品詞列パターンの類似性に基づく未知語品詞推定における品詞分類体系と品詞列長が与える効果

服部　峻†　福岡　知隆††　久保村千明†††　亀田　弘之†

† 東京工科大学コンピュータサイエンス学部　〒 192–0982 東京都八王子市片倉町 1404–1
†† 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
††† 山野美容芸術短期大学美容総合学科　〒 192–0396 東京都八王子市鑓水 530
E-mail: †{hattori,kameda}@cs.teu.ac.jp, ††g2110045e1@gss.teu.ac.jp, †††ckubomura@yamano.ac.jp

あらまし　Web 文書やブログ文書，チャット（対話）ログなどをテキスト解析し，様々な知識を抽出または発見する研究が盛んに行われている．例えば，形態素解析器によって Web テキストを形態素に区切り，特定の品詞の単語のみに反応して情報抽出を行いたい場合がある．しかしながら，Web テキストに限らず，近年のテキスト文書中には既存の形態素解析器にとって未知である語の割合が非常に増えて来ており，未知語に対しては既存の形態素解析器を用いた品詞推定が必ずしも正確ではないため，有益な知識の元を逃してしまっている．この問題に対して我々は，既存の形態素解析器のラッパーとして，形態素解析器が有する辞書に未登録である未知語に対しても，その未知語を含んだ入力文の品詞列パターンを適度に条件強化・緩和して用例コーパスから類似用例を検索し，その結果を用いて未知語の品詞を推定する手法を提案して来た．本稿では特に，品詞の分類体系，未知語に隣接する品詞列の形態素数が未知語品詞推定に与える効果（精度および実行速度）について詳しく検証する．
キーワード　未知語，品詞推定，形態素解析，対話システム

Effects of POS Categorization and Sequence Length in Unknown Word Category Inference based on Similarity between POS Sequence Patterns

Shun HATTORI†, Tomotaka FUKUOKA††, Chiaki KUBOMURA†††, and Hiroyuki KAMEDA†

† School of Computer Science, Tokyo University of Technology
1404–1 Katakura-machi, Hachioji, Tokyo 192–0982, Japan
†† Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology
1404–1 Katakura-machi, Hachioji, Tokyo 192–0982, Japan
††† The General Department of Aesthetics, Yamano College of Aesthetics
530 Yarimizu, Hachioji, Tokyo 192–0396, Japan
E-mail: †{hattori,kameda}@cs.teu.ac.jp, ††g2110045e1@gss.teu.ac.jp, †††ckubomura@yamano.ac.jp

Abstract　Recently, there has been an increase in research of extracting various knowledge from such text as Web/Blog documents and chatting (dialogue) logs. For example, researchers want to change Web text into morphemes and extract information in response to only the specific Part(s) Of Speech. However, researchers could miss valuable origins of knowledge, because recent text documents have more and more unknown words for the existing morphological analyzers and their word category inference for unknown words is not always accurate, As a solution to this problem, we have proposed a method of unknown word's category inference using the similar examples to an input sentence with a unknown word retrieved from a corpus of examples by expanding and/or relaxing a query of POS sequence surrounding the unknown word. This paper examines in detail effects of POS categorization and sequence length in unknown word's category inference based on similarity between POS sequence patterns.
Key words　Unknown Word, Word Category Inference, Morphological Analysis, Dialogue System

1. はじめに

Web 文書やブログ文書，チャット（対話）ログなどをテキスト解析し，様々な知識を抽出または発見する研究が盛んに行われている．例えば，実世界で提供されている製品やサービスなどの評判抽出 [1], [2]，実世界のある場所である期間に味わうことができる体験（イベント）のマイニング [3]〜[5]，語概念の階層構造（is-a/has-a 関係など）の抽出 [6], [7]，実世界オブジェクトの外観などの五感情報の抽出 [8]〜[11] など，実世界の様々な事象に関する知識を日々情報爆発し続ける Web から，特にブログなどの CGM からマイニングする研究が近年盛んに行われている．このような研究を行う際，形態素解析器によって Web テキストを形態素に区切り，特定の品詞の単語のみに反応して情報抽出を行いたい場合がよくある．しかしながら，Web テキストに限らず，近年のテキスト文書中には既存の形態素解析器にとって未知である語の割合が非常に増えて来ており [12]，未知語に対しては既存の形態素解析器を用いた品詞推定が必ずしも正確ではないため，有益な知識の元を逃してしまっている．

一方，近年の情報通信技術の進歩により，人間の対話相手として，人間に比べて膨大な情報の保持が可能なコンピュータが注目を浴びている．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て来ている．しかしながら，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，すなわ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，話し相手の人間に質問で返したり，頻繁に話題転換したりするなど，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，対話システムにおける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．システムが自動的に未知語の情報（品詞や意味など）を推定することで，既知語だけを含む発話に対してと同様に応答することが可能になると考えられる．

そこで我々は，入力文に出現した推定対象の未知語を直接的には含まないテキストデータ群を用いて，未知語の情報を推定する手法を研究している [13]〜[16]．未知語を用いた文例などの直接関係するデータ群を用いずに，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群を元に推定することで，例え全ての人間が知らないような単語であってもその情報の推定を可能とする手法を目指している．これまでの我々の研究において，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を行うシステムを作成している [13]．このシステムでは表層文字列に関して類似度を求めることで，多数存在する用例をより選別し，品詞推定の精度を高めている．多くの用例を抽出するために段階的な用例検索条件の緩和を行っており，その結果，用例検索では未知語とその直前・直後の単語だけを品詞並び条件とすることが有効であり，また，品詞推定時に使用する類似用例群は類似度の大きい極一部の用例のみを用いるべきであるという知見を得ている．さらに，未知語の品詞推定精度を向上させるため，品詞体系の見直しとして助詞の細分化を行い，用例検索における品詞並び条件の強化を図っている [14], [15]．しかしながら，入力文の品詞並びパターンで用例コーパスから類似用例を検索した後，入力文と各用例間の表層類似度に基づいて整順を行っているため，どうしても多くの計算時間が掛かってしまう．入力文 1 件を処理するのに数十秒を要してしまっており，実用性に欠けている．少なくとも，リアルタイム性が求められる対話システムでは使えないという問題が残っている．

