自動生成した言い換え文における動詞結合価誤りの自動検出手法

藤田篤[†]　乾健太郎[†]　松本裕治[†]

本稿では，語彙・構文的言い換えにおいて頻繁に生じる動詞結合価誤りの検出方法を提案する．われわれは，コーパスから獲得した大規模な正例に基づいて結合価の適格さを定量化するモデルと，人手で収集した小規模の負例に基づいて結合価の不適格さを定量化するモデルを構築し，これら 2 つのアンサンブルによって，精度の高い誤り検出器を実現した．また，能動学習の採用によって，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ことを確認した．

キーワード：言い換え，テキスト修正，言語モデル，誤り検出，動詞結合価，機械学習

Automatic Detection of Verb Valency Errors in Paraphrasing

FUJITA Atsushi[†]　INUI Kentaro[†]　MATSUMOTO Yuji[†]

This paper argues the issue of transfer errors in paraphrasing. Our previous investigation into transfer errors revealed that verb valency errors occur frequently, irrespective of the types of transfer. Motivated by this finding, we propose an empirical method to detect incorrect verb valences occurring in paraphrasing Japanese sentences. Our error detection model involves ensembling of two error detection models that are separately trained on a large collection of unlabeled positive examples and a small collection of labeled negative examples. An experiment showed that our ensemble method achieved 79.4% 11-point average precision, a 13.3 point improvement over the model trained only on positive examples. We also propose a selective sampling scheme to reduce the cost of labeling examples.

Keywords: parapharasing, text revision, language model, error detection, verb valency, machine learning

1　はじめに

近年，言い換えの自動生成は，自然言語処理のさまざまなタスクに応用可能な要素技術として，注目されるようになってきた [16, 1]．報告されている応用例としては，たとえば，機械翻訳 (MT; Machine Translation) [20]，質問応答 (QA; Question Answering) [14, 19]，情報検索 (IR; Information Retrieval) [12]，あるいは読解支援のための表現の簡単化 [3, 10] などが挙げられる．

QA や IR においては，大規模なテキストデータの中から関連情報をもれなく抽出したいという要求がある．このため，検索表現を拡張するために言い換えが用いられている．この文脈で重要なのは，さまざまな語形変化，表記の揺れ，同義表現などを生成することであり，生成される表現自体が言語的に適格である必要はない．一方，読解支援のように，人間が読むために言い換える場合，言い換えた後の表現が言語的に適格でなくてはならない．ところが，適格な言い換えを生成するために適用条件を書き尽くすことは困難である．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という試みもあるが [17]，任意の表現間の意味の違いを捉える，あるいはあらゆる文脈を考慮して言い換え表現対（言い換えパタン）を獲得するというという報告は現状では存在しない．したがって，言い換えパタンは概して不完全であり，さまざまな変換誤りが生じる可能性がある．

この問題に対し，われわれは，言い換えの自動生成にテキスト修正処理を導入し，その実現可能性と効果を調査している．これまでに，さまざまな種類の語彙・構文的言い換えにおける変換誤りの種類を整理し，各変換誤りが生じる傾向について調査した [5]．

文献 [5] で得られた知見の一つとして，トランスファ規則の種類にかかわらず動詞結合価に関する誤りが頻出するということが挙げられる．動詞結合価に関する誤りとは，例文 (1t) のような選択制限の違反，あるいは例文 (2t) のような動詞格構造の誤りである[1]．

(1) s. 資質と能力を持った「個人」が世代や**国境を超えた**ネットワークで結ばれる。
　　t.*資質と能力を持った「個人」が世代や**国境を上回った**ネットワークで結ばれる。

(2) 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。

[†]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
{atsush-f,inui,matsu}@is.aist-nara.ac.jp

[1]本稿の例中，s., t., r. は各々言い換え前の文，言い換え後の文，修正後の言い換え文を指す．

動詞結合価に関する誤りは，文中の動詞や名詞，慣用表現を別の語句に言い換えた場合，あるいは態交替，動詞交替によって格要素/格助詞が変化した場合に頻出する．とくに前者の語彙レベルの言い換えは，多くの言い換えアプリケーションにおいて重要な役割を担うため，この誤りは，最も優先して解消すべき問題の一つであるといえる．

このような背景から，本稿では，言い換えに必要な修正処理の一つとして，言い換え文中の動詞結合価に関する誤りを自動的に検出する方法について論じる．動詞結合価が適格であるか否かは個々の単語（動詞，名詞）に依存しているため，誤り検出にあたっていくつかの問題がある．これらを踏まえ，本稿では，個々の単語に基づく統計的なモデルを提案する．

