情報処理学会研究報告
IPSJ SIG Technical Report

用法の違いを考慮した類似単語の置換による学習データ生成とそれを用いた主題の違いに頑健な言語モデルの構築

清水信哉[†1]　鈴木雅之[†1]　齋藤大輔[†1]
峯松信明[†1]　広瀬啓吉[†1]

話し言葉の言語モデルを構築する際には，タスクにマッチした学習データが十分量得られないことが多い．そのデータの不足を補うため，学習データ中の単語を類似単語に置換することにより新たに学習データを生成するという手法を提案する．しかし，単語は文脈によって意味が変わるため，単に置換する単語のペアをあらかじめ決めておくことは適切ではない．そこで本稿では，係り受け関係に着目し，文脈に依存して置換する単語が変化するような指標を提案する．さらに，CSJ（日本語話し言葉コーパス）を用いた評価実験を行い，本手法の有効性を示した．

Training of robust language models by automatic sentence generation based on word replacing words with respect to their contexts

Shinya Shimizu ,[†1] Masayuki Suzuki,[†1]
Daisuke Saito ,[†1] Nobuaki Minematsu [†1]
and Keikichi Hirose[†1]

In some cases, a sufficient amount of training data is not available to build language models, especially those for spontaneous speech.To solve this problem, this paper proposes a new method of automatic generation of in-domain sentences based on word replacement. However, we cannot adopt a simple and static strategy for replacing words because the meaning of words easily vary depending on its context. Which word to be replaced with which word? Adequate sets of word pairs should be changed dynamically. Therefore, in this paper we propose a method of using a local dependency structure within a sentence as context, which can adequately select words to replace a given word with. Experimental results using CSJ (Corpus of Spontaneous Japanese) show the validity of the proposed method.

1. はじめに

連続音声認識においては，タスクに応じて適合した言語モデルを利用することが望ましいが，特に自由発話に近いタスクにおいて，タスク適合度の高い言語モデルを構築するのが難しいことが知られている[1)]．

その理由の一つは，そもそものモデル化が困難なことである．より自由発話に近いタスクでは，語彙や発話のスタイルが固定されておらず，言い直しなど不規則な表現も存在するためである．もう一つは，言語モデルの学習に用いることのできるデータ量が不十分なことである．新聞記事の読み上げやニュースでは，タスクに適合した書き起こしテキストを十分に用意して言語モデルの学習に用いることができる．それに比べて自由発話の書き起こしテキスト作成には多くの時間とコストがかかり，大量のテキストを学習データとして用いることは困難である．それに加え，対話か講演か，目下に対するものか目上に対するものかなど，タスクによって大きく発話スタイルが異なり，スタイルの適合した大量の書き起こしテキストを常に用意することはさらに困難である．そこで，数年に渡る新聞データなどから汎用性の高い言語モデルを構築し，少量のタスク適合データを用いて言語モデル適応を行う手法などが提案されている[2)]．

また，タスク適合度の高い言語モデルを構築する手法として，適合度の高いデータを擬似的に自動生成し，それによって言語モデルを学習する方法がある[3),4)]．本稿では，学習データの自動生成についての一手法を提案する．自動生成の際に注意すべきこととして，(1) 日本語として不適切な表現を生成しない，(2) タスクに適合しない語系列を生成しない，ということなどが挙げられる．提案手法では，小規模なコーパス中の語を類似語と置換することにより文を生成する．そのためまず，大規模だがタスク適合度の低いコーパスから格フレームという形で日本語に対する知識を獲得する．それを置換する語の選定基準として用いることにより，(1) を満たすような文生成を行う．また，小規模なコーパスとして発話スタイルの適合したコーパスを用いることにより，(2) の条件のうち，特にスタイルが適合しない語系列生成の回避を試みる．