この問題に対して我々は，既存の形態素解析器のラッパーとして，形態素解析器が有する辞書に未登録である未知語に対しても，その未知語を含んだ入力文の品詞並びパターンを適度に条件強化・緩和することで用例コーパスから類似用例を検索し，その結果をそのまま用いて未知語の品詞を推定することで，より高精度かつ高速な未知語品詞推定システムを提案する．本稿では特に，システムが使用する品詞の分類体系，未知語に隣接する品詞並びの形態素数 n が未知語品詞推定に与える効果（精度および実行速度）を詳しく検証する．

本論文の以下の構成を示す．2 章では，品詞並び条件の強化・緩和に基づく未知語品詞推定手法について提案する．3 章では，品詞の分類体系，未知語に隣接する品詞並びの形態素数 n が与える効果を評価実験によって検証する．最後に，4 章で本論文をまとめる．

2. 提案手法

本稿で提案する未知語の品詞推定手法は，MeCab や ChaSen などの既存の形態素解析器のラッパーとして働き，形態素解析器が有する辞書に未登録である未知語に対しても，その未知語を含んだ入力文の品詞並びパターンを適度に条件強化・緩和することで用例コーパスから類似用例を検索し，その結果をそのまま用いて未知語の品詞推定を行う．

2.1 概　　要

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．本稿のシステムが行う未知語処理は図 1 のように，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コーパスから類似用例検索し，その結果の類似用例群をそのまま用いて未知語品詞推定を行う．

Step 1. 品詞並び作成　入力文を形態素解析し，その結果に基づき，ある特定の（基本 13 分類あるいは細分化した）品詞体系において入力文の品詞並びパターンを作成する．

Step 2. 類似用例検索　ある特定の品詞体系において，入力文から得られた品詞並びパターンを未知語を含む任意の n 形態素まで緩和（削除）した品詞並び条件によって，用例データベースから類似用例を検索する．

Step 3. 未知語品詞推定　品詞並び条件で検索された類似用例群をそのまま用い，未知語に対応する箇所の品詞の頻度に基づいて未知語の品詞推定を行う．

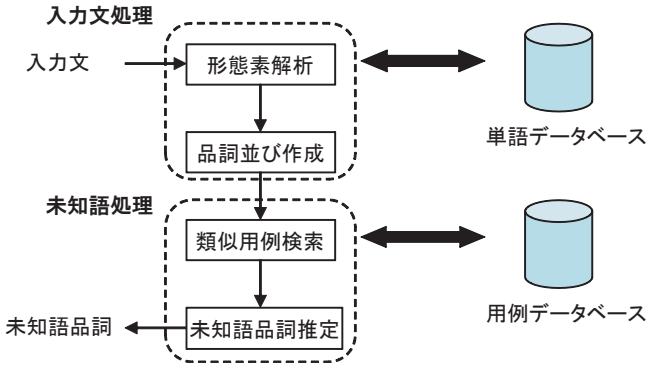

図 1　未知語品詞推定の処理の流れ

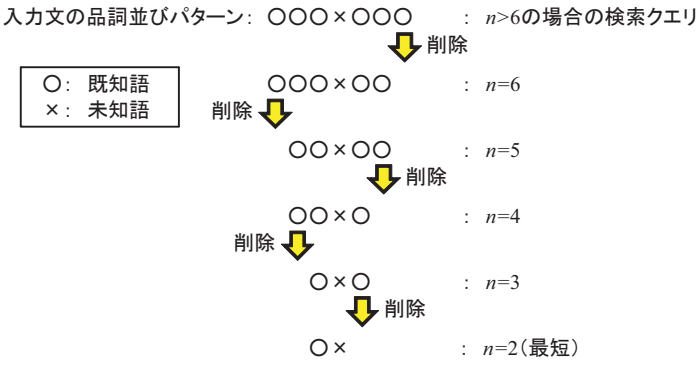

図 2　品詞並びパターンの条件緩和

2.2　データベース

提案システムは単語データベースと用例データベースの 2 種類のデータベースを持つ．単語データベースには形態素解析器（MeCab）が有する辞書（ipadic2.7.0）を用いる．用例データベースには Web 上に公開されている音声対話コーパス [17] を用い，発話文章を文単位に分解した 1812 文を利用している．

2.3　品詞分類体系

形態素解析器「MeCab」の品詞体系を用い，基本 13 分類に対して，各品詞の細分化を行うことで，Step 1 で作成される品詞並びパターンや，Step 2 で類似用例検索に用いられる品詞並び条件を強化する．本稿では以下のように 5 種類の品詞体系を新たに設定している．但し，各体系における品詞の掲載順は，MeCab のデフォルトの品詞 ID の順である．

基本 13 分類

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

名詞の細分化（名詞以外は細分化しない）

- サ変接続，ナイ形容詞語幹，一般，引用文字列，形容動詞語幹，固有名詞，数，接続詞的，接尾，代名詞，動詞非自立的，特殊，非自立，副詞可能．

動詞の細分化（動詞以外は細分化しない）

- 自立，接尾，非自立．

助詞の細分化（助詞以外は細分化しない）

- 格助詞，係助詞，終助詞，接続助詞，特殊，副詞化，副助詞，副助詞／並立助詞／終助詞，並立助詞，連体化．

記号の細分化（記号以外は細分化しない）

- アルファベット，一般，括弧開，括弧閉，句点，空白，読点．

ヒューリスティックに細分化

- 上述の助詞の細分化を適用．
- 名詞は，名詞または名詞-接尾にのみ細分化．
- 助動詞は，助動詞または助動詞-特殊・ダ/デスに細分化．
- 表層「て」かつ助詞-格助詞の場合，助詞-接続助詞に修正．

2.4　類似用例検索条件の段階的緩和

未知語を含む入力文と類似した用例をそれらの品詞並びパターンに基づいて検索するが，検索クエリが入力文全体の品詞並びパターンのままでは検索条件が厳し過ぎるため，十分な数の類似用例群が得られない場合が多い．そこで図 2 のように，入力文の形態素を文の端から段階的に削除（緩和）した結果の品詞並びパターンを検索クエリとする．品詞並びパターン条件として残す，未知語および未知語に隣接する形態素の数 n は 2 以上の任意の自然数である．また，入力文中における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，文末の形態素から順に削除して行く．文末が未知語であった場合は，先頭の形態素から順に削除して行く．それ以外の場合は，未知語の前後の形態素数が出来る限り等しく残るように，文の先頭と文末から形態素を交互に削除して行く．但し，未知語および未知語に隣接する品詞並びの形態素数 n が偶数の場合には未知語の前方にある形態素を優先して残している．