以下，2 節では，言い換えにおける誤り検出タスクの性質について述べる．3 節では，動詞結合価の性質について考察し，本稿で取り組むタスクを定義する．コーパスから得られる大規模な正例と人手で作成した小規模な負例に基づくわれわれの誤り検出モデルについて 4 節で述べ，5 節で誤り検出の評価実験について述べる．学習データを効率的に獲得するために導入した能動学習について 6 節で述べ，7 節でまとめる．

2 言い換えにおける誤り検出

われわれは，言い換えの自動生成を次の 3 ステップに分解して考えている．すなわち，

トランスファ: 言い換えパタン，あるいは規則を適用し，ある入力テキストに対して可能な言い換え候補の集合を生成する．

誤り検出: 各言い換え候補を評価し，言語的な誤り，および言い換えとしての誤りを検出する．

修正: 検出した変換誤りを，可能ならば修正し，そうでなければ言い換え候補を棄却する．

言い換えにおける誤り検出・修正処理は，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] と類似しているように見える．どちらも，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，異なる側面もある．言い換えシステムは，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して構築されることが多い．そして，このようなシステムで用いられる言い換えパタンは,「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，デコーディングのように候補を比較してより好ましい候補を選択するだけではなく，誤り検出処理では，適切な言い換え候補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．

また，誤り検出・修正処理の実現にあたって，われわれは，構文レベルの依存構造を仮定している．理由は次の通りである．

- 日本語は比較的語順が自由な言語である．したがって，n-gram などの表層よりも依存構造を扱う方が正確な言語モデルを構築できると考えられる．
- われわれが開発を進めている言い換えエンジン KURA [21] を含めて，既存の日本語言い換えシステムの多くが，依存構造に基づくトランスファ方式で言い換えを実現している．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．

3 動詞結合価に関する誤りの検出

3.1 評価対象

本稿では，トランスファによって置換，あるいは依存構造を変更された単語を含む動詞格構造を対象とする．ここで，動詞格構造とは，一つの動詞とその格要素からなる構造を指すものとする．以下，動詞格構造中の個々の 3 つ組 $\langle$ 動詞 v, 格助詞 rel, 名詞 $n\rangle$ を単に $\langle v, rel, n\rangle$ で表す．

例文 (1t), (2t) は，$\langle$ 超える, を, 国境 $\rangle$, $\langle$ 貫く, に, チームプレー $\rangle$ という，動詞格構造中の一つの $\langle v, rel, n\rangle$ が不適格な例である．一方，次の例文 (3t) は，$\langle$ ある, が, 言葉 $\rangle$, $\langle$ ある, に, 各地 $\rangle$ が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．

(3) 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。

これらの事例から，任意の動詞格構造について結合価の適格/不適格を人手で分類するならば，次のような決定木に準ずるであろう．

動詞結合価の適格性判定

- $\langle v, rel, n\rangle$ が 1 つでも不適格ならば不適格.
- すべての $\langle v, rel, n\rangle$ 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.

ところで，文献 [5] において収集した動詞結合価に関する誤り事例 162 事例のうち，例文 (3t) のように各 $\langle v, rel, n\rangle$ が適格でも兄弟格要素の共起が適格でない事例は 8 事例であった．われわれは，この問題はそれほど深刻ではないと考え，個々の $\langle v, rel, n\rangle$ 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．

動詞結合価の適格性判定（本稿における判定基準）

- $\langle v, rel, n\rangle$ が 1 つでも不適格ならば不適格.
- すべての $\langle v, rel, n\rangle$ が適格ならば適格.

3.2 動詞結合価の誤り検出の難しさ

単純には，人手で構築した動詞結合価辞書に照らすことで，動詞結合価の誤りを検出できると考えられる．ただし，この手法では，以下に示す問題がある．

1. 動詞結合価辞書の規模拡大・洗練にはコストがかかる．コーパスを用いて半自動的に構築・拡充する試みもあるが作業の難易度は高い [6].
2. 既存の結合価辞書は，慣用句などの例外を扱えるほど整備されていない．また，選択制限の指定に用いられている名詞シソーラスも，言い換えにおける単語間の細かな用法の違いを捉えるには粗い[2]．また，受身や使役など，動詞にヴォイスが後続する場合に結合価がどう変化するかも，動詞結合価辞書には記載されていない．
3. 動詞結合価の適格/不適格を決定的に分けることしかできない．緩やかなマッチング手法を考えるとしても，名詞のシソーラス上の距離を用いると，2. の問題に帰結してしまう．

Utsuro ら，Miyata らは，ある名詞がある動詞の下位範疇を満たすか否かを最大エントロピー法，ベイジアン・ネットワークなどの確率モデルで推定している [22, 15]. これらの手法を応用すれば統計的，確率的アプローチによって 1. および 3. の問題は解消できるが，人手で作成したシソーラスを用いるのであれば，2. の問題が残る．