†1 東京大学
The University of Tokyo

2. 学習データの自動生成による言語モデル学習

学習データ量の不足を補うためにタスクに適合した言語データを自動生成する手法として，いくつかのものが提案されている．例えば秋田らは，統計的話し言葉変換によって話し言葉認識のための言語モデルを学習するという手法を提案している[4]．この手法ではまず話し言葉の性質をコーパスから抽出し，話し言葉でないコーパスから話し言葉へのスタイル変換アルゴリズムを構築する．スタイル変換アルゴリズムとしては機械翻訳の枠組みを用い，文書スタイルと話し言葉スタイルのパラレルコーパスを用いて学習を行っている．それにより，話し言葉ではないコーパスから話し言葉のコーパスを生成し，N-gram 言語モデルを学習する．また太田らは，フィラー書き起こしのないコーパスからのフィラー付き言語モデル学習を提案している[3]．これは，フィラーの挿入アルゴリズムを用いて，フィラーの書き起こされていないコーパスからフィラーを含むコーパスを生成し，N-gram 言語モデルを学習するという手法である．これらの手法は一般的なコーパスから特定の発話スタイルへ，具体的には話し言葉でないコーパスから話し言葉への変換を試みたものである．

一方，本稿で提案する手法では，タスクに対し発話スタイルは適合しているがデータ量が少なく主題も偏ってしまっているような学習データしか得られない場合に，学習データ中の単語を類似単語と置換することにより言語データを自動生成し学習データを増加させる．それにより，スタイルを維持したままデータ量を増やすと同時に主題の偏りを緩和することを試みる．また，本手法による文生成アルゴリズムはスタイルの変換を学習するのではなく，単に単語の置換を学習するだけである．そのため，スタイルごとに文生成アルゴリズムを学習する必要が無いという利点も存在する．

3. 用法の違いを考慮した単語の置換

単語の置換により適切に文を生成するためには，どの単語をどの単語に置換するのが適当かを示す指標が必要になる．その指標として，あらかじめ単語間の類似度を定義しておきそれを利用することがまず考えられる．しかし，単語は文脈によって意味，用法が変化する．ある単語がある文脈では別のある単語とほぼ同義であり，置換可能であったとしても，別の文脈では同義にならず，置換するのが適当ではないということが起こる．ここで言う文脈とは本来，前後数単語から主題，さらには話者や周囲の状況などの非言語情報をも含めたものであるが，それらを全て考慮するのは不可能であり，文脈としてどの情報を採用するかが問題となる．このような意味，用法の曖昧性解消のためには様々な手法が検討されているが，日本語においては，例えば動詞についてみると，直前の格要素（その動詞に係る名詞とその格）を考慮することによりその用法をおおよそ決定できるとされている[5]．

そこで本手法では，名詞と動詞の係り受け関係を文脈として利用する．すなわち，例えば同じ名詞でも，どの動詞と係り受け関係にあるかに応じて置換可能な単語が変化するものとする．さらに，名詞と動詞の係り受けに関する知識として格フレームを用いる．さらに格フレームに対し格ごとに確率的潜在意味解析 (Probabilistic Latent Semantic Analysis, 以下 PLSA)[6] を適用してクラスタリングを行い，係り受けを考慮した置換単語の選択基準に用いる．

3.1 格フレーム

格フレームとは，用言と名詞の関係を，用言の用法ごとに記述したものである．本稿では表層格 9 格 (ガ格，ヲ格，ニ格，カラ格，ヘ格，ト格，ヨリ格，マデ格，デ格) のみを用いた簡易な格フレームを構築し利用した．さらに用言と名詞のペア抽出に関しても，(本来は係り受け解析だけでは不十分であるが，) 今回は単に係り受け解析を行い，直接の係り受け関係にあるものを抽出した．近年，web を用いて大規模な格フレームが構築されており[7]，これを用いることも可能である．

