品詞並び検索条件の段階的緩和による用例ベース未知語品詞推定

Example-based Inference of Unknown Word Category by Stepwise Query Relaxation of Part-of-Speech Sequence

福岡知隆 [1*]　服部峻 [2]　久保村千明 [3]　亀田弘之 [2]
Tomotaka Fukuoka [1], Shun Hattori [2], Chiaki Kubomura [3], and Hiroyuki Kameda [2]

[1] 東京工科大学大学院 バイオ・情報メディア研究科
[1] Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology
[2] 東京工科大学 コンピュータサイエンス学部
[2] School of Computer Science, Tokyo University of Technology
[3] 山野美容芸術短期大学 美容保健学科
[3] School of Beauty and Health Science, Yamano College of Aesthetics

Abstract: In this paper, we evaluate the accuracy of three kinds of our example-based inference of unknown word category with stepwise query relaxation in example search. To search an example database for examples similar to an input sentence with an unknown word, we use the part-of-speech (POS) sequence of the input sentence. Our query relaxation is to stepwisely short the POS sequence by several strategies. A word category is probabilistically adopted based on its proportion in the top n similar examples ranked by their matching rates. A matching rate is calculated by the ratio of surface-based matching between an input sentence and a similar example. The number n groups our inference of unknown word category into three phases. The experimental results show that the accuracy of our inference increases more than three-fold by stepwise query relaxation of POS sequence.

1　はじめに

近年の情報通信技術の進歩により，人間の対話相手が増えた．人間に比べて膨大な情報の保持が可能なコンピュータである．チャットなどでの雑談相手，Web 上での商品の説明，介護における話し相手など，多岐にわたり人間はコンピュータと対話を行うようになった．

しかし，人間同士の対話と比較すると，コンピュータの返答結果や対話の過程は劣っている場合が多い．その原因の一つが円滑性（発話者の意図に沿い，対話が速やかに行われること）の欠如である．コンピュータのデータベース内に情報が存在しない単語，即ち未知語に遭遇した場合にその現象は著しい．既存の処理では未知語に対して人間への質問や話題転換が頻繁に起ってしまい，対話の円滑さが損なわれる場合がある．

この問題を解決するため，未知語処理を改善し人間とコンピュータ間の対話をより自然で円滑にする必要がある．一つの解決手法として，システムが自動的に未知語の情報を推定することで，既知語だけの発話と同様に応答することが可能になると考えられる．以下に述べる手法では，人間の発話における未知語の検出が成功していることを前提としている．

システムが未知語の品詞や意味などの情報を推定する手法の多くは，未知語に直接関連する情報，例えば，その未知語を含む文や文書などを辞書や新聞コーパスから獲得して，未知語の情報を推定している．

このような手法では Web などの膨大な量の情報源を利用して行われるため，システムが未知語の用例を獲得し、情報を推定できる可能性は大きい．しかし，未知語が新しい造語や，ごく一部の人間の間でしか使われない用語の場合，未知語の用例が情報源に存在しないなどうまく対応ができない可能性がある．また，Web を情報源とする場合は情報の信頼性にも問題がある．

そこで我々は，対象の未知語を含まない情報を用いて，未知語の情報を推定する手法を研究している．未知語の用例などの直接関係する情報を用いずに，入力文との類似性を品詞並びパターンや文中における単語間の共起パターンなどに基づいて評価し，類似検索した結果の類似用例を元に推定することで，例えすべての人間が知らない単語であってもその情報の推定を可

*連絡先：東京工科大学大学院バイオ・情報メディア研究科
〒 192–0982 東京都八王子市片倉町 1404–1
E-mail: g2110045e1@gss.teu.ac.jp

能とする．また，情報源としては，不特定多数者が作成する Web などの情報ではなく，システム管理者が精査した対話の用例データなどを用いる．

このような考えに基づき，未知語を含んだ入力文と品詞並びが類似した用例を用いて未知語の品詞推定処理を行うシステムを作成した [1]．このシステムではまず，システムの保持するデータベースの中から推定に必要な用例だけを抽出し，その上で抽出した用例群を用いて品詞推定を行う．類似した用例の抽出はシステムのパターンデータベースの中から入力文の品詞並びを用いたクエリによる検索により行われる．しかし，入力文の品詞並びをそのままクエリにしたのでは検索条件が厳し過ぎるため，検索される類似用例数が少なく，その結果，未知語品詞の誤推定や，推定不可能となる場合があるという問題があった．