2.5　未知語品詞推定

入力文の品詞並びパターンで用例データベースから検索された類似用例群を全てそのまま用いる．これまでの我々の研究のように，入力文と用例間で表層の類似度を計算・整順を行う過程は経ない．また，検索結果の類似用例群の中で，入力文中の未知語の位置に相当する品詞の頻度をカウントすることによって，各品詞の推定比率を求めた上で，以下の 2 種類の手法により未知語の品詞を決定する．これまでの我々の未知語品詞推定手法では後者を採用していた．

（1）　最大確率の品詞を解として採択する

（2）　比率に基づいて確率的に品詞推定を行う

3.　評 価 実 験

提案した未知語品詞推定手法において，使用する品詞体系，未知語に隣接する品詞並び条件の形態素数 n などが与える効果（精度および実行速度）を検証する．以下に実験環境を示す．

PC：　IBM ThinkPad X40

OS：　Microsoft WindowsXP Professional Version 2002 SP3

CPU：　Intel Pentium M 1.2GHz

メモリ：　1.49GB RAM

言語：　Java 1.6.0_21

未知語を含んだ入力文を 21440 文用意し，品詞体系や条件緩和の形態素数 n などを変えて，提案した未知語品詞推定手法の精度および計算時間（各々3 回ずつ試行した平均）を求める．評価用の入力文は，漫画作品やライトノベル，マイクロブログなどを参考に作成している．システムが保持する用例コーパスは従来の日本語文法および対話形式により近いものを，未知語を含む評価用の入力文コーパスは従来の日本語文法および対話

形式とはより異なるものを選んでいる．21440 件の入力文における未知語の正解品詞の割合は以下の通りであり，図 3 は形態素数毎のヒストグラムである．

- 名詞：9612 文（45%）
- 動詞：1358 文（6.3%）
- 形容詞：3550 文（17%）
- 副詞：547 文（2.6%）
- 感動詞：1122 文（5.2%）
- 助詞：1058 文（4.9%）
- 助動詞：3723 文（17%）
- 連体詞：22 文（0.1%）
- 接頭詞：414 文（1.9%）
- 接続詞：34 文（0.2%）

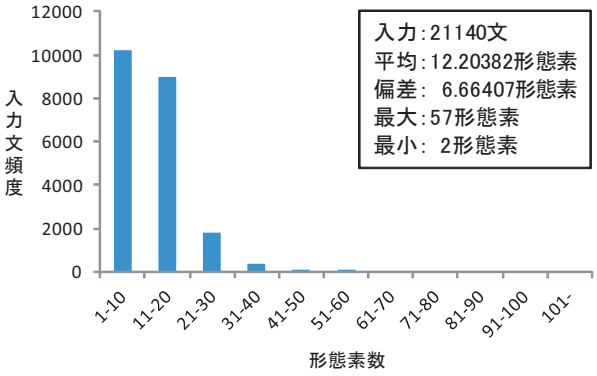

図 3　評価用入力文の形態素数毎ヒストグラム

3.1　品詞分類体系と品詞列長への依存性

類似用例検索条件の品詞分類体系（2.3 節）と，未知語および未知語に隣接する品詞列の形態素数 n（2.4 節）を変化させて，提案した未知語品詞推定手法の精度および計算時間を求める．

表 1 から表 6 は，未知語品詞推定ステップにおいて最大確率の品詞を解として採択する手法を用いた場合の各品詞体系における未知語品詞推定の平均精度を未知語に隣接する品詞並び条件の形態素数 n 毎に比較している．太字は，入力文の未知語の品詞毎に精度が最大となる形態素数 n を示している．また，図 4 から図 9 は，品詞体系，未知語に隣接する品詞並び条件の形態素数 n，未知語品詞推定ステップの 2 種類の手法を変化させて，推定精度および計算時間を比較している．形態素解析器「MeCab」（実際には Java 版の Sen を使用している）は未知語を全て名詞と推定するため，本稿の入力文 21440 件の内，名詞の未知語に対しては推定精度が 1.00 であるが，その他の品詞の未知語に対しては推定精度が 0.00 であり，平均精度は 0.448 である．また，平均計算時間は 0.227 ミリ秒であった．

形態素解析器「MeCab」の平均精度を上回っているのは，品詞分類体系として名詞の細分化，助詞の細分化，或いは，ヒューリスティックに細分化を用いて，品詞並びパターン条件緩和の形態素数 $n = 3$，最大確率の品詞を解として採択する手法で未知語品詞推定を行った場合である．提案手法の最良値はヒューリスティックに細分化を用いた場合の 0.509 であり，形態素解析器「MeCab」の平均精度を約 14%改善している．また，形態素解析器「MeCab」では名詞の未知語しか抽出できないが，本提案手法では助詞，助動詞，感動詞，動詞も精度良く抽出できており，名詞以外のキーワード抽出を必要とするようなアプリケーションには非常に有用であると考える．さらに，これまで我々が開発したシステム [14] では，入力文 1 件の処理に計算時間が数十秒ほど掛かっていたが，数ミリ秒と約 1000 倍高速化できている．次に，品詞毎の推定精度について見ていく．

動詞の未知語を推定したい場合，品詞分類体系としてヒューリスティックに細分化，品詞並びパターン条件緩和の形態素数 $n = 3$，最大確率の品詞を解として採択する手法を用いると最良値 0.331 を得ることができる．例えば，「今日の晩飯について**ククッ**てみる。」「キャストについて**グフッ**てみて下さい。」などの動詞の未知語を正しく推定できている．一方で，「いつでも行けるように**スタンバッ**とくんだぞ。」「**ダフッ**てはいけないが、かと言ってそれを意識し過ぎてトップしてもいけない。」などの動詞の未知語は推定失敗してしまっている．前者は検索結果の用例数が数百件と多いが，後者は数件または数十件と少ないことが推定失敗の主な原因であると考えられる．

