類似単語の置換に基づく学習データ自動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, 齋藤大輔, 鈴木雅之, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

連続音声認識においては、タスクに応じて発話スタイルや主題などに適合した言語モデルを利用することが望ましい。そのためには、これらの要因に適合したコーパスを大規模に用意し言語モデルを構築できれば理想的であるが、現実には非常に困難である。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ことや、未知表現のためのディスカウントなどの手法が検討されている。

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 [1, 2]。本稿では、タスク適合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は、(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。以上の提案手法を、日本語話し言葉コーパス (以下 CSJ) を用いた実験により評価を行い、本手法の有効性を示す。

2 文の自動生成によるデータ量の増加

タスクに適合した言語データを擬似的に生成することによりデータ量の不足を補うという手法として、様々なものが提案されている。例えば太田らは、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構築を提案している [1]。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。また、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構築するという手法を提案している [2]。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。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。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のコーパスへの変換を試みたものである。

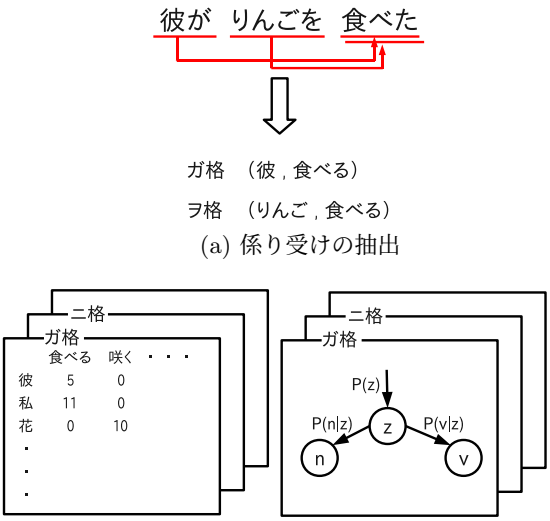
それに対し本稿で提案する手法は、対象とするタスクに対し発話スタイルが適合している言語データ中の語を類似単語と置換することにより文を生成し、学習データを増加させるという手法である。それによりスタイルを保持したままデータの量を増やし、主題の偏りを緩和することが可能となる。スタイルと一口に言っても、大きくは話し言葉か否か、細かくは講演か対話かなど様々に分かれる。その意味で、よりスタイルの適合した言語データは少量しか得られず、主題も限られてしまうことがある。そういった場合に本手法が有効となる。また、本手法による文生成アルゴリズムはスタイルの変換を学習するのではなく、単に単語の置換を学習するだけである。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 類似単語の置換による文生成

類似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。しかし、単語は文脈によって意味が変わる。ある二単語がある文脈ではほぼ同義で置換可能だったとしても、別の文脈では置換するのが適当でないということが起こる。そのため、文脈に依存して置換可能な単語が変化するような指標を用いることが望ましい。ここで言う文脈は本来、前後数単語から主題、さらには話者や周囲の状況など非言語情報をも含めたものであるはずだが、それらを全て考慮するのは不可能である。そこで、何をもって文脈とするかが問題となる。

そこで本手法では、名詞と動詞の係り受け関係を

*Training of robust language models by automatic sentence generation based on word replacement. by SHIMIZU Shinya, SAITO Daisuke, SUZUKI Masayuki, MINEMATSU Nobuaki and HIROSE Keikichi(The Univ. of Tokyo)



(a) 係り受けの抽出

(b) 名詞-動詞行列　(c) 格ごとの PLSA

Fig. 1　格フレームの抽出と PLSA

文脈として利用する。そのために、名詞と動詞の係り受けに関する情報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 (以下 PLSA) を用いて、係り受けを考慮した置換しやすさを学習する。

3.1　格フレーム

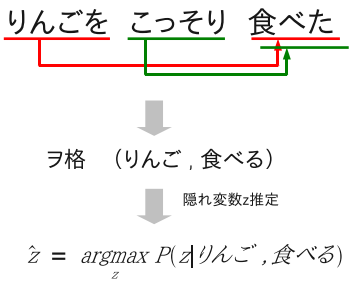
格フレームとは、用言と名詞の関係を、用言の用法ごとに記述したものである。本稿では表層格 9 格 (ガ格、ヲ格、ニ格、カラ格、ヘ格、ト格、ヨリ格、マデ格、デ格) のみを用いた簡易な格フレームを構築し利用した。さらに用言と名詞のペア抽出に関しても、(本来は係り受け解析だけでは不十分であるが、) 今回は単に係り受け解析を行い、係り受け関係にあるものを抽出した。近年、web を用いて大規模な格フレームが構築されており [3, 4]、これを用いることも可能である。