3.2 格フレーム抽出と PLSA によるクラスタリング

まず，格フレーム抽出の手順を**表 1**(a) に示す．学習データ中の各文に対し係り受け解析を行い，係り受け関係にある二節が以下の三つの条件を満たす時のみ，格フレームとして抽出した．

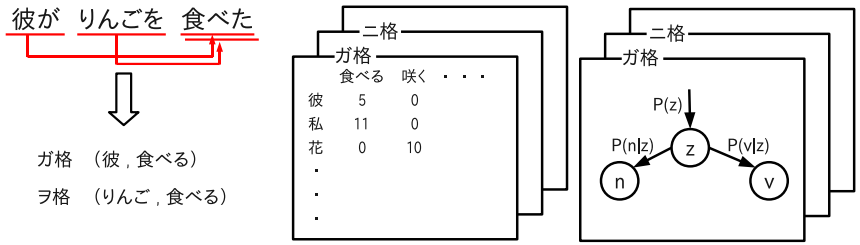
- 係り元の節に「名詞+助詞」の形がある
- その助詞は日本語表層格 9 格のいずれか，または「は」である
- 係り先は動詞又はサ変接続名詞+「する」の活用形，である

ここで，サ変接続名詞+「する」というのは,「心配する」「集合し」などである．これを擬似的に動詞として扱う．さらに以上の条件を満たすとき，さらに以下の二つの処理を行って格フレームとして加算するものとした．

- 助詞「は」をガ格として扱う
- 動詞の原形化

ここで，本来であれば助詞「は」は一律「が」に変更して良いものではないが，今回は簡易なものとして一律の変更を行った．

以上の処理の後，図 1(b) のような名詞-動詞行列が格ごとにできる．これに PLSA を適用する．PLSA は，文書と単語の出現確率を，トピックという隠れ変数により特徴付けてモ



(a) 係り受けの抽出　(b) 名詞-動詞行列　(c) 格ごとの PLSA

図 1　格フレームの抽出と PLSA

デル化する手法として提案された[6]．これはクラスタリング手法の一つとも考えることができ，文書と単語に限らず，任意の複数変数の共起のクラスタリングに用いることができる．今回は係り受け関係にある名詞と動詞の共起のクラスタリングに用いるものとし，図 1(c) のように，格ごとに PLSA を定義し，EM アルゴリズムによりパラメータの推定を行う．

3.3　学習された PLSA パラメータを用いた置換手順

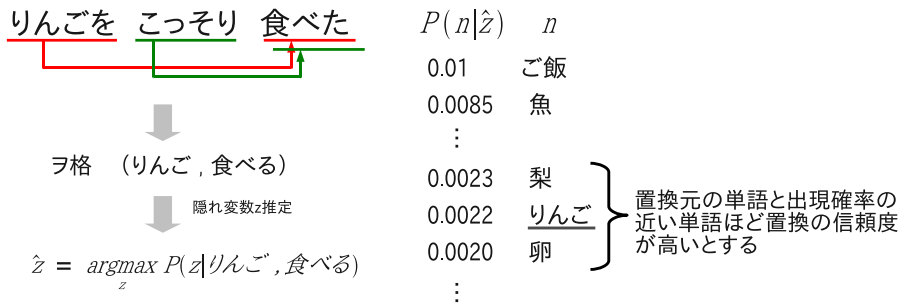
実際に置換を適用する手順を示す．まず**図 2**(a) のように類似語置換を行う小規模データ中の各文を係り受け解析し，格フレームの抽出時と同じ手順で名詞と動詞のペアを抽出しそこから隠れ変数 z を推定する．隠れ変数 z の推定は以下の式に従う．

$$\begin{aligned}\hat{z} &= \underset{z}{\operatorname{argmax}}\, P(z|n,v) \\ &= \underset{z}{\operatorname{argmax}}\, P(z)P(n|z)P(v|z)\end{aligned} \tag{1}$$