形容詞の未知語を推定したい場合，品詞分類体系として助詞の細分化，品詞並びパターン条件緩和の形態素数 $n = 4$，比率に基づいて確率的に品詞推定する手法を用いても最良値 0.025 しか得ることができず，提案手法では非常に不十分である．例えば，「おおっ、こいつは**スゲェ**ぜ！」「いくら何でも、これ以上騒ぎになるのは**マズイ**なぁ。」「**カワイイ**とこあるじゃん…」などの形容詞の未知語を正しく推定できている．

副詞の未知語を推定したい場合も提案手法では非常に不十分である．品詞分類体系としてヒューリスティックに細分化，品詞並びパターン条件緩和の形態素数 $n = 3$，比率に基づいて確率的に品詞推定する手法を用いて最良値 0.066 を得ることしかできていない．例えば，「あんたが真道のように**イキナリ**彼女を抱き上げたら、叫ばれちゃうところよ。」「中は**シットリ**つやつやで旨そうっすね !! 」「赤神のことは**スッパリ**諦めたさ！」などの副詞の未知語を正しく推定できている．

感動詞の未知語を推定したい場合，品詞分類体系として名詞の細分化，品詞並びパターン条件緩和の形態素数 $n = 3$，最大確率の品詞を解として採択する手法を用いると最良値 0.400 を得ることができる．例えば，「オレてっきり同い歳だとばっかり、**スイマセン**！」「**ハァ**、お前そんなことで？」「**ゲッ**、呂布っ！」などの感動詞の未知語を正しく推定できている．

連体詞の未知語を推定したい場合，品詞分類体系としてヒューリスティックに細分化，品詞並びパターン条件緩和の形態素数 $n = 3$，比率に基づいて確率的に品詞推定する手法を用いても最良値 0.120 しか得ることができず，提案手法では未だ不十分である．例えば，「**ソウイウ**意味ではなくてですな。」などの連体詞の未知語を正しく推定できている．

接頭詞の未知語に対しても提案手法は非常に不十分である．接続詞の未知語に対しても多くの品詞分類体系では不十分であるが，名詞の細分化を用いた場合にだけ 0.206 と悪くない推定精度である．例えば，「**ソリャ**無いよ！」などが推定できている．

最後に，品詞並びパターン条件緩和の品詞列長 n について見てみると，形態素数 $n = 3$ の場合，つまり，推定対象の未知語および前後一形態素を含む品詞並びパターンに基づいて用例データベースから用例検索した場合に，品詞分類体系に概ね依らず，推定精度が最良かつ計算時間も最小になっている．

表 1　基本 13 分類と最大確率の品詞採択手法における推定精度の比較

$n=$	2	3	4	5	6	7	8	9	10
名詞	**0.79**	0.69	0.66	0.63	0.48	0.35	0.24	0.17	0.13
動詞	0.00	0.17	**0.25**	0.23	0.23	0.16	0.10	0.08	0.07
形容詞	0.00	0.00	0.00	0.01	**0.01**	0.01	0.00	0.00	0.00
副詞	0.00	0.00	0.02	0.02	**0.02**	**0.02**	0.02	0.01	0.01
感動詞	0.34	**0.39**	0.33	0.26	0.19	0.13	0.10	0.07	0.06
助詞	**0.68**	0.64	0.63	0.60	0.50	0.38	0.27	0.18	0.15
助動詞	0.21	0.38	**0.46**	0.44	0.36	0.24	0.16	0.12	0.10
連体詞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
接頭詞	0.00	0.00	**0.01**	0.00	0.00	0.00	0.00	0.00	0.00
接続詞	**0.09**	0.03	0.03	0.03	0.00	0.06	0.06	0.00	0.00
Avg.	**0.44**	0.44	0.44	0.42	0.33	0.24	0.16	0.12	0.09

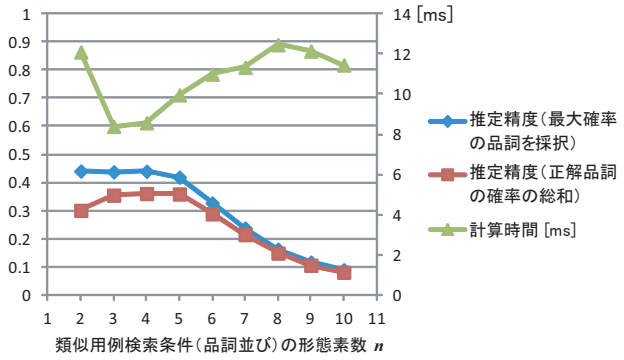

図 4　基本 13 分類による未知語品詞推定における精度と計算時間

表 2　名詞の細分化と最大確率の品詞採択手法における推定精度の比較

$n=$	2	3	4	5	6	7	8	9	10
名詞	**0.78**	0.67	0.60	0.50	0.28	0.15	0.09	0.07	0.06
動詞	0.00	0.18	**0.28**	0.23	0.13	0.07	0.04	0.03	0.03
形容詞	0.00	0.01	**0.02**	0.01	0.01	0.00	0.00	0.00	0.00
副詞	0.00	0.00	**0.03**	0.02	0.02	0.01	0.01	0.01	0.01
感動詞	0.34	**0.40**	0.31	0.18	0.11	0.08	0.06	0.06	0.06
助詞	**0.66**	0.66	0.62	0.51	0.34	0.21	0.12	0.07	0.07
助動詞	0.27	**0.52**	0.43	0.43	0.28	0.16	0.11	0.08	0.07
連体詞	0.05	0.05	0.05	**0.09**	0.00	0.00	0.00	0.00	0.00
接頭詞	0.00	0.01	**0.01**	0.01	0.00	0.00	0.00	0.00	0.00
接続詞	**0.21**	0.12	0.09	0.06	0.00	0.00	0.00	0.00	0.00
Avg.	0.45	**0.46**	0.41	0.35	0.21	0.12	0.07	0.05	0.05