4 提案モデル

4.1 アプローチ

3.2 項で挙げた 2. の問題を避けるために，シソーラスのような人手で構築された知識を用いず，個々の単語そのものを扱う．動詞は，ヴォイスが付属して構文的な受身，使役の形を取る場合は，元の動詞と区別する．

ここで，誤り検出器を構築するためのデータに関する制約がある．誤り検出を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して捉えるのであれば，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考えられる．しかしながら，正例がコーパスから大規模に獲得できるのに対して，負例は大規模には入手できない．さらに，個々の単語そのものを扱うとなると，$\langle v, rel, n\rangle$ の共起空間は非常に大きくなるため，事例は非常にスパースである．ゆえに，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ても有効に働くとは期待できない．

[2]NTT 日本語語彙大系で約 2,700 クラス．われわれは，文献 [4] において，より細かい粒度のシソーラスの同概念語間の言い換えでさえも結合価誤りが生じることを観察している．また，(ほぼ) 同じ意味の単語であるからといって，用法まで同じであるとは限らない．

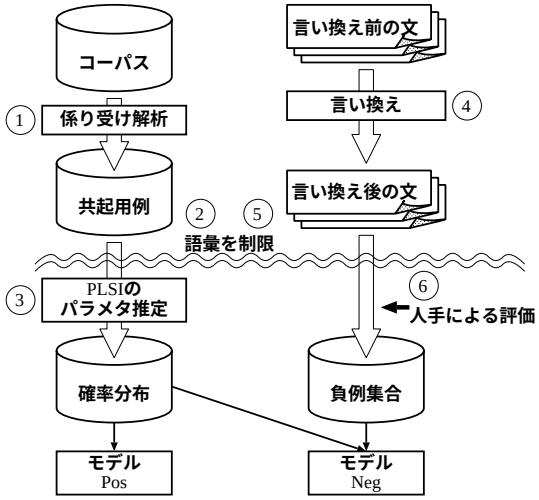

図 1: 提案モデル

大規模な正例とごく少数の負例を効率よく組み合わせる手法として，われわれは，正例，負例のそれぞれのみを用いた誤り検出モデルを別々に構築し，これら 2 つを組み合わせる手法を提案する．以降，正例のみに基づく誤り検出モデルを Pos，負例のみに基づく誤り検出モデルを Neg，および両モデルのアンサンブルモデルを Ens と表す．以下本節では，各モデルについて詳述する（図 1 も参照されたい）．

4.2 統計に基づく誤り検出モデル Pos

モデル Pos は正例のみを用いて構築する．コーパスから獲得できる正例は極めて大規模であるため，さまざまな単語について 3 つ組 $\langle v, rel, n\rangle$ の生起確率 $P(\langle v, rel, n\rangle)$ を見積もった場合，ある程度信頼できると考えられる．$P(\langle v, rel, n\rangle)$ を推定するさまざまな手法の中で，今回は，分布クラスタリング [18] に基づいて単語の共起を潜在的な意味からの同時発生とみなす PLSI (Probabilistic Latent Semantic Indexing) [8] を検討した．また，ベースラインとして，頻度のディスカウントに基づく Good-Turing 法を用いた．

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

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

ここで，z は共起の潜在的な意味クラス（隠れクラス）を指す[3]．式中の確率的パラメタ $P(\langle v, rel\rangle|z)$，$P(n|z)$，$P(z)$ は，EM アルゴリズムによって推定できる [8].

$P(\langle v, rel, n\rangle)$ 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，$\langle v, rel, n\rangle$ 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 ($Prob$)，$\langle v, rel\rangle$ と n の相互情報量 (MI)，および $\langle v, rel\rangle$ と n の

[3]本タスクの文脈では，『$\langle v, rel, n\rangle$ の共起の意味』を表すものと考える．トランスファ前後の語の意味を比較するのとは異なる．

Dice 係数 ($Dice$) の 3 つを検討した．すなわち，モデル Pos は，任意の入力 $\langle v, rel, n\rangle$ に対して，$Prob$，MI，もしくは $Dice$ をスコアとして出力する．

$$MI(\langle v, rel, n\rangle) = \log \frac{P(\langle v, rel, n\rangle)}{P(\langle v, rel\rangle)P(n)}$$

$$Dice(\langle v, rel, n\rangle) = \frac{2 \times P(\langle v, rel, n\rangle)}{P(\langle v, rel\rangle) + P(n)}$$

4.3 負例に基づく誤り検出モデル Neg

3 つ組 $\langle v, rel, n\rangle$ の共起空間は非常に広大であるため，$P(\langle v, rel\rangle)$，もしくは $P(n)$ が低い場合，Pos が推定する $\langle v, rel, n\rangle$ の尤度は信頼度が低くなる．この欠点は，負例を用いることで補うことができる可能性があるが，Good-Turing 法や PLSI の確率計算に負例を用いることはできない．

