた事例集合は，高頻度語のみを対象としており，また比較的小規模であった．このため，新規の学習事例に対しては，パラメータの再推定が必要になる可能性がある．上記の 5 つのパラメータは厳密には独立ではないが，4.3 節で示した実

験結果から，各パラメータは，デヴェロップメントデータを用いて，一つづつ，経験的に推定して差し支えないと考える．以下，各パラメータの性質と推定方法を考察する．

4.4.1 隠れクラス数 $|Z|$

図 3 および図 4 において，11 点平均精度に対する $|Z|$ の影響を観察すると，最も高い精度を得る $|Z|$ 付近で精度は比較的安定している．このことから，準最適なパラメータであれば容易に推定できると結論づける．

4.4.2 *Pos* の尺度

新規の語彙，評価データに対しても，*MI* が最適であるという保証はなく，尺度の再選択が必要になる可能性がある．選択の指標としては次の 2 つが考えられる．すなわち，(i) 誤り検出精度の高さ，および (ii) $|Z|$ に対して精度が敏感であるか否か．前者は，本論文における負例の選択的収集と誤り検出で尺度を使い分けたように，*R-P* 曲線における特性も参照すべきである．後者については，ロバスト性を考えると，$|Z|$ に対して過敏ではないことが望ましい．

4.4.3 参照する最近傍事例数 k

共起空間が広大，かつ学習に利用可能な負例がスパースであるため，より多くの最近傍事例を参照すると精度が低下する．したがって，よほど大規模な負例を収集しない限りは $k = 1$ を選択して差し支えない．

4.4.4 *Pos* と *Neg* の線形結合の重み β

今回は，*Pos* と *Neg* の線形結合の重みは，$\beta = 0.5$ に固定していた．しかし実際，選択的負例収集における各時点で，$0.4 \leq \beta \leq 0.6$ で最も高い精度を得ており，β の値は一定ではなかった．すなわち，最適な値を推定すべきである．学習のどの時点でも，この β の範囲での 11 点平均精度の差は 0.5〜1.5% あり，*R-P* 曲線においても有意な差があった．すなわち，他の $|Z|$ や k などのパラメータに比べて，精度に対して大きく影響していた．したがって，他の比較的安定しているパラメータを決定し，*Com* を構築したうえで推定すべきと考える．この場合，β は *Pos* と *Neg* の混合に関する唯一のパラメータであるため，容易に推定できる．

4.4.5 各モデルが「誤り」と出力するスコアの閾値について

4.2 節で述べたとおり，本論文では，*R-P* 曲線および 11 点平均精度によってモデルを定量評価した．ただし，最終的には，モデルが出力するスコアについて，そのスコア以下の事例を誤りと見なす閾値を決定しなくてはならない．この閾値は，必要な再現率，あるいは精度を達成するという目的に応じて選択する方が自然である．すなわち，他のすべてのパラメータを決定してモデルを構築し，デヴェロップメントデータに対して描画した *R-P* 曲線から決定すればよいと考える．

5. おわりに

本論文では，統計と用例に基づく動詞格構造の誤りの検出モデルを提案した．コーパスから獲得した大規模な正例に基づく統計的言語モデル *Pos* に，人手で収集した小規模の負例に基づいて格構造の不適格さを定量化するモデル *Neg* を混合することで，誤り検出モデル *Com* を構築した．評価実験において，*Com* は，*Pos* に比べて 12.1 ポイント高い，11 点平均精度 81.3%を得た．また，負例の効率的収集を目的とし，能動学習を導入した．ラベル付けの優先度を示す選好関数を提案し，誤り検出に対して貢献度が高い負例を効率良く収集できる可能性を示した．

不適格な動詞格構造の自動検出については，さらに次に示すような調査が考えられる．

- 負例のみを用いる *Neg* は，典型的な負例と正例に囲まれた特異な負例を区別できていない．正負例両方を用い，事例間の距離が正負例の弁別に対してどれだけ敏感であるかをとらえるようにモデルを改良する．
- 言い換えの自動生成というタスク全体における貢献度を評価する．