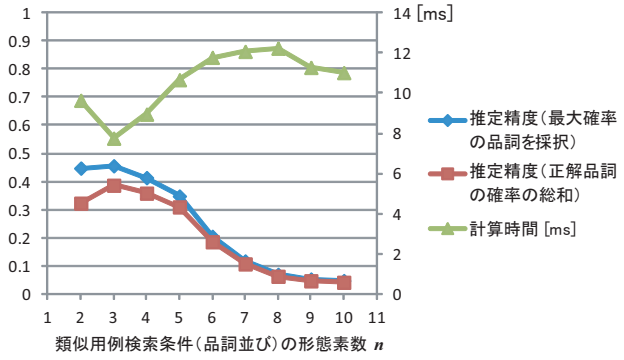

図 5　名詞の細分化による未知語品詞推定における精度と計算時間

表 3　動詞の細分化と最大確率の品詞採択手法における推定精度の比較

$n=$	2	3	4	5	6	7	8	9	10
名詞	**0.78**	0.69	0.66	0.63	0.49	0.35	0.24	0.16	0.12
動詞	0.00	0.21	**0.23**	0.19	0.16	0.10	0.06	0.05	0.04
形容詞	0.00	0.00	0.01	**0.01**	0.01	0.00	0.00	0.00	0.00
副詞	0.00	0.00	**0.03**	0.02	**0.03**	0.02	0.02	0.01	0.01
感動詞	0.34	**0.39**	0.33	0.25	0.18	0.13	0.09	0.07	0.06
助詞	**0.67**	0.63	0.62	0.56	0.45	0.33	0.23	0.16	0.14
助動詞	0.22	0.37	**0.45**	0.44	0.34	0.22	0.14	0.11	0.09
連体詞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
接頭詞	0.00	0.00	**0.01**	0.00	**0.01**	0.00	0.00	0.00	0.00
接続詞	**0.09**	0.03	0.03	0.00	0.03	0.06	0.06	0.00	0.00
Avg.	0.44	**0.44**	0.44	0.41	0.32	0.22	0.15	0.11	0.08

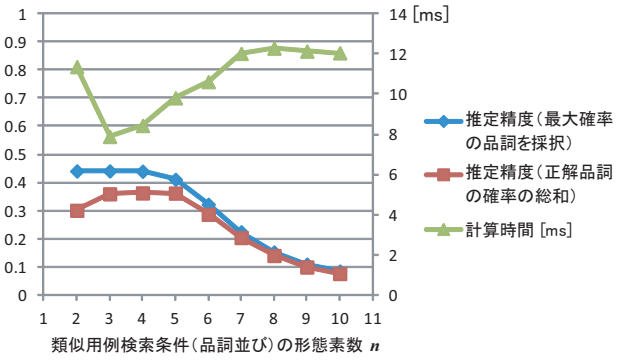

図 6　動詞の細分化による未知語品詞推定における精度と計算時間

表 4　助詞の細分化と最大確率の品詞採択手法における推定精度の比較

$n=$	2	3	4	5	6	7	8	9	10
名詞	**0.74**	0.72	0.64	0.51	0.30	0.17	0.10	0.07	0.06
動詞	0.08	0.21	**0.23**	0.19	0.15	0.07	0.04	0.03	0.03
形容詞	0.00	0.02	**0.02**	0.02	0.01	0.00	0.00	0.00	0.00
副詞	0.00	0.00	0.02	**0.04**	0.03	0.01	0.01	0.01	0.01
感動詞	0.34	**0.39**	0.33	0.23	0.14	0.09	0.06	0.05	0.04
助詞	**0.69**	0.66	0.64	0.57	0.43	0.28	0.17	0.11	0.10
助動詞	0.21	**0.58**	0.44	0.39	0.27	0.14	0.09	0.08	0.08
連体詞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
接頭詞	0.00	0.00	**0.01**	0.01	**0.01**	0.01	0.00	0.00	0.00
接続詞	**0.09**	0.03	0.03	**0.09**	0.00	0.00	0.00	0.00	0.00
Avg.	0.43	**0.49**	0.43	0.35	0.22	0.12	0.07	0.06	0.05

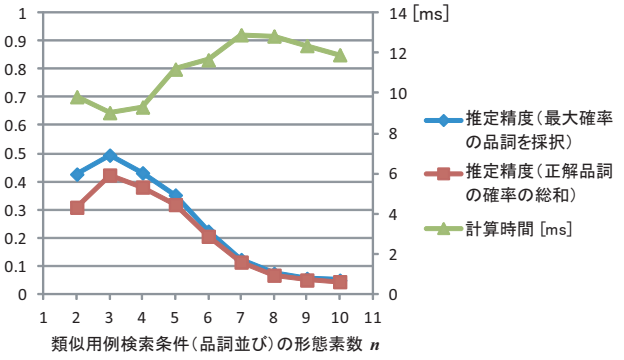

図 7　助詞の細分化による未知語品詞推定における精度と計算時間

表 5　記号の細分化と最大確率の品詞採択手法における推定精度の比較

$n=$	2	3	4	5	6	7	8	9	10
名詞	**0.76**	0.68	0.62	0.56	0.41	0.28	0.18	0.11	0.06
動詞	0.00	0.15	**0.23**	0.20	0.20	0.11	0.07	0.04	0.03
形容詞	0.00	0.00	0.00	0.01	**0.01**	0.01	0.00	0.00	0.00
副詞	0.00	0.00	0.02	**0.02**	0.02	0.01	0.01	0.00	0.00
感動詞	0.32	**0.37**	0.30	0.23	0.15	0.11	0.08	0.06	0.04
助詞	**0.68**	0.51	0.47	0.42	0.31	0.22	0.14	0.08	0.06
助動詞	0.21	0.28	**0.40**	0.19	0.15	0.10	0.07	0.06	0.05
連体詞	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
接頭詞	0.00	0.00	**0.01**	0.00	0.00	0.00	0.00	0.00	0.00
接続詞	**0.09**	0.03	0.03	0.03	0.00	0.06	0.06	0.00	0.00
Avg.	**0.43**	0.41	0.40	0.33	0.25	0.17	0.11	0.07	0.04