本稿で提案するシステムは，この未知語推定における用例検索において，クエリである入力文の品詞並びに対して入力文の単語を削除することで検索条件の緩和を行い，検索される類似用例数を増加させる．複数の条件緩和手法を提案し，それらの組み合わせ，または手法の条件緩和の程度を変化させることで，段階的な条件緩和を行う．そして，それぞれの緩和段階における検索結果を評価することで，最適な条件緩和手法を獲得する．また，品詞推定手法においても，三つの手法を用いて推定精度の評価を行う．それぞれの条件緩和手法により得られた類似用例群を用いて表層的に最も類似した用例に着目した手法と抽出結果から確率的に品詞の推定を行う手法，またこれらを組み合わせた手法による品詞推定結果の推定精度の比較を行う．

2　提案手法

本章では入力文中の未知語の品詞推定を行うシステムの概要と，提案する手法について述べる．

2.1　システム概要

今回提案する未知語の品詞推定手法は，未知語の意味推定における要素の一つとなる．未知語の品詞を推定することにより，システムが未知語の意味推定を行うときに利用する類似用例の絞り込みが可能となる．

本稿でシステムが行う未知語処理は類似した用例の検索と品詞の推定である．入力文処理から得られた情報を元に，入力文と類似した用例を検索し，品詞の推定を行う（図 2.1）．入力文処理において MeCab[2] を用いた形態素解析を行うが，入力文中の未知語の検出は MeCab に依存しており，今回提案する手法は未知語が正しく検出される前提の下で行われている．

類似用例検索手法における条件緩和手法について **2.2** で述べる．**2.3** で検索された類似用例を用いた三つの品詞推定手法を定義する．**3** で類似用例検索数と品詞推定精度をそれぞれの手法を組み合わせて比較する．

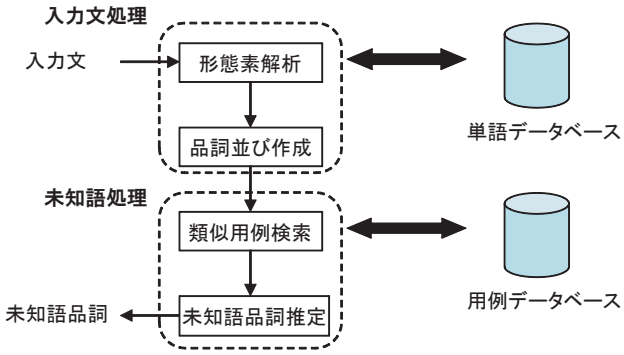

図 2.1: システムの処理の流れ

2.1.1　単語データベース

本システムでは形態素解析ツール MeCab で使用されている単語データベース ipadic2.7.0 を使用する．品詞は 13 種類（表 2.1），単語数は 231900 個である．

表 2.1: MeCab における品詞の分類

名詞	動詞	助詞	助動詞
形容詞	副詞	接続詞	感動詞
接頭詞	連体詞	記号	フィラー
その他			

2.1.2　用例データベース

Web 上に公開されている音声対話コーパス [3] を使用し，用例データベースを作成する．用例データベースの関係スキーマは次の三つである．

用例データベース　（用例，品詞並び，出現頻度）

用例は，一文から成り立つ会話文の連結例である．品詞並びは，MeCab を用いて用例の形態素解析をした結果，用例中の単語それぞれに表 2.1 の品詞を割り当て，単語の並び順にしたものである．頻度は，それぞれの用例が過去に何度出現したかを表すものであり，今回は用例データベースの作成に使用したコーパス中に何度出現したかを表す．

用例データベースに用例を保存する際に単語データベースに存在しない単語，即ち未知語を文中に含む用例は除外した．また，保存する用例は一文から成り立つものとし，複数の文から成り立つ用例は文単位に分割してから保存した．その結果，コーパスから抽出された用例は 2279 個であり，重複を除外すると 1279 行として用例データベースに保存された．用例データベースにおける頻度ごとの用例数を図 2.2 に示す．