そこで，負例のみに基づく誤り検出モデル Neg を構築した．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-最近隣検索法 (k-Nearest Neighbor) を採用した．入力の $\langle v, rel, n\rangle$ と任意の学習済事例 $\langle v', rel', n'\rangle$ との距離を計算するための素性として，Pos の構築において得られる，隠れクラス $z \in Z$ への帰属確率分布 $P(z|\langle v, rel, n\rangle)$ を用いた．また，任意の事例間の距離は，確率分布同士の分布類似度で与えた．分布類似度の尺度としては，Jensen-Shannon divergence，および α-skew divergence を用いた．文献 [13] において，この 2 つの分布類似度は優れた実験結果をもたらすと報告されている[4]．

以下，入力の $\langle v, rel, n\rangle$，および任意の学習済事例 $\langle v', rel', n'\rangle$ の，隠れクラスへの帰属確率分布をそれぞれ $q = P(z|\langle v, rel, n\rangle)$，$r = P(z|\langle v', rel', n'\rangle)$ とし，各分布類似度の定義を示す．

Jensen-Shannon divergence (JS)

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

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

ここで，$ave_{q,r}$ は q と r の平均であり，D は Kullback-Leibler (KL) divergence である．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．

α-skew divergence (S_α)

$$S_\alpha(q, r) = D\left(q \,\|\, \alpha \cdot r + (1 - \alpha)\, q\right)$$

S_α では，q と r の重み付き平均を KL divergence の参照側確率分布としており，式中のパラメタ α ($0 \le \alpha \le 1$) がその重みである．ただし，事前に最適な α の値を推定することはできない [13].

モデル Neg が入力の $\langle v, rel, n\rangle$ に対して与えるスコア $Score_{Neg}$ は，$\langle v, rel, n\rangle$ とその k 個の最近隣事例との分布類似度 d_i (JS または S_α) の重みつき平均とした．

$$Score_{Neg} = \frac{1}{k}\sum_{i=1}^{k} \lambda_i\, d_i\,(q, r)$$

ここで，i は類似度の順位であり，λ_i は，個々の近隣事例に対する重みである．

4.4 Pos と Neg のアンサンブルモデル Ens

モデル Pos は正例のみを，モデル Neg は負例のみを用いて構築しているため，それぞれ異なる性質を持つと想像できる．この 2 つのモデルを組み合わせることによって各々が相補的に働くと予想し，アンサンブルモデル Ens を構築した．今回は，モデル Pos，Neg が出力するスコア自身を用いて Ens 自身のスコア $Score_{Ens}$ を決定した．

モデル Pos は確率 ($0 \le Prob \le 1$)，あるいは共起尺度 ($-\infty \le MI, Dice \le \infty$)，モデル Neg は分布類似度 ($0 \le Score_{Neg} \le \infty$) というように，出力されるスコアの尺度は異なる．そこでまず，各モデルのスコアを信頼度 C_{Pos}，C_{Neg} ($0 \le C_{Pos}, C_{Neg} \le 1$) に写像する．具体的には，訓練データの交差検定によって，モデルが出力するスコアに対するモデルの精度を信頼度とする写像関数を導出する．Ens は，次式に示す C_{Pos} と C_{Neg} の重み付き平均をスコアとして出力する．

$$Score_{Ens} = \beta C_{Pos} + (1 - \beta) C_{Neg}$$

ここで，β ($0 \le \beta \le 1$) はモデルの重みである．

5 誤り検出実験

5.1 モデルの訓練および評価事例の作成

モデルの訓練，および評価事例の作成の手順を以下に示す．図 1 も参照されたい．

1. 新聞記事 19 年分[5]の係り受け解析結果[6]から，のべ 53,157,450 組，異なり 7,993,331 組の $\langle v, rel, n\rangle$ を収集した．名詞は 65,384 語，動詞は 33,884 語であった[7].
2. のべ 2,000 回以上出現した名詞 3,365 語，動詞 2,516 語に語彙を制限した．格助詞も，頻度の高い “が”，“を”，“に”，“で”，“へ”，“から”，“より” の 7 つと

[4] “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．

[5] 毎日新聞 9 年分，日経新聞 10 年分，のべ 25,061,504 文.

[6] 係り受け解析には CaboCha [11] を用いた.