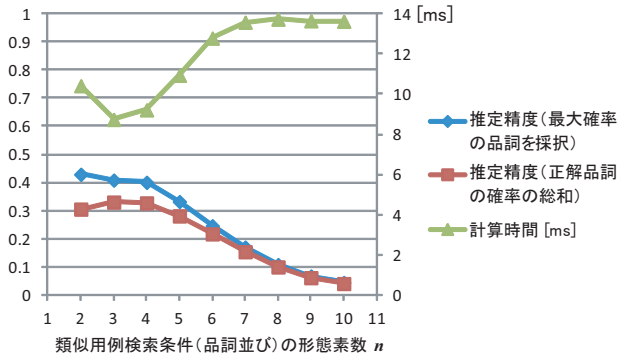

図 8　記号の細分化による未知語品詞推定における精度と計算時間

表 6　ヒューリスティック細分化と最大確率品詞採択における精度比較

$n=$	2	3	4	5	6	7	8	9	10
名詞	**0.79**	0.78	0.63	0.45	0.26	0.13	0.07	0.05	0.05
動詞	0.12	**0.33**	0.33	0.20	0.10	0.05	0.02	0.02	0.02
形容詞	0.00	**0.02**	0.02	0.01	0.01	0.00	0.00	0.00	0.00
副詞	0.00	0.05	**0.06**	0.04	0.02	0.01	0.00	0.00	0.00
感動詞	0.34	**0.39**	0.33	0.21	0.12	0.07	0.05	0.05	0.04
助詞	0.52	**0.64**	0.60	0.48	0.31	0.18	0.11	0.07	0.07
助動詞	0.21	**0.47**	0.40	0.34	0.21	0.12	0.08	0.07	0.07
連体詞	0.00	**0.09**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
接頭詞	0.00	**0.01**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
接続詞	**0.09**	0.06	0.06	0.06	0.03	0.00	0.00	0.00	0.00
Avg.	0.44	**0.51**	0.42	0.31	0.18	0.09	0.06	0.04	0.04

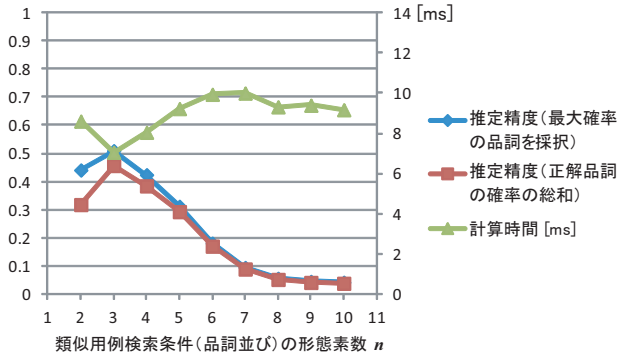

図 9　ヒューリスティック細分化による未知語品詞推定精度と計算時間

3.2　類似度計算・整順処理の有無への依存性

本稿のように，入力文中の未知語に隣接する形態素数 n の品詞並びパターンを条件に用例データベースから検索された類似用例群を全てそのまま用いて未知語品詞推定を行う（2.4 節）のではなく，これまでの我々の研究 [14] のように，入力文と類似用例間で表層の類似度を計算・整順を行って上位 k 件の類似用例群のみにフィルタリングした上で未知語品詞推定を行う場合の精度および計算時間を求める．但し，入力文 x と類似用例 y との間の表層類似度 $w(x,y)$ は，入力文を未知語に隣接する形態素数 n に条件緩和した後の表層の集合 $S(x)$，及び，類似用例から入力文の品詞並び条件に類似した一部を取り出した後の表層の集合 $S(y)$ を用いて，以下の Jaccard 係数で計算する．

$$w(x,y) := \frac{|S(x) \cap S(y)|}{|S(x)| + |S(y)| - |S(x) \cap S(y)|}$$

表 7 は，ヒューリスティック細分化の品詞分類体系，及び，類似用例検索条件（品詞並び）の形態素数 $n=3$ を用い，未知語品詞推定ステップにおいて類似度計算・整順処理を行った上で上位 k 件の類似用例群における最大確率の品詞を解として採択する手法の平均推定精度を比較している．また，図 10 は上位 k 件毎に推定精度と計算時間を比較している．

上位 k 件に依存せず，ほぼ一定の推定精度と計算時間である．平均精度は $k=2$ のとき最良値 0.542 となり，類似度計算・整順処理無しと比べて 6%改善しており，未知語の品詞毎に見ても改善しているが，計算時間が 2.8 倍に大きく悪化してしまう．

表 7　類似度計算・整順処理有り未知語品詞推定における精度比較

$k=$	1	2	3	4	5	10	50	100	200
名詞	0.69	0.76	0.78	0.77	0.77	**0.78**	0.78	0.77	0.77
動詞	0.36	**0.39**	0.39	0.39	0.38	0.37	0.37	0.36	0.34
形容詞	0.03	**0.07**	0.05	0.05	0.05	0.03	0.02	0.02	0.02
副詞	0.03	**0.09**	0.07	0.08	0.06	0.05	0.05	0.05	0.05
感動詞	0.29	0.37	0.37	0.36	0.37	0.38	0.39	**0.39**	**0.39**
助詞	0.52	**0.69**	0.63	0.63	0.65	0.61	0.69	0.67	0.66
助動詞	0.53	**0.63**	0.59	0.59	0.55	0.56	0.52	0.50	0.48
連体詞	0.11	0.08	0.12	0.14	**0.17**	0.09	0.09	0.09	0.09
接頭詞	0.02	0.03	**0.04**	0.03	0.01	0.01	0.01	0.01	0.01
接続詞	0.07	**0.19**	0.10	0.11	0.08	0.06	0.06	0.06	0.06
Avg.	0.47	**0.54**	0.54	0.53	0.53	0.53	0.52	0.51	0.51

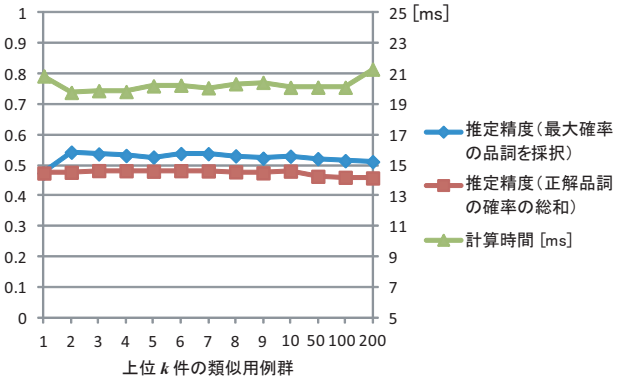

図 10　類似度計算・整順処理有り未知語品詞推定の精度と計算時間

3.3　用例データベースのサイズと種類への依存性

用例データベース（2.2 節）のサイズと種類を変化させて，本稿で提案した未知語品詞推定手法の精度および計算時間を求める．音声対話コーパス [17] は，発話文章を文単位に分解した 1812 文からなり，図 11 は形態素数毎のヒストグラム，図 12 は文字列長毎のヒストグラムである．一方，インタビュー形式による日本語会話データベース（上村コーパス）[18] は，発話文章を文単位に分解した 20728 文からなり，図 14 は形態素数毎のヒストグラム，図 15 は文字列長毎のヒストグラムである．