3.2　格フレーム抽出と PLSA

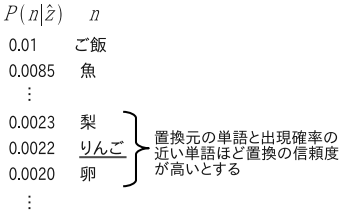
格フレーム抽出の手順を Fig. 1 に示す。まず、(a) のように係り受け解析を行い、係り受け関係にある二節が以下の条件を満たす時のみ、格フレームとして抽出した。

- 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

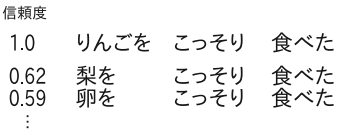
サ変接続名詞+「する」というのは、「心配する」「集合し」などである。これを擬似的に動詞として扱う。以上の条件を満たすとき、さらに以下の処理を行って格フレームとして加算するものとした。



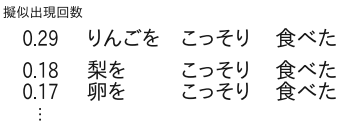
(a) 係り受け解析と隠れトピックの推定



(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

Fig. 2　置換の適用

- 助詞「は」をガ格として扱う
- 動詞の原形化

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、Fig. 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 とは、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デル化する手法であるが [5]、これは文書と単語に限らずとも適用可能なモデルである。今回は、名詞と動詞の出現確率をトピックという隠れ変数に基づいてモデル化する。Fig. 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 Fig. 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、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。

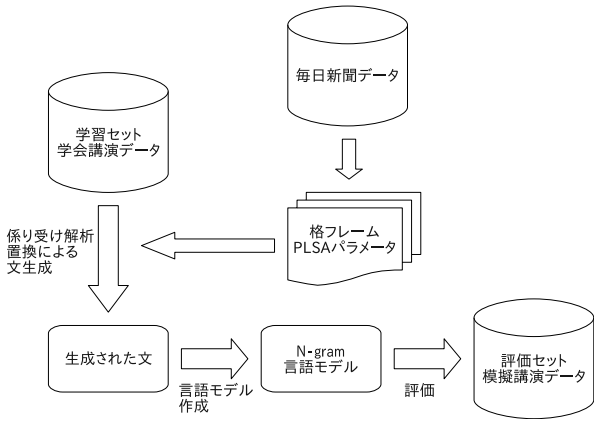


Fig. 3　実験の概要

$$\hat{z} = \underset{z}{\operatorname{argmax}}\, P(z|n,v) \quad (1)$$

$$= \underset{z}{\operatorname{argmax}}\, P(z)P(n|z)P(v|z) \quad (2)$$

そして Fig. 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仮定する。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。

$$Sim(p,q) = \frac{1 - \frac{|p-q|}{p}}{1 + \frac{|p-q|}{p}} \quad (3)$$

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しか置換されないのでは効果が薄いことが予想されることから、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

$$R(n,m,\hat{z},v) = P(\hat{z}|n,v) \cdot Sim(P(n|\hat{z}), P(m|\hat{z})) \quad (4)$$

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。そうして、置換により生成された文とその信頼度を並べたものが Fig. 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、Fig. 2(d) のように擬似的な文の出現回数とした。また、同様の手法で動詞 v の置換も可能であるが、今回は名詞のみとした。

Table 1　実験環境

言語モデルの種類	単語 trigram 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	testset perplexity

4　評価実験

4.1　実験条件

実験の概要を Fig. 3 に、実験条件を Table 1 に示す。日本語の知識としての格フレーム抽出には毎日新聞を用いる。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) を用いる。本実験では、評価データそのものは一切用いずに、発話スタイルの合致した言語モデル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 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。そうして得られた結果により最適な N を決定後、学習セットごとに提案手法の効果を確認することとする。

なお、形態素解析器として ChaSen[6]、係り受け解析器として CaboCha[7] を用いた。

4.2　実験結果

まず、生成文数 N による変化を Fig. 4 に示す。どの学習セットを用いた場合にも perplexity(以下 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、$N = 5$ として学習セット間の比較を行う。その結果を Fig. 5 に示す。Fig. 5 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