[7] 格の交替現象を直接扱うため，“れる/られる”，“せる/させる”，などの接尾辞を後続する場合は，これを含めて動詞一語としている．

した．1. で得た 3 つ組のうち，異なりで 3,628,345 組がこの制限語彙で表現されており，$\langle v, rel \rangle$ の異なりは 16,899 種，$\langle v, rel \rangle$ と n の共起行列要素充填率は 6.38%であった．

3. 2. で得た 3 つ組を PLSI 学習パッケージ[8] に入力し，確率的パラメタを推定した．隠れ変数の個数 $|Z|$ は 100，200，500，1,000，1,500，2,000 とした．
4. 評価事例（かつモデル Neg の訓練に用いるための負例）を作成するために，日経新聞 2000 年分の記事からランダムに取り出した 90,000 文を言い換えエンジン KURA [21] に入力し，さまざまな種類の約 28,000 のトランスファ規則 [5] を網羅的に適用して言い換え事例 7,167 事例を生成した．
5. 4. で生成した言い換え事例から，動詞格構造が変化しており，かつ動詞格構造のすべての名詞，動詞，格助詞が 2. で制限した語彙に含まれる 3,169 事例を取り出した．
6. 最後に，5. で得た事例，および事例に含まれる 3,706 組の $\langle v, rel, n \rangle$ を人手で適格/不適格に分類した[9]．結果，事例としては正例 2360，負例 809 を，$\langle v, rel, n \rangle$ としては正例 2,854，負例 852 を得た．ここで得た事例を誤り検出実験の評価に用い，$\langle v, rel, n \rangle$ の負例のみをモデル Neg の学習に用いた．

5.2 評価方法

動詞結合価に関する誤り検出実験の結果は，以下に示す再現率 R，および精度 P で評価した．

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

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

モデルが出力するスコア（確信度）に対して閾値を設けることで，モデルを 2 値分類器として扱うことができる．すなわち，閾値以下の値を持つ事例を変換誤り，閾値を超える値を持つ事例を適格な事例として分類することになる．また，この閾値を変化させることで，ラベル付けする事例の数を制御し，再現率- 精度カーブ（R-P カーブ）を描画することができる．個々の R-P カーブは，$R = 0.0, 0.1, \ldots, 1.0$ の 11 点平均精度で評価した．

5.3 実験結果

5.1 項で作成した 3,169 事例を用いて動詞結合価の誤り検出の実験を行った．以下，各誤り検出モデルごとに 11 点平均精度を示し考察する．また，比較のために，既存の動詞結合価辞書を用いた誤り検出の精度も示す．

表 1: モデル Pos の 11 点平均精度 (PLSI)

尺度 \ $\lvert Z \rvert$	100	200	500	1,000	1,500	2,000
$Prob$	0.6247	0.6267	0.6283	0.6340	0.6360	0.6331
MI	0.6426	0.6462	0.6402	0.6515	0.6606	**0.6608**
$Dice$	0.6459	0.6466	0.6440	0.6482	0.6553	0.6547

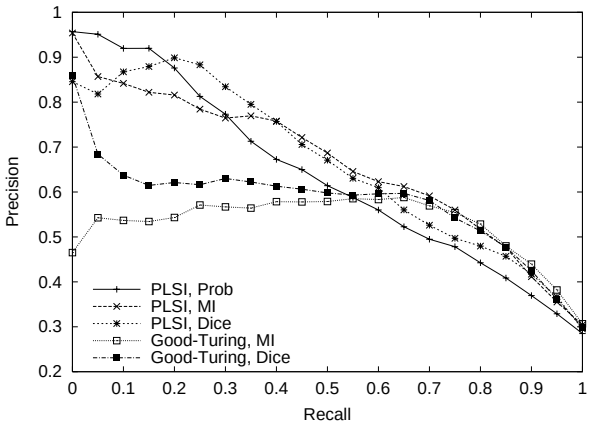

図 2: モデル Pos の R-P カーブ ($|Z| = 2,000$)

5.3.1 Pos の誤り検出能力

モデル Pos の誤り検出能力の評価結果を表 1，および図 2 に示す．MI と $Dice$ はほぼ同等の 11 点平均精度を得ており，共起確率そのものを分類に用いた場合 ($Prob$) よりもわずかながら優れていた．比較的低コストな Good-Turing 法を用いて共起確率を推定した場合，11 点平均精度は，$Prob$ が 0.3951[10]，MI が 0.5178，$Dice$ が 0.5793 であった．図 2 が示す通り，PLSI を用いて共起確率を推定した方が Good-Turing 法よりも顕著に優れた精度を得ており，計算コストをかける価値はあったといえる．また，最も高い 11 点平均精度を得たのは，$|Z| = 2,000$，PLSI，MI であった．

5.3.2 Neg の誤り検出能力

モデル Neg は 5 分割交差検定によって評価した．分布類似度の重みは，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，さまざまなパラメタの組み合わせによって得られた 11 点平均精度を表 2 および表 3 に示す．また，各パラメタに関する考察結果を以下に示す．