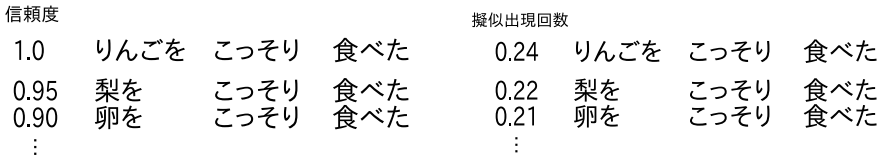
そして図 2(b) のように，推定した隠れ変数 $\hat{z}$ について，$P(n|\hat{z})$ の高い順に名詞 n を並べる．ここで，置換元の単語，ここでは「りんご」と出現確率の近い単語ほど，トピック $\hat{z}$ の下で置換の信頼度が高いという仮定を置く．予備実験により，$\hat{z}$ の下での置換元単語の出現確率を p，置換先単語の出現確率 q として，その類似度 $Sim(p,q)$ を以下のように設定した．

$$Sim(p,q) = \frac{1 - \frac{|p-q|}{p}}{1 + \frac{|p-q|}{p}} \tag{2}$$

これは，$Sim(p,p) = 1, Sim(p,0) = 0$，また $|p-q|$ に対し単調減少になるような関数となっている．ただし，$2p < q$ のとき $Sim(p,q) = 0$ とする．これにより，$0 \leq Sim(p,q) \leq 1$ となる．単に出現確率の高い単語ほど置換の信頼度が高いとした場合，トピックごとに常に



(a) 係り受け解析と隠れトピックの推定　(b) 置換単語の選択



(c) 置換により生成された文　(d) 生成された文とその擬似出現回数

図 2　置換の適用

決まった数単語が置換されることになり，出現率の低い単語は置換されない．そもそも出現確率の高い単語は出現確率の低い単語に比べればデータ不足の問題が少なく，出現確率の高い単語にのみ置換される場合，効果が低くなることが予想される．これらを考慮して，上記のように $Sim(p,q)$ を定義した．

さらにこれを用いて $n, v, \hat{z}$ の下での名詞 n から名詞 m への置換の信頼度 $R(n,m,\hat{z})$ を

$$R(n,m,\hat{z},v) = P(\hat{z}|n,v) \cdot Sim(P(n|\hat{z}), P(m|\hat{z})) \tag{3}$$

と定義した．$P(\hat{z}|n,v)$ を乗じた理由は，$P(\hat{z}|n,v)$ が大きいほどトピック推定の信頼度が高いと考えられるからである．以上の方法を用いて，置換により生成された文とその信頼度を並べたものが図 2(c) である．そしてその上位 N 文を採り，信頼度を合計 1 になるように正規化し，図 2(d) のように擬似的な文の出現回数とした．

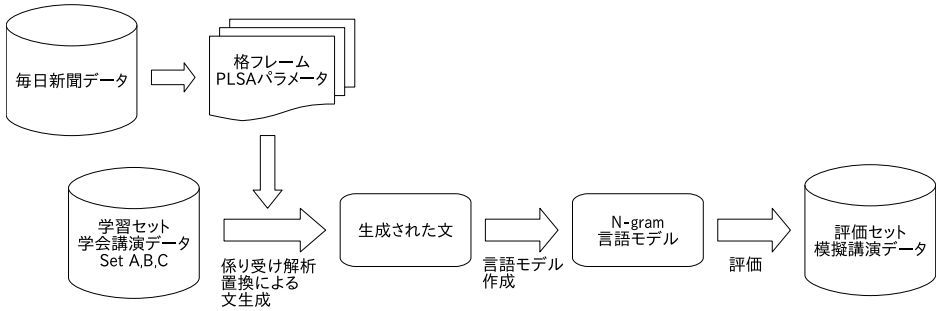


図 3 実験の概要

表 1 実験環境

言語モデルの種類	単語 3-gram
	with backoff smoothing
ディスカウント手法	Witten-Bell 法
語彙数	50000
格フレーム抽出用データ	毎日新聞'91-'95,'97-'98
	約 6000 万単語
N-gram 学習用データ	set A : CSJ 工学系学会講演　約 200 万単語
	set B : CSJ 人文系学会講演　約 90 万単語
	set C : CSJ 社会系学会講演　約 60 万単語
評価データ	CSJ 模擬講演　約 400 万単語
潜在トピック数	100
評価方法	test set perplexity