図 13 は，用例データベースとして音声対話コーパスと（のサブセット），品詞分類体系としてヒューリスティック細分化，及び，類似用例検索条件（品詞並び）の形態素数 $n = 3$ を用い，未知語品詞推定ステップにおいて類似度計算・整順処理を行わないで検索された全ての類似用例群における最大確率の品詞を解として採択する手法の推定精度と計算時間を，用例データベースのサイズ毎に比較している．一方，図 16 は上村コーパス（のサブセット）を用いて同様に比較している．但し，サブセットは無作為抽出によって作成している．

二つのグラフより，提案した未知語品詞推定手法の計算時間は用例データベースのサイズに対して線形に変化し，一方，推定精度は（指数）曲線を描いている．用例データベースのサイズが 1400 文の辺りで推定精度はピーク値 0.512 を取り，それ以上にサイズを大きくしてもほとんど変化していない．上村コーパス（のサブセット）での推定精度のピーク値は 0.509 であり，

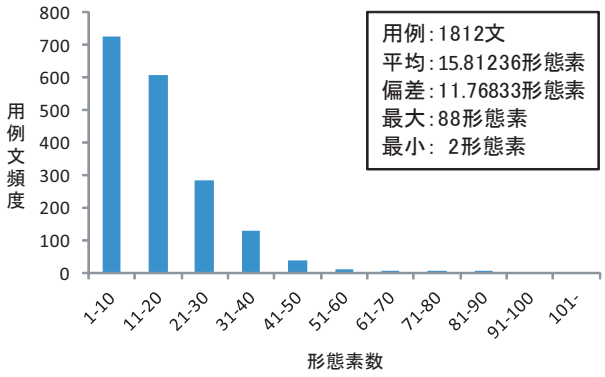

図 11　音声対話コーパスの用例文の形態素数毎ヒストグラム

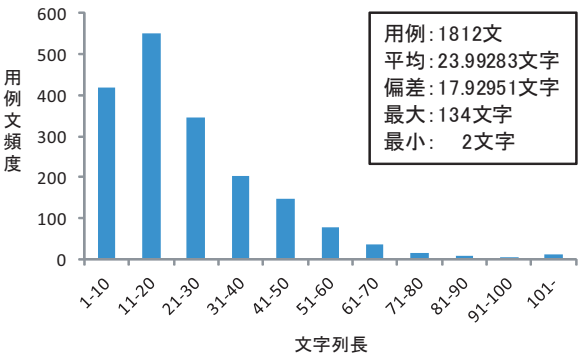

図 12　音声対話コーパスの用例文の文字列長毎ヒストグラム

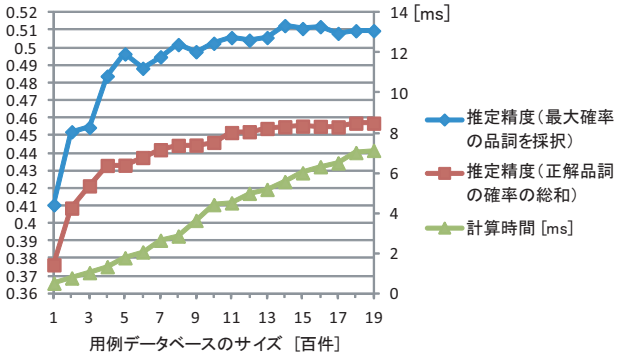

図 13　音声対話コーパスのサイズに依る推定精度と計算時間

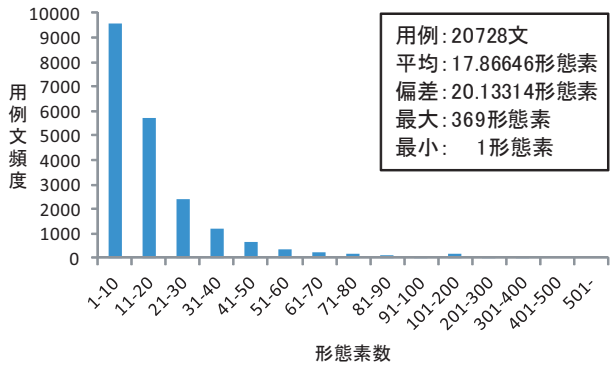

図 14　上村コーパスの用例文の形態素数毎ヒストグラム

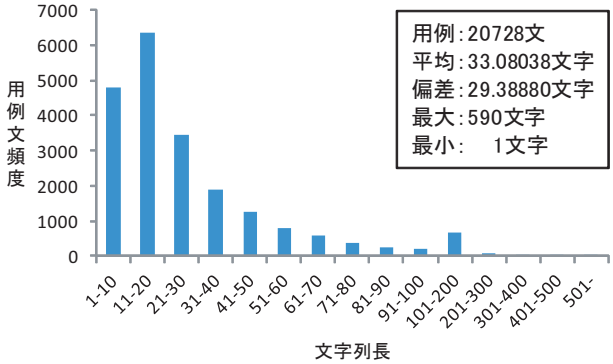

図 15　上村コーパスの用例文の文字列長毎ヒストグラム

図 16　上村コーパスのサイズに依る推定精度と計算時間

音声対話コーパスの場合とほとんど変わらず，推定対象のテキストに特化させていない一般的なコーパスを用いる限りは用例データベースの種類はほとんど影響が無いと言える．

音声対話コーパスのサブセットのサイズが 1400 文の辺りで推定精度はピーク値を取るが，音声対話コーパスのフルセットを用いた場合の推定精度とほとんど変わらない一方で，計算時間は 5.6 ミリ秒と 1.3 倍高速化することができるため，本提案手法における最良なパラメータ設定であると考える．とは言え，元々の形態素解析器「MeCab」による計算時間 0.227 ミリ秒と比べると格段に遅い．ユーザが利用する事前に行っておくことができるデータマイニング等では問題無いかもしれないが，リアルタイム応答が必要となるような対話システムへの適用を考えると，推定精度は維持しつつ，さらなる高速化が必要である．