隠れクラス数について ：Pos と同様に，PLSI の隠れクラス数 $|Z|$ を大きくするほど高い精度を得た．表 2

[8] http://cl.aist-nara.ac.jp/~taku-ku/software/plsi/

[9] 人手による事例評価は 3.1 項 1 つ目の決定木に基づく．すなわち，兄弟格要素の共起も考慮している．

[10] Good-Turing 法において共起確率をスコアとして用いることは，共起事例の頻度を用いるのと同義である．

表 2: モデル Neg の 11 点平均精度 ($k = 1$)

尺度 \ $\|Z\|$	100	200	500	1,000	1,500	2,000
JS	0.6002	0.6213	0.6790	0.6961	0.7115	0.7134
$S_{0.01}$	0.6366	0.6533	0.6923	0.6976	0.7140	0.7165
$S_{0.25}$	0.6105	0.6284	0.6856	0.6996	0.7138	**0.7185**
$S_{0.5}$	0.6060	0.6230	0.6821	0.7004	0.7145	0.7177
$S_{0.75}$	0.6030	0.6250	0.6807	0.6978	0.7137	0.7150
$S_{0.99}$	0.6082	0.6341	0.6843	0.6990	0.7090	0.7117

表 3: モデル Neg の 11 点平均精度 ($|Z| = 2,000$, $S_{0.25}$)

k	1	2	3	5	10
11 点平均精度	**0.7185**	0.7044	0.6930	0.6816	0.6730
k	15	20	25	30	
11 点平均精度	0.6697	0.6674	0.6655	0.6638	

が示す通り，精度は $|Z| = 2,000$ では収束の傾向を見せている．

分布類似度の尺度について ：S_α は α の値によっては JS を上回る精度をもたらした．ただし，$|Z|$ と最も優れた精度をえる α の値に相関は見られない．これらは，文献 [13] と同様の観察結果である．

参照する最近隣事例の個数について ：$|Z| = 1,500$〜$2,000$ では，試行したすべての尺度のもとで $k = 1$ で最も良い 11 点平均精度を得た．また，$|Z| = 100$〜$1,000$ でも，ほとんどの場合，$k = 1$ または $k = 2$ で最も良い 11 点平均精度を得た．共起空間が非常に広大かつスパースであるため，複数の最近隣事例を参照したことがノイズになった，また，$|Z|$ を大きくする（より多くの素性を考慮する）ことによって距離の計算が精密になり，ノイズの影響が大きくなったと解釈できる．

表 2 および表 3 に示したさまざまなパラメタの組み合わせの結果，$|Z| = 2,000$，$S_{0.25}$，$k = 1$ で最も高い 11 点平均精度を得た．

5.3.3 Ens の誤り検出能力

誤り検出モデル Pos，Neg のアンサンブルモデル Ens の構築には，それぞれの評価において最も高い 11 点平均精度を得たパラメタを採用した．すなわち，両モデルについて $|Z| = 2,000$，Pos については PLSI，MI，Neg については $S_{0.25}$，$k = 1$ である．

Neg と同様に 5 分割交差検定による評価を，表 4，および図 3 に示す．表 4 は，Pos，Neg（が出力するスコアに対する信頼度）に対する重み β を変化させた場合の 11 点平均精度を示している．現状のパラメタでは，Neg をやや重視した場合 ($\beta = 0.4$) に最も高い 11 点平均精度，約 79.4% (0.7937) を得た．これは，Pos の最高精度を約 13.3 ポイント，Neg の最高精度を約 7.5 ポイント上回っている．図 3 からは，まず，低再現率の範囲では Neg の方が Pos よりも優れており，高再現率の範囲では Pos の方が Neg よりも優れているという特徴を持っていたことが分かる．これに対して Ens は，あらゆる再現率の範囲で両者をしのぐ精度を得ており，Pos と Neg が相補的に作用しているといえる．

表 4: モデル Ens の 11 点平均精度 ($|Z| = 2,000$, PLSI, MI, $S_{0.25}$, $k = 1$)

β	0.1	0.2	0.3	0.4	0.5
11 点平均精度	0.7506	0.7751	0.7899	**0.7937**	0.7904
β	0.6	0.7	0.8	0.9	
11 点平均精度	0.7778	0.7520	0.7235	0.6899	

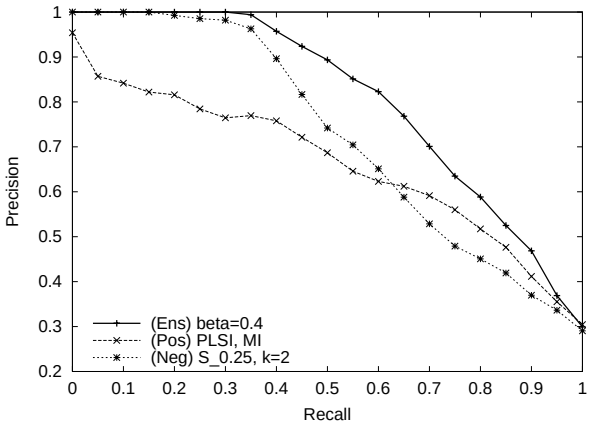

図 3: モデル Ens の R-P カーブ ($|Z| = 2,000$, $\beta = 0.4$)