2.2　類似用例検索条件の段階的緩和手法

品詞並びを用いて入力文と類似した用例を検索するが，クエリが入力文の品詞並びのままでは検索条件が厳し過ぎて，十分な数の類似した用例が得られない場

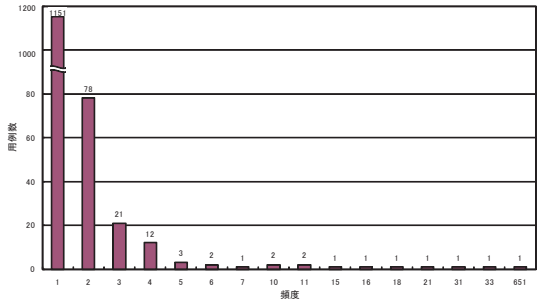

図 2.2: 用例数の頻度分布

合がある．以下の四つの手法を用いてクエリに特定の処理を施し，検索条件を段階的に緩和する．また，一つの用例の中に複数の入力文と類似した品詞並びが存在する場合，それぞれ別の用例として扱うものとしている．そのため，データベースに保存されている用例数よりも検索された用例が多い場合がある．

(1) 句点の削除 入力文の句点を削除したクエリを用いる．句点が存在した場合，品詞並びが一致するのは文末だけとなるが，句点を削除することで，文の途中に存在する品詞並びの検索が可能となる．

(2) 読点を基点とした削除 文中に読点が存在した場合，読点を基点として用例を分割する．分割された用例それぞれの内，未知語を含む部分の品詞並びをクエリとする（図 2.3）．

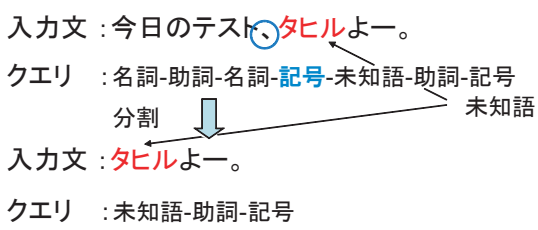

図 2.3: 読点を基点とした削除による緩和

(3) 組み合わせによる緩和 手法（1）と（2）を組み合わせた結果をクエリとする．

(4) 文の先頭・文末から単語を削除 入力文の単語を文の端から段階的に削除した結果をクエリとする．削除する単語数 k はパラメータで任意だが，クエリは最低でも三つの単語の品詞並びとなるものとする．また，未知語の位置により場合分けされる．文の先頭が未知語であった場合は，文末から単語を削除する．文末が未知語であった場合は，先頭から削除する．それ以外の場合は未知語の前後の単語が残るように，文の先頭と文末の交互に単語を削除していく（図 2.4）．

2.3 品詞推定手法

システムは類似用例検索結果を元に品詞推定を行う．本稿では品詞推定手法として以下の三種類の手法を定義し，次章においてその品詞推定精度の比較を行う．

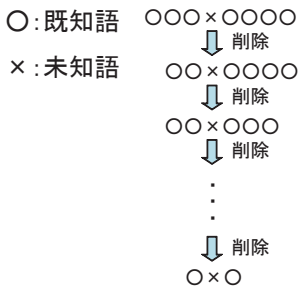

図 2.4: 文の先頭・文末からの単語削除による緩和

(a) 用例の表層文字列の一致率による推定 類似した用例群のそれぞれに対して，入力文の単語群の表層文字列との一致率を計算する．最もその値が大きい用例を用いて，入力文の未知語と同じ位置にある単語の品詞を未知語の品詞とする．一致率 w の計算には，X を入力文に出現する単語の集合，Y を候補文に出現する単語の集合とするダイス係数を用いる．ただし，一つの文中に同一の単語が複数回出現した場合は別の要素として扱う．

$$w = 2 \times \frac{|X \cap Y|}{|X| + |Y|}$$

以下に具体例を示す．

入力文の単語

X:音声/対話/という/こと/は/喋れ/ます/か/?

類似した用例の単語

Y_1:音声/対話/という/こと/が/重要/です/か/?
Y_2:音声/対話/という/こと/は/話せ/ます/か/?