4. 評価実験

4.1 実験条件

実験の概要を**図 3** に，実験条件を**表 1** に示す．格フレーム抽出には毎日新聞データベースを用いる．また，評価データは CSJ 模擬講演であり，これと発話スタイルは類似しているが主題の異なるデータとして CSJ 学会講演 (工学系, 人文系, 社会系，以下 set A,B,C) を用いる．また置換対象として，名詞のみを置換した場合，動詞のみを置換した場合，名詞と動詞の両方を置換した場合について比較を行う．本実験では，通常の言語モデルの適応とは異なり，タスクに対しスタイル，主題ともに適合したデータは用いずに，発話スタイルの適合したデータを用いてその語彙的頑健性の向上を狙う．

これに加え，比較用として毎日新聞データ及び評価データである模擬講演データを用いた言語モデルも作成した．ただし模擬講演データを用いる際には，データの 1/10 を評価データ，残りを学習データとして用いる (タスク closed，データ open)．

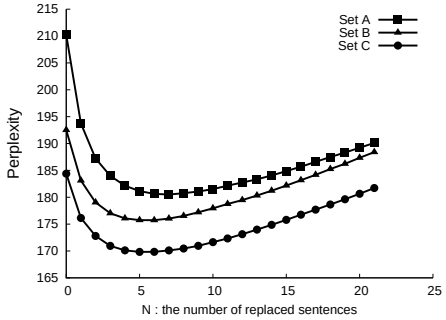
置換による文生成時には，3.3 節に示した通り，set A から C の各文ごとに信頼度上位 N 文を生成文として追加する．まず，N を変化させたときの性能の変化を確認する．そして得られた結果により最適な N を決定後，再度学習セットごとの提案手法の効果を確認する．評価尺度としては Perplexity(以下 PP) を用いるが，同時に 3-gram hit rate の変化についても確認する．

なお，形態素解析器として ChaSen[8]，係り受け解析器として CaboCha[9] を用いた．

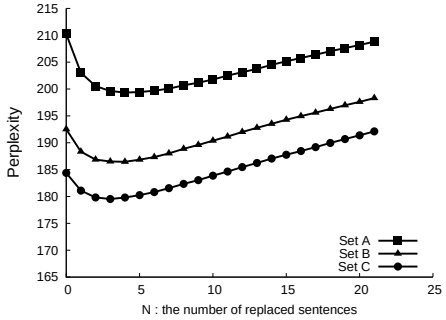
4.2 実験結果

まず，生成文数 N による PP の変化を**図 4** に，3-gram hit rate の変化を**図 5** に示す．どの学習セット，どの置換対象を用いた場合にも，N が増えるに従って 3-gram hit rate は上昇している．一方，PP は $N = 5$ 付近で最小となり，その後上昇している．これにより，$N = 5$ として再度学習セット間の比較を行う．その結果を**図 6** に示す．図 6 の Newspaper, Closed は，それぞれ毎日新聞データ，模擬講演データを学習データとして用いたものである．毎日新聞データは書き言葉であって，評価データである模擬講演データと大きく語彙，文体などが異なるため，PP は非常に大きい．逆に模擬講演データは当然のことながら評価セットに非常に近く，PP は非常に小さくなっている．この模擬講演データによる結果が今回の最高ラインとなる．通常の，置換を用いない trigram と提案手法を比べてみると，set A から C のすべての学習データにおいて性能の向上が見られる．置換の対象について比較すると，どの学習データを用いた場合にも動詞のみを置換した場合，名詞のみを置換した場合，両方を置換した場合の順に性能が向上しており，両方を置換した場合にはそれぞれ 14.8%,9.5%,8.5%の PP 削減が得られている．これにより，提案手法の有効性が確認された．特に set A（工学系学会講演）において最も効果が大きいが，これは評価セットの模擬講演に対し，工学系学会講演が最も主題のミスマッチが大きいためであると考えられる．