5.4 比較実験

比較のため，5.1 項で作成した事例集合に対する結合価辞書ベースの誤り検出精度を示す．

われわれが構築したモデルと異なり，動詞結合価辞書が，評価用事例集合中のすべての動詞を評価できるとは限らない．実際，NTT 日本語語彙大系 [9] にエントリを持つ動詞は，全 $\langle v, rel, n\rangle$（のべ）中，89.3%であった（被覆率 89.3%）．そして，下位範疇および選択制限を満たすか否かで個々の $\langle v, rel, n\rangle$ の適格/不適格を判定し，それらの組み合わせによって事例の適格/ 不適格を判定した結果，誤り検出の再現率は 59.7%，精度は 42.8%であった．高頻度語のみ用いたため被覆率は高いが，誤り検出能力が高いとは言えない．

6 能動学習の導入

負例を用いることによって，統計のみに基づくモデルと比較して誤り検出の精度を向上させることができた．しかしながら，前節で示した実験では高頻度語のみを対象としていたということを無視するわけにはいかない．より多くの語を扱えるように Neg，および Ens を訓練するには，より大規模かつ多様な負例が必要になる．したがって，負例を収集するコストをいかにして削減するかが次の課題となる．われわれはこの課題に対して，能動学習を試行した．

6.1 選択的サンプリング

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．

われわれの誤り検出モデル Neg においては，次の性質を満たす 3 つ組 $\langle v, rel, n\rangle$ が，有益な学習事例になると考えられる．すなわち，

選好基準 1 正例ではない．

選好基準 2 すでに収集済の負例との類似度が低い（距離が大きい）．

これらを考慮して，ラベル付けの候補となる 3 つ組に対して優先度を与え，もっとも優先度が高い 3 つ組に人手で適格/不適格のラベルを付与する．以下，3 つ組を便宜的にサンプルと呼ぶ．

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，サンプルと既にラベル付けされている負例との類似度は，Neg によって算出できる．Pos が出力する，サンプルの生起確率を p，Neg が出力する，サンプルとその最近隣事例（負例）の距離を s として，次のような選好関数を作成した．

$$Preference \quad = \quad -s\log(p)$$

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 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予測し，学習済の負例との距離 s を重視するため，p の対数を用いてスケーリングしている．

6.2 実験

能動学習は次に示す手順で行った．

1. 5.1 項で収集した 3,706 組の $\langle v, rel, n\rangle$ からランダムに 100 サンプルを取り出し，初期の学習事例とした．内訳は，正例 84，負例 16 であった．残りの 3,606 サンプルを能動学習の候補とした．
2. 学習の候補の各々に対して，Pos および Neg によって，p，s を算出し，先に示した選好関数によって優先度を付与する．優先度を算出するための Neg パラメタとしては $|Z| = 2,000$，$S_{0.25}$ を用いた．
3. もっとも優先度が高いサンプルを取り出し，適格/不適格のラベルを付与する．不適格（負例）であれば，Neg の学習事例に追加し 2. に戻る．適格（正例）であれば，次に優先度が高いサンプルを取り出す．
4. 候補が無くなるまで 2.，3. を繰り返す．