計算結果

$w_1 : 2 \times \frac{6}{9+9} \fallingdotseq 0.67$
$w_2 : 2 \times \frac{8}{9+9} \fallingdotseq 0.89$ （Y_2 が最類似用例と判断）

また，一致率が同値であった場合は品詞並びを抽出した用例の頻度が大きい方が優先される．

(b) 類似した用例の比率による確率的推定 システムは類似した用例を検索した段階で，それぞれの用例を用いた場合に推定される品詞を獲得する．その結果，推定される品詞の比率が求められる．この手法では，その品詞の比率から確率的に品詞を推定する．図 2.5 の例では，システムは未知語の品詞を 65 %の確率で動詞と推定する．

(c) 一致率の上位 n 件の用例による確率的推定 検索された類似した用例の内，一致率が大きい順に用例を n 個選択する．選択された用例の中での推定品詞の比率を用いて確率的に品詞の推定を行う．一致率が同値の用例は品詞並びを抽出した用例の頻度が大きい方を上位とする．手法（a）と（b）

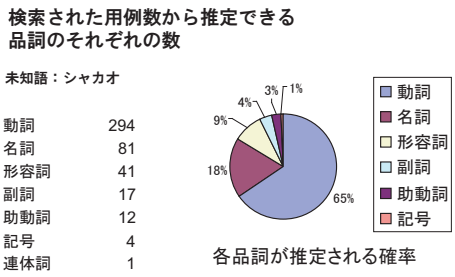

図 2.5: 入力文から検索される用例の比率の例

はこの手法において，n を特定の数値とすることで実現可能である．n が 1 である場合は，一致率が最も大きい用例が選択され，その用例の品詞が選択される確率は 100 %であり，手法（a）に相当する．n が検索された用例数と同じならば，すべての用例を用いた品詞の比率を用いて確率的に推定され，手法（b）に相当する．

3 実験

未知語を含んだ入力文を [4] を参考に 28 個用意し（表 3.1），それぞれの条件緩和による類似用例数の変化と品詞推定手法ごとの推定精度を比較する．ただし，未知語の品詞は著者が主観的に決定したものである．

3.1 類似した用例の検索結果比較

入力文が品詞並びそのままのクエリ，句点の削除を行ったクエリ，読点による分割を行ったクエリ，二つの削除手法を組み合わせによる条件緩和を行ったクエリによって類似した用例を検索した結果を図 3.1 と図 3.2 に示す．図 3.1 は 28 個の入力文すべての検索結果であり，図 3.2 はその一部を拡大したものである．

文の先頭・文末から削除する単語数ごとの平均用例検索数と単語の削除が有効な入力文の数を図 3.3 に示す．

これらの結果から，すべての検索条件緩和手法で類似用例検索数の向上が見られた．しかし，28 個の入力文に対して類似した用例を検索できたのは，手法（4）において単語の削除数が 8 以上の場合であり，それ以外の手法は十分な用例検索数を獲得できたとはいえない．

3.2 品詞推定手法ごとの推定精度の比較

品詞推定手法において，入力文そのままと，条件緩和手法（1），(2），(3）による品詞推定数，その内の正答数，推定精度を図 3.4 に示す．n が 1 の場合は推定手法（a）の結果であり，n が max の場合は手法（b）の結果であり，それ以外は手法（c）の結果の一例である．また，$n = 1$ の場合には正答数は品詞推定が正しく行われた入力文の数であるが，それ以外の場合は，正答数はそれぞれの入力文において，正しい推定が行われる確率を有効数字 2 桁で表した数字の和である．

条件緩和手法（4）による品詞推定結果を $n = 1$ の場合（a）を図 3.5 に，n が最大の場合（b）を図 3.6 に，$n = 10$ の場合を図 3.7 に，$n = 100$ の場合を図 3.8 にそれぞれ示す．28 個すべての入力文のクエリを形態素 3 個まで削除し，n を 1 から 100 まで変化させた品詞推定精度の推移を図 3.9 に示す．

これらの結果から，今回評価を行った手法（b）による比率を元にした確率的な推定手法は精度の低下を招いていることが判明した．しかし，図 3.9 のグラフから n が 10 以下の結果が高い品詞推定精度を示す場合がある．即ち，確率的に品詞推定を行う場合，品詞推定精度の向上には厳選した用例だけを利用する必要がある．

図 3.5〜図 3.8 のグラフより，手法（4）の条件緩和手法では削除単語数 k を増やし，類似用例検索数を増せば，一部では品詞推定精度が若干低下することもあるが，全体的には品詞推定精度が向上している．類似用例検索における条件緩和には文の先頭・文末からできる限り単語を削除することが最も有効である．即ち，類似用例の検索においては未知語の周囲単語を用いるだけで良いという知見が得られた．