4. まとめと今後の課題

Web 文書やブログ文書，チャット（対話）ログなどをテキスト解析し，様々な知識を抽出または発見する研究が盛んに行われている．しかしながら，Web テキストに限らず，近年のテキスト文書中には既存の形態素解析器にとって未知である語の割合が非常に増えて来ており，未知語に対しては既存の形態素解析器を用いた品詞推定が必ずしも正確ではないため，有益な知識の元を逃してしまっている．この問題に対して我々は，既存の形態素解析器のラッパーとして，形態素解析器が有する辞書に未登録である未知語に対しても，その未知語を含んだ入力文の品詞列パターンを適度に条件強化・緩和することで用例コーパスから類似用例を検索し，その結果を用いて未知語の品詞を推定する手法を提案した．品詞の分類法，未知語に隣接する品詞列の形態素数が未知語品詞推定に与える効果（精度および実行速度）を検証した結果，形態素解析器「MeCab」の未知語品詞推定手法の平均精度を約 14%改善し，かつ，これまでの我々の手法 [14] よりも約 1000 倍高速化して実用的な計算時間（入力文 1 件当たりの平均処理時間約 10 ミリ秒）を実現できている．平均では形態素解析器「MeCab」の精度をわずかに上回った程度であり，計算時間が大きく掛かってしまうため，依然として未だ不十分であると考えるが，名詞以外にも動詞や感動詞，助詞，助動詞の未知語も精度良く取得することができること，また，抽出したい品詞の種類に合わせてパラメータを最適化することも可能であることの 2 点において優れている．

今後は，用例データベースとして異なるコーパスや異なるサイズを用いた場合の推定精度および実行速度の比較評価を行う予定である．また，動詞や形容詞など，キーワード抽出のニーズは高いが依然として推定精度が不十分な品詞への対応策や，用例データベースとして Web を活用することも検討して行く．

謝　　辞

本研究は科学研究費助成事業（学術研究助成基金助成金）若手研究（B）（研究代表者 : 服部峻，課題番号 : 23700129）「ウェブから時空間依存データを抽出するウェブセンサに関する研究」の助成を受けたものである．ここに記して謝意を表す．

文　　献

[1] 鈴木 泰裕，高村 大也，奥村 学: “WebLog を対象とした評価表現抽出,” 第 6 回セマンティックウェブとオントロジー研究会, SIG-SWO-A401-02 (2004).

[2] 藤村 滋，豊田 正史，喜連川 優: “文の構造を考慮した評判抽出手法,” 電子情報通信学会 第 16 回データ工学ワークショップ (DEWS2005), 6C-i8 (2005).

[3] Tezuka, T., Kurashima, T., and Tanaka, K.: “Toward Tighter Integration of Web Search with a Geographic Information System,” Proceedings of the 15th International World Wide Web Conference (WWW’06), pp.277–286 (2006).

[4] 倉島 健，藤村 考，奥田 英範: “大規模テキストからの経験マイニング,” 電子情報通信学会 第 20 回データ工学ワークショップ (DEWS2008), A1-4 (2008).

[5] Inui, K., Abe, S., Morita, H., Eguchi, M., Sumida, A., Sao, C., Hara, K., Murakami, K., and Matsuyoshi, S.: “Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents,” Proceedings of the 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI’08), pp.314–321 (2008).

[6] 服部 峻，田中 克己: “性質継承と概念の再帰的適用に基づく Web からの概念階層抽出,” 情報処理学会論文誌（トランザクション）データベース，Vol.1, No.3 (TOD40), pp.60–81 (2008).

[7] Hattori, S., and Tanaka, K.: “Extracting Concept Hierarchy Knowledge from the Web based on Property Inheritance and Aggregation,” Proceedings of the 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI’08), pp.432–437 (2008).

[8] 服部 峻，手塚 太郎，田中 克己: “文書中の地物画像を言語的記述で代替するための地物の外観情報の Web からの抽出,” 情報処理学会論文誌（トランザクション）データベース，Vol.48, No.SIG11 (TOD34), pp.69–82 (2007).

[9] Hattori, S., Tezuka, T., and Tanaka, K.: “Mining the Web for Appearance Description,” Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA’07), LNCS Vol.4653, pp.790–800 (2007).

[10] 服部 峻，田中 克己: “Web 抽出した特異な色名と色特徴量変換に基づく特異画像の Web 検索,” 情報処理学会論文誌（トランザクション）データベース，Vol.3, No.1, pp.49–63 (2010).

[11] Hattori, S.: “Cross-Language Peculiar Image Search Using Translation between Japanese and English,” Proc. of the First IRAST International Conference on Data Engineering and Internet Technology (DEIT’11), pp.418–424 (2011).

[12] 服部 峻，亀田 弘之: “Web テキストにおける未知語の頻度調査,” 電子情報通信学会 思考と言語研究会 (SIG-TL), 信学技報, Vol.110, No.63, TL2010-2, pp.7–12 (2010).

[13] 福岡 知隆，服部 峻，久保村 千明，亀田 弘之: “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,” 第 90 回 人工知能学会 知識ベースシステム研究会 (SIG-KBS), 人工知能学会研究会資料, SIG-KBS-B001-04, pp.23–30 (2010).

[14] 福岡 知隆，服部 峻，久保村 千明，亀田 弘之: “品詞並び検索条件の緩和と強化による用例ベース未知語品詞推定に関する諸検討,” HAI シンポジウム 2010 (HAI’10), 3D-2 (2010).

[15] Tomotaka Fukuoka, Shun Hattori, Chiaki Kubomura, and Hiroyuki Kameda: “Example-based Inference of Unknown Word Category by a Surrounding POS Sequence,” Proceedings of the 12th Conference of the Pacific Association for Computational Linguistics (PACLING’11), #38 (2011).

[16] 福岡 知隆, 服部 峻, 久保村 千明, 亀田 弘之: “単語間の意味カテゴリー距離に基づく用例ベース未知語意味カテゴリー推定,” 第 10 回 情報科学技術フォーラム (FIT’11), 4F-4 (2011).

[17] 重点領域研究音声対話コーパス，`http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/` (2011).

[18] 上村隆一: 平成 8-10 年度文部省科学研究費補助特定領域研究「人文科学とコンピュータ」公募研究「日本語会話データベースの構築と談話分析」（研究代表者 : 上村隆一）の成果による．