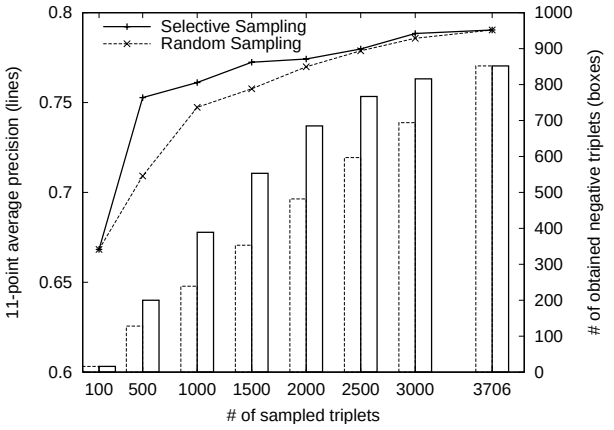

図 4: Ens の学習曲線 （棒グラフ:収集した負例の数，折れ線グラフ:11 点平均精度）

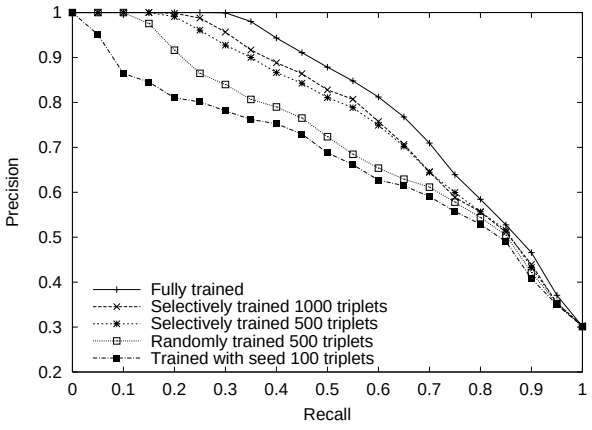

図 5: Ens の R-P カーブ

6.3 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ランダムサンプリングによる学習の結果を図 4 に示す．図 4 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，11 点平均精度を表している．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している．また，学習の各時点での R-P カーブを図 5 に示す．

図 4 の棒グラフが示す通り，われわれが用いた選好関数によって，ランダムサンプリングと比較して優先的に負例を選択できている．また，折れ線グラフからは，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うことが分かる．とくに，学習の初期段階で顕著な精度の向上が見られる．ランダムサンプリングでは 3 つ組 1,500 個にラベル付けした時点で 11 点平均精度 75%を達成しているのに対し，選択的サンプリングでは，3 つ組 500 個にラベル付けした時点で早くも同じ精度に到達している．11 点平均精度の伸びは，サンプル数 100〜500 に比べて，500〜3,000 では緩やかになっている．これは，学習の候補の中で最も有効

な 3 つ組のほとんどが学習初期で選択されたからだと解釈できる．また，図 5 の R-P カーブからも，選択的サンプリングの方がランダムサンプリングに比べて早い段階で顕著に良い精度を得ていることが分かる．

7 おわりに

本稿では，語彙・構文的言い換えにおいて頻繁に生じる動詞結合価誤りの検出方法を提案した．われわれは，コーパスから獲得した大規模な正例に基づいて結合価の適格さを定量化するモデル Pos，人手で収集した小規模の負例に基づいて結合価の不適格さを定量化するモデル Neg，およびこれら 2 つをアンサンブルした誤り検出モデル Ens を構築した．評価実験では，11 点平均精度で 79.4%を得た．これは，正例のみを用いた誤り検出モデルよりも 13.3 ポイント高い精度である．また，能動学習の採用し，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．

引き続き，次に示す課題に取り組む予定である．

- 言い換え全体に対する貢献度を評価する．
- 大規模語彙でモデルを構築し，誤り検出精度，能動学習において同様の傾向が得られるか否かを検証する．

参考文献

[1] ACL. *The 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP)*, 2003.

[2] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of machine translation: parameter estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[3] John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 269–270, 1999.

[4] 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.

[5] 藤田篤, 乾健太郎. 語彙的言い換えに必要な知識の部品化. 情報処理学会自然言語処理研究会予稿集, NL-149-5, pp. 31–38, 2002.

[6] Sanae Fujita and Francis Bond. A method of adding new entries to a valency dictionary by exploiting existing lexical resources. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pp. 42–52, 2002.

[7] Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. Fast decoding and optimal decoding for machine translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 228–235, 2001.

[8] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 50–57, 1999.

[9] 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.

[10] Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. Text simplification for reading assistance: a project note. In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP)*, pp. 9–16, 2003.

[11] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[12] 黒橋禎夫. 言葉の意味を計算機で扱う. 言語処理学会第 6 回年次大会チュートリアル資料, pp. 21–28, 2000.

[13] Lillian Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics*, pp. 65–72, 2001.

[14] Dekang Lin and Patrick Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, Vol. 7, No. 4, pp. 343–360, 2001.

[15] Takashi Miyata, Takehito Utsuro, and Yuji Matsumoto. Bayesian network models of subcategorization and their MDL-based learning from corpus. In *Proceedings of the 4th Natural Language Processing Pacific Rim Symposium (NLPRS)*, pp. 321–326, 1997.

[16] NLPRS. *NLPRS'01 Workshop on Automatic Paraphrasing: Theories and Applications*, 2001.

[17] Hiroyuki Okamoto, Kengo Sato, and Hiroaki Saito. Preferential presentation of Japanese near-synonyms using definition statements. In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP)*, pp. 17–24, 2003.

[18] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 183–190, 1993.

[19] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 215–222, 2002.

[20] 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.

[21] Tetsuro Takahashi, Tomoya Iwakura, Ryu Iida, Atsushi Fujita, and Kentaro Inui. KURA: a transfer-based lexico-structural paraphrasing engine. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS) Workshop on Automatic Paraphrasing: Theories and Applications*, pp. 37–46, 2001.

[22] Takehito Utsuro, Takashi Miyata, and Yuji Matsumoto. Maximum entropy model learning of subcategorization preference. In *Proceedings of the 5th Workshop on Very Large Corpora (VLC)*, pp. 246–260, 1997.