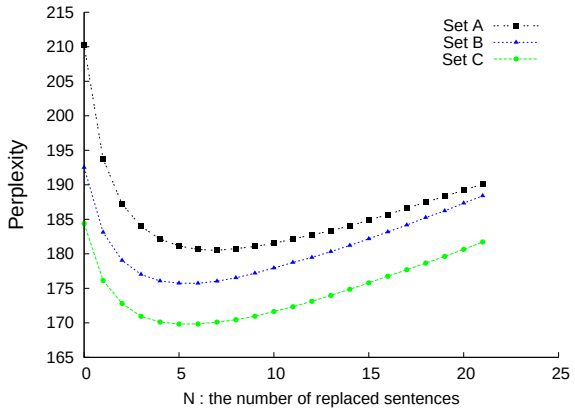


Fig. 4　生成文数 N による perplexity の変化

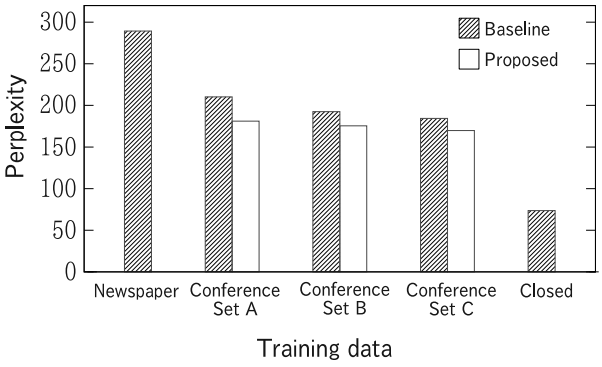


Fig. 5　学習セットごとの提案手法による変化

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 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、それぞれ、13.8%,8.8%,7.9%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。なお、特に set A の工学系学会講演で最も効果が大きいが、これは工学系学会講演が最も評価セットである模擬講演とドメインが離れていることに起因すると考えられる。

また、個別の生成例について Table 2 に示す。左の列にある数字は置換の信頼度である。(a),(b) は共に名詞「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) はまた別の例であるが、いずれの表にも共通して、「さ」「ら」「さん」など、そもそも置換先単語として妥当でないものが含まれているのが分かる。また、(b) の「後」や (c) の「員」など、妥当でないと思われるものが含まれている。これらの原因の一部は、形態素解析時の誤りによるものである。

Table 2　生成例

(a) ニ格 (声, 合わせる)

原文	えー **声**	に合わせていくというようなことが
0.97	えー **批判**	に合わせていくというようなことが
0.85	えー **要望**	に合わせていくというようなことが
0.78	えー **さ**	に合わせていくというようなことが
0.77	えー **要請**	に合わせていくというようなことが
0.67	えー **ニーズ**	に合わせていくというようなことが

(b) ヲ格 (声, 聞く)

原文	えー **声**	を聞いた時の印象を
0.96	えー **言葉**	を聞いた時の印象を
0.96	えー **後**	を聞いた時の印象を
0.96	えー **感想**	を聞いた時の印象を
0.96	えー **音**	を聞いた時の印象を
0.96	えー **情報**	を聞いた時の印象を

(c) ヲ格 (人, 調べる)

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

5　まとめ

言語モデルの作成において、タスクに適合したデータの量が少ない場合に、大規模データから格フレームという形で日本語に対する知識を獲得し、それを用いて学習データ中の語を置換して文を生成するという手法を提案した。さらに CSJ を用いた評価実験を行い、PP で平均 10.2% の改善を得た。

今後の課題としては、置換の信頼度の定義や隠れトピック数などを固定せず最適なものを求めること、より大規模な格フレームを用いること、名詞だけでなく動詞も置換することなどを検討する予定である。さらに、そもそも置換先として妥当ではないような単語をいかに排除するのかについても検討してみたい。

参考文献

[1] 太田 他, 情報処理学会研究報告, SLP2007-75, pp.1-6 (2007)
[2] Y.Akita and T.Kawahara, In Proc. ICASSP, Vol.4, pp33-36 (2007)
[3] 河原大輔, 黒橋禎夫, 自然言語処理, Vol. 12, No. 2, pp. 109-132 (2005)
[4] 河原大輔, 黒橋禎夫, 情報処理学会, 自然言語処理研究会, pp. 67-73 (2006)
[5] T.Hofmann, In Proc. ACM SIGIR, pp 50-57 (1999)
[6] 松本 他, 情報処理学会論文誌, Vol. 41, No. 11, pp. 1208-1214 (2000)
[7] 工藤拓, 松本裕治, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834-1842 (2000)