4 むすび

本稿では品詞並びをクエリとした類似用例検索における条件緩和手法の提案と，その結果を用いた三つの品詞推定手法の推定精度の評価を行った．その結果，本稿で提案したほとんどの条件緩和手法と品詞推定手法において，用例の検索数と品詞推定の精度の向上が見られた．特に条件緩和手法（4）は他の手法のおよそ 2 倍の推定可能数と推定精度を示した．しかし，他の手法の用例検索数は，条件緩和により増加しても依然として低い水準のままである．よって，今回提案した条件緩和手法の中で有用なのは手法（4）だけである．

品詞推定手法は，確率的な手法において，使用する用例に一致率の高い一部だけを用いることで最も大きい品詞推定精度を得ることが可能となっている．

検索条件の緩和手法として（1），(2），(3）は不十分であるといえるため，今後は，手法（4）による条件緩和を行うものとし，用例データベース作成に使用するコーパスや実験対象の入力文セットを変更したり，品詞を細分化した場合の推定精度の比較評価する．

参考文献

[1] 福岡知隆, 税田竜一, 久保村千明, 服部峻, 亀田弘之:文の類似性を用いた未知語処理手法の提案とそれに基づく円滑な対話応答システムの作成, 情報処理学会第 72 回全国大会, 6X-2, pp.2-619–620 (2010/3).

[2] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecabsourceforge.net/ (2010/9/1).

[3] 重点領域研究 音声対話 コーパス, http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/ (2010/5/24).

[4] 大修館, 第 3 回「もっと明鏡」大賞 みんなで作ろう国語辞典: http://www.taishukan.co.jp/meikyo_campaign3/happyo/can3_release.pdf (2010/7/7).

表 3.1: 実験に使用した未知語を含む入力文

番号	入力文 品詞並び	未知語 品詞
1	昨日はガチで怒られた。 名詞-助詞-未知語-助詞-動詞-動詞-助動詞-記号	ガチ 名詞
2	今朝中央線で、隣がシャカオでさ。 名詞-名詞-名詞-助詞-記号-名詞-助詞-未知語-助動詞-助詞-記号	シャカオ 名詞
3	今日のテスト、タヒルよー。 名詞-助詞-名詞-記号-未知語-助詞-記号	タヒル 動詞
4	山田から田中にチェンソーした。 名詞-助詞-名詞-助詞-未知語-動詞-助動詞-記号	チェンソー 名詞
5	チェンソーして欲しい。 未知語-動詞-助詞-形容詞-記号	チェンソー 名詞
6	また MONPE からクレームが来たよ。 接続詞-未知語-助詞-名詞-助詞-動詞-助動詞-助詞-記号	MONPE 名詞
7	ヤスレよー。 未知語-助詞-記号	ヤスレ 動詞
8	誰かヤスラなきゃー。 名詞-助詞-未知語-助動詞-感動詞-記号	ヤスラ 動詞
9	おい、コソアドしてたのかよ。 感動詞-記号-未知語-動詞-動詞-助動詞-名詞-助詞-助詞-記号	コソアド 名詞
10	最近の若者は DENKOKU のほうがうまくいくんですよ。 名詞-助詞-名詞-助詞-未知語-助詞-名詞-助詞-形容詞-動詞-名詞-助動詞-助詞-記号	DENKOKU 名詞
11	こっちからバカヤローカイサンしてやったよ。 名詞-助詞-未知語-動詞-助動詞-助詞-記号	バカヤローカイサン 名詞
12	まさにハレンチなのだろう。 副詞-未知語-助動詞-名詞-助動詞-助動詞-記号	ハレンチ 名詞
13	おっかけひと筋だった彼女がついに RIAKOI に目覚めた。 動詞-名詞-名詞-助動詞-助動詞-名詞-助詞-副詞-未知語-助詞-動詞-助動詞-記号	RIAKOI 名詞
14	あいつ今、お菓子の数をギソッたよな？ 名詞-名詞-記号-名詞-助詞-名詞-助詞-未知語-助動詞-助詞-助詞-記号	ギソッ 動詞
15	宿題が終わんなくってクマった。 名詞-助詞-動詞-形容詞-助詞-未知語-助動詞-記号	クマッ 動詞
16	テスト中に答えが分からなくなりセパった。 名詞-名詞-助詞-名詞-助詞-動詞-助動詞-動詞-未知語-助動詞-記号	セパッ 動詞
17	その瞬間、教室中がナギった。 連体詞-名詞-記号-名詞-名詞-助詞-未知語-助動詞-記号	ナギッ 動詞
18	どんな些細なことでもペリル性格。 連体詞-名詞-助動詞-名詞-助動詞-助詞-未知語-名詞-記号	ペリル 動詞
19	社長の息子だから生まれつきペリッている。 名詞-助詞-名詞-助動詞-助詞-名詞-未知語-助詞-動詞-記号	ペリッ 動詞
20	ついポチッちゃった。 副詞-未知語-動詞-助動詞-記号	ポチッ 動詞
21	数学の時間はずっとモソッてたんだよ。 名詞-助詞-名詞-助詞-副詞-未知語-助詞-名詞-助動詞-助詞-記号	モソッ 動詞
22	あいつ最近イキリコブタになっているよね。 名詞-名詞-未知語-助詞-動詞-助詞-動詞-助詞-助詞-記号	イキリコブタ 名詞
23	めっちゃオシムだったな〜。 副詞-未知語-助動詞-助動詞-助詞-記号-記号	オシム 名詞
24	ずっと外にいたから KANIKAMA や。 副詞-名詞-助詞-動詞-助動詞-助詞-未知語-助動詞-記号	KANIKAMA 名詞
25	あのシャメラマンの取った写真は絶品だ。 連体詞-未知語-助詞-動詞-助動詞-名詞-助詞-名詞-助動詞-記号	シャメラマン 名詞
26	うわ〜、今日マジちょ〜ダルビッシュ！ 感動詞-記号-記号-名詞-名詞-名詞-記号-未知語-記号	ダルビッシュ 名詞
27	HAMINGUAUTO する人って。 未知語-動詞-名詞-助詞-記号	HAMINGUAUTO 名詞
28	近頃はボウインボウショクがはげしくて困ります。 名詞-助詞-未知語-助詞-形容詞-助詞-動詞-助動詞-記号	ボウインボウショク 名詞