文献 6) では，言い換えにおける変換誤りの多くが，動詞格構造誤りのように共起の問題，あるいは活用型のように抽象化してとらえられるレベルの問題であることを示した．たとえば，形容詞や名詞述語を主辞とする構造も，本論文で扱った動詞格構造と同様のモデルで適格/不適格を判定できる可能性がある．しかし，本論文で述べたモデルでは検出できない変換誤りも残されている．たとえば，例文 (6) を見てみよう．

(6) s. **遺伝子**を抑えることで成長を抑制できる．
t. *__因子__を抑えることで成長を抑制できる．

例文 (6t) では 〈 抑える, を, 因子 〉 という 3 つ組の共起は適格であるため動詞格構造の誤りとしては検出できない．しかし,「生物の遺伝形質を規定する」という対象としている分野の情報が欠けることによって意味がぼやけてしまうため，言い換えとしては適格ではない．これは，類義語間の細かい意味の違いが捨象されているために生じる変換誤りだと考えられる．岡本ら[20] は，推敲支援における類義語の順位づけを目的として，類義語間の静的な意味の違いを国語辞典の語釈文から抽出する手法を提案している．適格な言い

換えを生成するためには，本論文で扱った統計的アプローチ，および岡本らのアプローチの先に，句や節といった任意の表現間の意味の違い，文脈における適合性をとらえるための枠組みを構築する必要がある．

謝辞 本論文の査読者の方々には，論文を精読していただき，有益なご指摘をいただきました．深く感謝いたします．

参 考 文 献

1) ACL: *The 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (*IWP*) (2003).
2) Brown, P.F., Pietra, S.A.D., Pietra, V.J.D. and Mercer, R.L.: The mathematics of machine translation: parameter estimation, *Computational Linguistics*, Vol.19, No.2, pp.263–311 (1993).
3) Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S. and Tait, J.: Simplifying text for language-impaired readers, *Proc. 9th Conference of the European Chapter of the Association for Computational Linguistics* (*EACL*), pp.269–270 (1999).
4) 日本電子化辞書研究所：EDR 電子化辞書仕様説明書，日本電子化辞書研究所 (1995).
5) 藤田 篤，乾健太郎：語釈文を利用した普通名詞の同概念語への言い換え，言語処理学会第 7 回年次大会発表論文集，pp.331–334 (2001).
6) 藤田 篤，乾健太郎：語彙・構文的言い換えにおける変換誤りの分析，情報処理学会論文誌，Vol.44, No.11, pp.2826–2838 (2003).
7) Germann, U., Jahr, M., Knight, K., Marcu, D. and Yamada, K.: Fast decoding and optimal decoding for machine translation, *Proc. 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.228–235 (2001).
8) Gildea, D.: Probabilistic models of verb-argument structure, *Proc. 19th International Conference on Computational Linguistics* (*COLING*), pp.308–314 (2002).
9) Hofmann, T.: Probabilistic latent semantic indexing, *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR*), pp.50–57 (1999).
10) 池原 悟，宮崎正弘，白井 諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林 良彦 (編)：日本語語彙大系：CD-ROM 版，岩波書店 (1997).
11) Inui, K., Fujita, A., Takahashi, T., Iida, R. and Iwakura, T.: Text simplification for reading assistance: a project note, *Proc. 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (*IWP*), pp.9–16 (2003).
12) 鍛治伸裕，黒橋禎夫，佐藤理史：国語辞典に基づく平易文へのパラフレーズ，情報処理学会自然言語処理研究会予稿集，NL-144-23, pp.167–174 (2001).
13) 河原大輔，黒橋禎夫：用言と直前の格要素の組を単位とする格フレームの自動構築，自然言語処理，Vol.9, No.1, pp.3–19 (2002).
14) Keller, F., Lapata, M. and Ourioupina, O.: Using the Web to overcome data sparseness, *Proc. 2002 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.230–237 (2002).
15) 近藤恵子，佐藤理史，奥村 学：格変換による単文の言い換え，情報処理学会論文誌，Vol.42, No.3, pp.465–477 (2001).
16) Lapata, M., Keller, F. and McDonald, S.: Evaluating smoothing algorithms against plausibility judgements, *Proc. 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.346–353 (2001).
17) Lee, L.: On the effectiveness of the skew divergence for statistical language analysis, *Proc. 8th International Workshop on Artificial Intelligence and Statistics*, pp.65–72 (2001).
18) 宮田高志，宇津呂武仁，松本裕治：Bayesian Network による下位範疇化の確率モデルおよびその学習，情報処理学会自然言語処理研究会予稿集，NL-119-12, pp.77–84 (1997).
19) NLPRS: *Workshop on Automatic Paraphrasing: Theories and Applications* (2001).
20) 岡本紘幸，斎藤博昭：文脈を考慮した日本語類義表現の言い換え，言語処理学会第 9 回年次大会発表論文集，pp.97–100 (2003).
21) 大野 晋，浜西正人：角川類語新辞典，角川書店 (1981).
22) Pereira, F., Tishby, N. and Lee, L.: Distributional clustering of English words, *Proc. 31st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.183–190 (1993).
23) 白井 諭，池原 悟，河岡 司，中村行宏：日英機械翻訳における原文自動書き換え型翻訳方式とその効果，情報処理学会論文誌，Vol.36, No.1, pp.12–21 (1995).
24) Takahashi, T., Iwakura, T., Iida, R., Fujita, A. and Inui, K.: KURA: a transfer-based lexico-structural paraphrasing engine, *Proc. 6th Natural Language Processing Pacific Rim Symposium* (*NLPRS*) *Workshop on Automatic Paraphrasing: Theories and Applications*, pp.37–46 (2001).
25) 徳田昌晃，奥村 学：日本語から手話への機械翻訳における手話単語辞書の補完方法について，