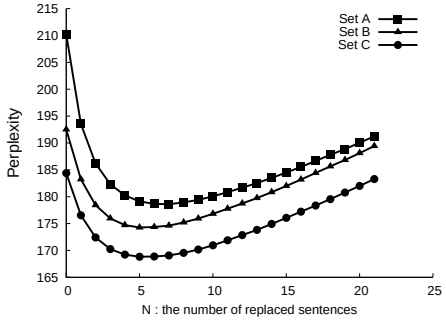
また，動詞の置換と名詞の置換について比べてみると，名詞の置換の方が効果が高いことが分かる．これについては，(1) 語彙数の違い，(2) 動詞の音便などに対応できていないことによる日本語として不適切な文の生成，(3) 代名詞と係り受け関係にある場合の問題，などが考えられる．(1) について，まず通常使われる名詞の種類は動詞の種類に比べ遥かに多いことが知られている．実際，今回の実験においては毎日新聞 10 年分から出現頻度上位 50000 単語を語彙として採用しているが，そのうち名詞が 38324 単語と約 77%を占めるの



(a) 名詞のみを置換

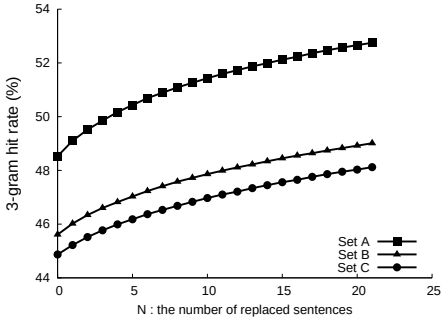


(b) 動詞のみを置換

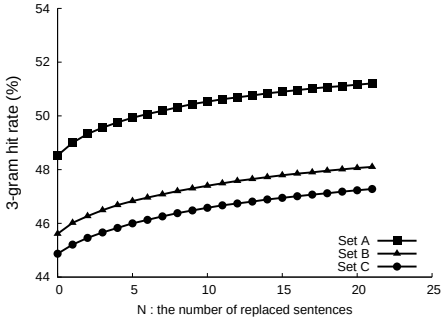


(c) 名詞と動詞の両方を置換

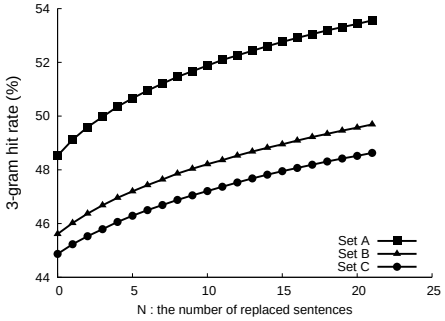
図 4 生成文数 N による perplexity の変化



(a) 名詞のみを置換



(b) 動詞のみを置換



(c) 名詞と動詞の両方を置換

図 5 生成文数 N による 3-gram hit rate の変化

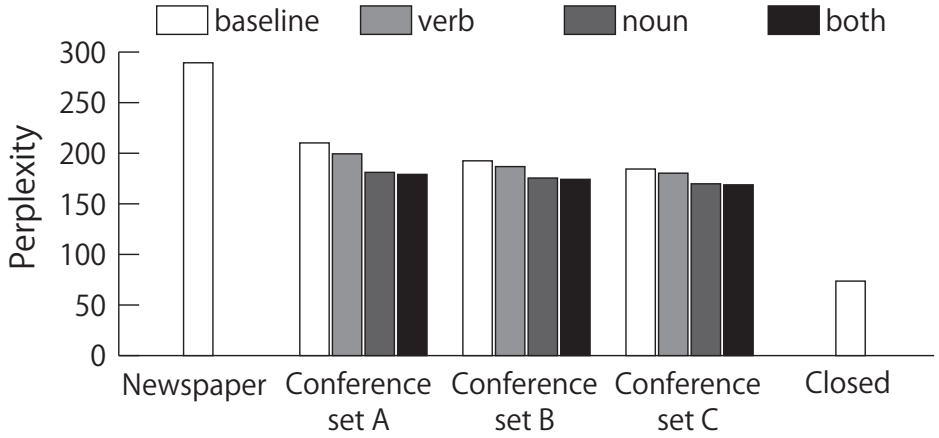


図 6 学習セットごとの提案手法による変化

に対し，動詞は 8881 単語と名詞の 4 分の 1 以下である．そのため，名詞の方が動詞に比べて言語モデル全体に及ぼす影響が大きく，また名詞の方がよりデータ不足の問題が大きく，本手法がより有効に働いたと考えられる．また (2) については，音便や濁音化など，動詞には名詞には無い変化があり，今回の実装ではそれに対応できていないことにより，不適切な語系列が生成されたことが言語モデルの性能を悪化させる方向に働いたと考えられる．さらに (3) について，話し言葉においては，代名詞が非常に多く現れる．それゆえ，動詞と係り受け関係にある名詞も代名詞であることが多い．今回の枠組みでは，文脈として係り受け関係にある単語のみを考慮しているため，ある動詞を置換する際，その動詞と係り受け関係にある名詞が代名詞である場合に文脈の情報をあまり利用できていないことになる．これも，動詞の置換の効果が低い原因の一つであると考えられる．