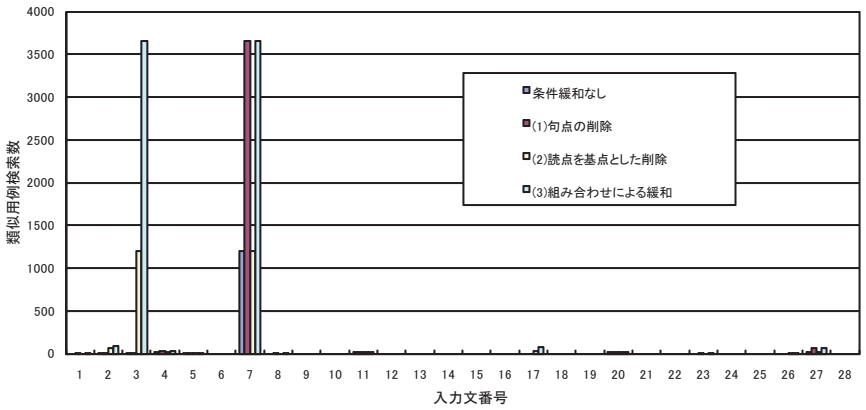

図 3.1: 類似用例検索の条件緩和による検索結果の比較（全体）

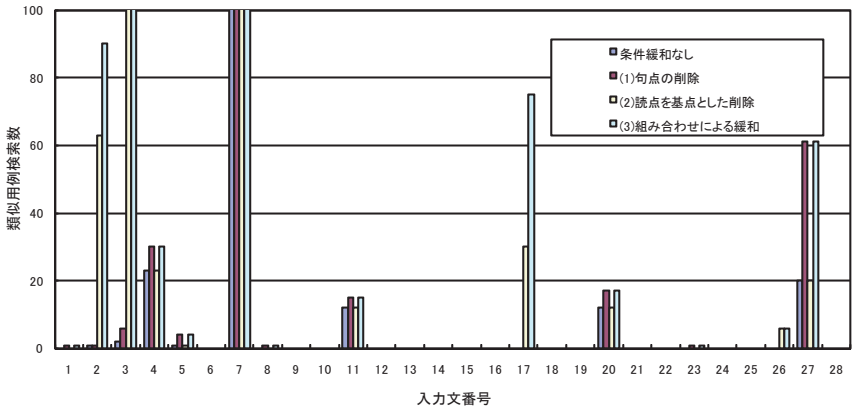

図 3.2: 類似用例検索の条件緩和による検索結果の比較（一部拡大）

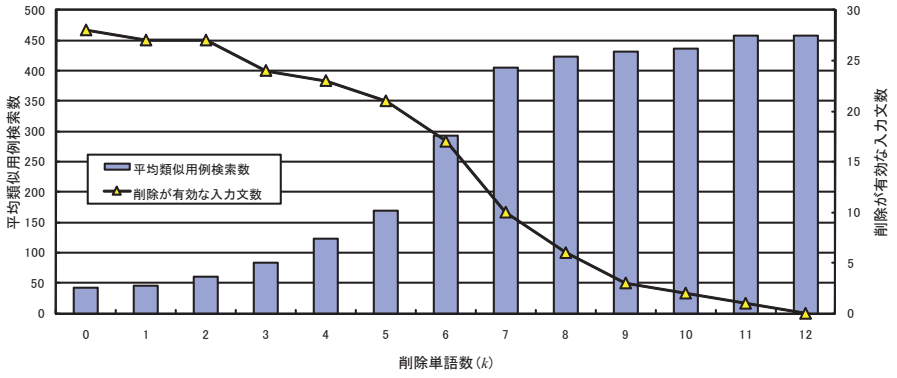

図 3.3: 文の先頭・文末から削除する単語数と検索された平均用例数