情報処理学会論文誌，Vol.39, No.3, pp.542–550 (1998).
26) Torisawa, K.: An unsupervised learning method for associative relationships between verb phrases, *Proc. 19th International Conference on Computational Linguistics* (*COLING*), pp.1009–1015 (2002).
27) 宇津呂武仁，宮田高志，松本裕治：最大エントロピー法による下位範疇化の確率モデル学習および統語的曖昧性解消による評価，情報処理学会自然言語処理研究会予稿集，NL-119-11, pp.69–76 (1997).
28) 山本和英：換言と言語変換の協調による機械翻訳モデル，言語処理学会第 8 回年次大会発表論文集，pp.307–310 (2002).

(平成 15 年 8 月 25 日受付)
(平成 16 年 2 月 2 日採録)

藤田　篤

1977 年生．2000 年九州工業大学情報工学部卒業．2002 年同大学大学院情報工学研究科博士前期課程修了．同年，奈良先端科学技術大学院大学情報科学研究科博士課程入学．現在に至る．自然言語処理の研究に従事．

乾　健太郎（正会員）

1967 年生．1995 年東京工業大学大学院情報理工学研究科博士課程修了．同年より同研究科助手．1998 年より九州工業大学情報工学部助教授．1998 年〜2001 年科学技術振興事業団さきがけ研究 21 研究員を兼任．2001 年より奈良先端科学技術大学院大学情報科学研究科助教授．現在に至る．博士（工学）．自然言語処理の研究に従事．人工知能学会，ACL 各会員．

松本　裕治（正会員）

1955 年生．1977 年京都大学工学部情報工学科卒業．1979 年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984 年〜1985 年英国インペリアルカレッジ客員研究員．1985 年〜1987 年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993 年より奈良先端科学技術大学院大学教授．現在に至る．工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM 各会員．