また，個別の生成例について**表 2** に示す．左の列にある数字は置換の信頼度である．まず名詞の置換について見ると，(a),(b) は共に「声」を置換するものだが，格と係り先が違うことにより，置換する単語が異なっているのが分かる．さらに (c) は「人」を置換する例であるが，いずれの表にも共通して,「さ」「ら」「さん」など，そもそも置換先の名詞の候補として妥当でないものが含まれているのが分かる．これらの原因の一つは，形態素解析の誤りによるものであると考えられる．話し言葉，特にフィラーや言い直しなどの不規則な表現が多い文においては，形態素解析の誤りが増えることが多い．また (b) の「後」などは，

表 2 生成例

(a) 名詞「声」ニ格 (声, 合わせる)

原文			
0.97	えー	**声**	に合わせていく
0.85	えー	**批判**	に合わせていく
0.85	えー	**要望**	に合わせていく
0.78	えー	**さ**	に合わせていく
0.77	えー	**要請**	に合わせていく
0.67	えー	**ニーズ**	に合わせていく

(b) 名詞「声」ヲ格 (声, 聞く)

原文			
0.96	えー	**声**	を聞いた時の
0.43	えー	**言葉**	を聞いた時の
0.26	えー	**後**	を聞いた時の
0.23	えー	**感想**	を聞いた時の
0.23	えー	**音**	を聞いた時の
0.18	えー	**情報**	を聞いた時の

(c) 名詞「人」ヲ格 (人, 調べる)

原文			
0.27	どういう	**人**	を調べるかと言うと
0.27	どういう	**者**	を調べるかと言うと
0.05	どういう	**さん**	を調べるかと言うと
0.04	どういう	**ら**	を調べるかと言うと
0.02	どういう	**女性**	を調べるかと言うと
0.02	どういう	**員**	を調べるかと言うと

(d) 動詞「買う」ガ格 (自分, 買う)