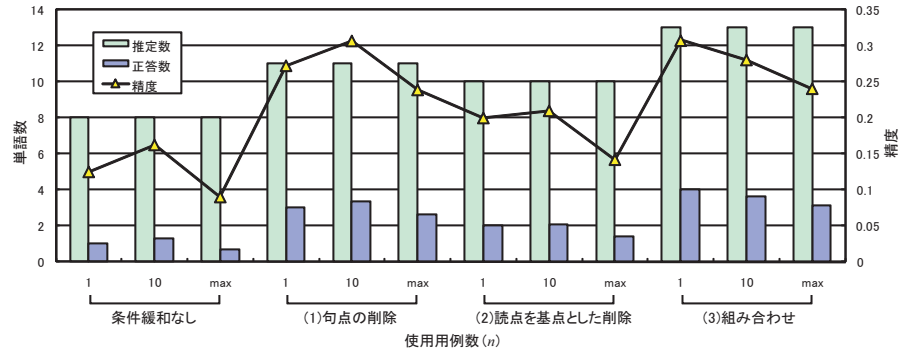

図 3.4: 品詞推定手法ごとの品詞推定精度の比較

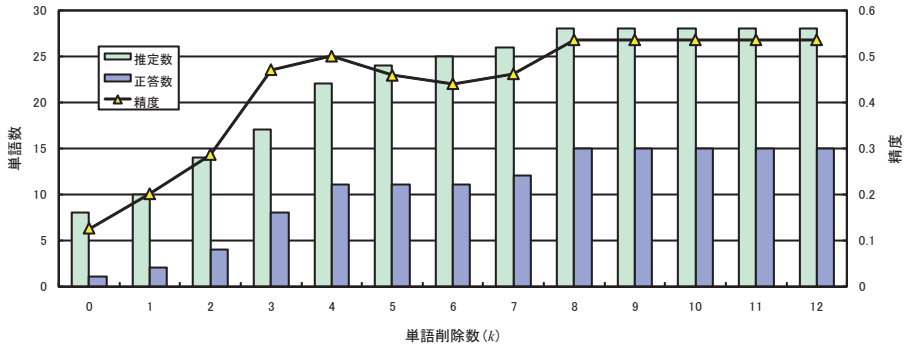

図 3.5: 条件緩和手法（4）における $n = 1$ の品詞推定結果

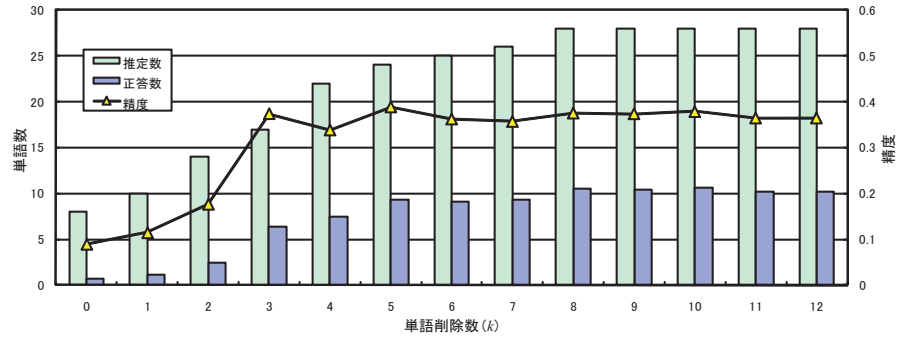

図 3.6: 条件緩和手法（4）における n が検索された用例数と同じ場合の品詞推定結果