原文			
0.33	自分が	**買う**	としたら
0.32	自分が	**読む**	としたら
0.32	自分が	**いう**	としたら
0.32	自分が	**住む**	としたら
0.31	自分が	**使う**	としたら
0.30	自分が	**答える**	としたら

(e) 動詞「買う」ヲ格 (車, 買う)

原文			
0.46	日本の車を	**買う**	と
0.16	日本の車を	**売る**	と
0.16	日本の車を	**使う**	と
0.12	日本の車を	**包む**	と
0.11	日本の車を	**扱う**	と
0.10	日本の車を	**売り出す**	と

(f) 動詞「用いる」ヲ格 (機械, 用いる)

原文			
0.77	えー機械を	**用い**	て
0.77	えー機械を	**持ち込み**	て
0.58	えー機械を	**積み**	て
0.48	えー機械を	**備え**	て
0.45	えー機械を	**含み**	て
0.41	えー機械を	**たたき**	て

文脈によっては置換先の名詞になり得るが，この文脈では妥当でない．

さらに，動詞について見ると，(d),(e) は共に「買う」を置換するものだたが，やはり格と係り先の違いにより，置換する単語が異なっている．さらに (f) は動詞の音便変化等に対応できていない例である．動詞「持ち込む」が助動詞「て」に接続する際には，単に連用形「持ち込み」に活用させるのでは不適切で，撥音便「持ち込ん」に変化させる必要がある．それだけでなく，助動詞「て」についても濁音化にした「で」に変化させるのが適切である．しかし，今回の実装ではこれらに対応できていない．

5. ま と め

言語モデル作成における学習データの不足を補うため，単語を置換することにより学習データを自動生成するという手法を提案した．その際係り受けに着目し，予め格フレームを抽出し名詞と動詞をクラスタリングしておくことにより，用法の変化を考慮した置換を試みた．さらに CSJ を用いた評価実験により，名詞と動詞を置換した場合に PP で平均 10.9%の改善を得，本手法の有効性を確認することができた．ただし，名詞のみ置換した場合にも PP で平均 10.2%の改善が得られており，両方を置換した場合における動詞の寄与分はわずかであった．

今後の課題としては，まず名詞のみを置換した時に比べ名詞と動詞の両方を置換した場合の改善がわずかである理由の調査が挙げられる．また，音便変化や濁音化に対応すること，生成文として追加する文数 N を文ごとに変化させることなどによる改善が見込まれる．さらに置換の指標についても，本稿で用いた格フレームと PLSA の組み合わせに限らず検討してみたい．

参 考 文 献

1) 河原達也：話し言葉音声認識の概観，電子情報通信学会技術研究報告. SP, 音声，Vol.100, No.523, pp.1–5 (2000).
2) Federico, M.: Bayesian estimation methods for n-gram language model adaptation, *Fourth International Conference on Spoken Language Processing* (1996).
3) 太田健吾，土屋雅稔，中川聖一：フィラーの書き起こしのないコーパスからのフィラー付き言語モデルの構築，情報処理学会研究報告. SLP， Vol.2007, No.75, pp.1–6 (2007).
4) 秋田祐哉，河原達也：統計的機械翻訳の枠組みに基づく言語モデルの話し言葉スタイルへの変換，電子情報通信学会技術研究報告. NLC， Vol.105, No.494, pp.19–24 (20051215).
5) 河原大輔，黒橋禎夫：格フレーム辞書の漸次的自動構築，自然言語処理， Vol.12, No.2, pp.109–132 (2005).
6) Hofmann, T.: Probabilistic latent semantic indexing, *In Proc.ACM SIGIR*, pp. 50–57 (1999).
7) 河原大輔，黒橋禎夫：高性能計算環境を用いた Web からの大規模格フレーム構築，情報処理学会 自然言語処理研究会，pp.67–73 (2006).
8) 松本裕治他：形態素解析システム『茶筌』，情報処理， Vol.41, No.11, pp.1208–1214 (2000).
9) 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).