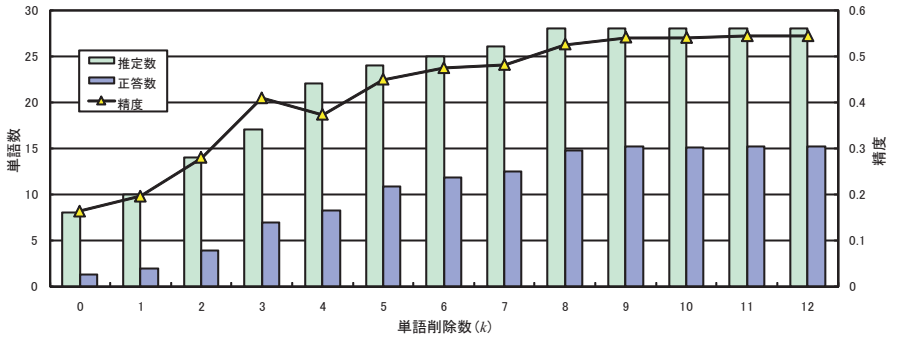

図 3.7: 条件緩和手法（4）における $n = 10$ の品詞推定結果

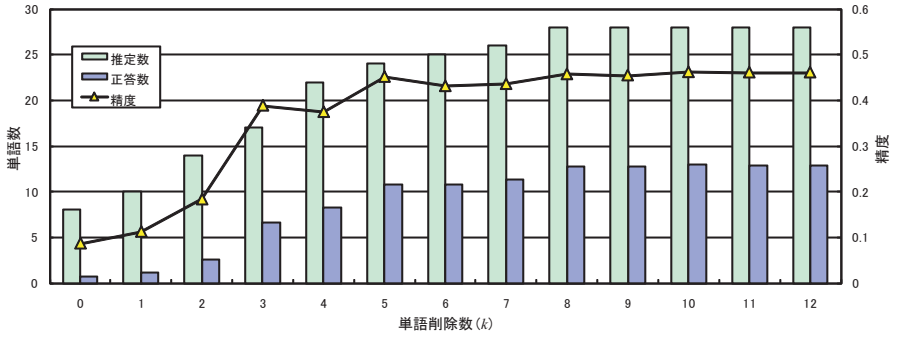

図 3.8: 条件緩和手法（4）における $n = 100$ の品詞推定結果

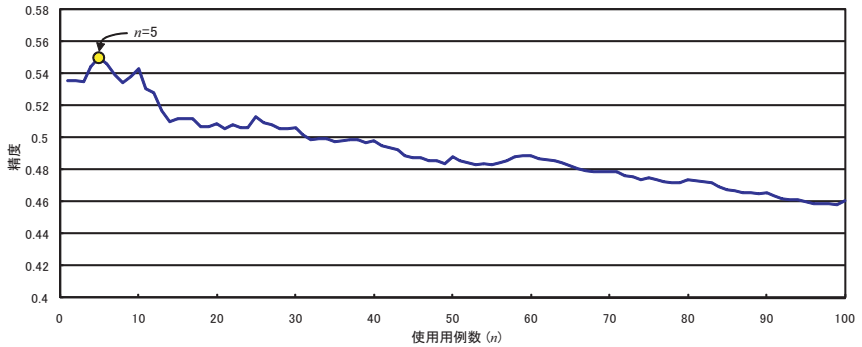

図 3.9: 条件緩和手法（4）における n を 1 から 100 まで変化させた品詞推定精度の推移