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

Studies of Example-based Inference of Unknown Word Category by Query Relaxation and Reinforcement 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 our example-based inference of unknown word category. In the previous papers, we already become able to get much more similar examples to an input sentence by query relaxation of its POS sequence in example searches. To get our word category inference more accurate, we reinforce a query of POS sequence by subdividing the kinds of particles more finely. Then we investigate the effect of the other factors to the accuracy: formula for similarity, size of example database, kinds of corpus to compose the example database, and the number of similar examples used to the inference. We rearrange these factors, and evaluate each word category inference. As a result, we got conclusion that the query reinforcement is not definitely effective to improve the accuracy but effective to reduce the computational time, and how each factor effects the accuracy.

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こってしまい，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，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．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上手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

*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

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．

このような手法の一つに，N-gram を用いて未知語の品詞を確率的に求める手法が存在するが，我々はより品詞推定精度を向上させるため，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]．

このシステムでは表層文字列における類似度を求めることで，多数存在する用例を選別し，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，用例検索は未知語とその前後の単語の品詞並びをクエリ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一部のもののみを用いるべきであるという知見を得た．

本稿では，まず，このシステムにおける未知語の品詞推定精度を向上させるため，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た．また，三種類の類似度の計算式を用いた場合，システムの保持するデータベースの規模，データベースに用いるコーパス，品詞推定に使用する類似用例数ごとの品詞推定精度の比較を行った．

2 提案手法

本稿で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使用する類似用例の絞り込みが可能となる．

また，提案手法は形態素解析器 MeCab により未知語の検出が正確に行われていることを前提としている．

2.1 提案システムの処理手順

システムは以下の手順で未知語の品詞推定を行う．

step1：入力文情報取得 MeCab により入力文を形態素解析し，その結果に基づき，条件強化を行った入力文品詞並びを作成する．

step2：類似用例検索条件の緩和 得られた品詞並びを未知語を含む三形態素まで削除することにより検索条件を緩和する．

step3：未知語品詞推定 緩和した入力文品詞並び条件を検索クエリとして用例を検索し，それぞれの用例の入力文との類似度を計算し，任意の個数の類似用例を用いて品詞推定する．

2.2 システムのデータベース

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二つのデータベースを持つ．単語データベースには ipadic2.7.0 を用いる．用例データベースは Web 上に公開されている二種類の音声対話コーパス [2, 3] を用い，必要な許諾を得て，発話文章を文単位に分解したものを利用した．

具体的にはコーパス [2] からランダムに発話文を選択し，用例数が 100 から 1200 までの 12 セット（パターン a），及び，コーパス [3] からランダムに発話文を選択し，用例数が 100 から 1200 までの 12 セット（パターン b）と 500 から 14500 までの 29 セット（パターン c）を用例データベースとした．また，用例検索条件の強化を実現するため，品詞を 13 個（表 1）として強化なしとした場合と助詞を細分化し 22 個（表 2）として強化ありとした場合を作成した．

表 1: MeCab における品詞の分類

名詞	動詞	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
助詞			

表 2: 助詞を細分化した品詞の分類

名詞	動詞	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
格助詞	**副助詞**	**終助詞**	**係助詞**
接続助詞	**並立助詞**	**副詞化**	**連体化**
副助詞／並立助詞／終助詞			**特殊**

また，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，会話文の中の一文である．品詞並びは，MeCab を用いて用例の形態素解析をした結果，それぞれの単語に品詞を割り当て，そ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延べで何度出現したかを表す．

2.3 用例検索条件緩和

前の研究 [1] で得た知見に基づき，入力単語を削除することで用例検索条件の緩和を行った．未知語を含み，入力文が三形態素になるように削除する．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，未知語とその後に二形態素が続く．文末が未知語であった場合は，二形態素が続き，その後ろに未知語が存在する．それ以外の場合は未知語とその前後の単語が存在する．

2.4 類似度計算式

類似度の計算式は以下の三つの式のいずれかを用いる．いずれの式も文中の表層文字列を要素としている．w を類似度，X を入力文に出現する単語群，Y を類似文に出現する単語群とする．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．

ダイス係数

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

入力文は条件緩和を行った後の三形態素であり，類似文は用例から入力文の品詞並びに類似した一部を取り出したものである．

Jaccard 係数

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

入力文は条件緩和を行った後の三形態素であり，類似文は用例から入力文の品詞並びに類似した一部を取り出したものである．

コサイン類似度

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

入力文は条件緩和を行う前の原文であり，類似文は用例そのものである．

2.5 品詞推定手法

品詞並び条件で検索された用例の内，類似度が大きい順に類似用例を任意の自然数である n 個選択する．それぞれの類似用例検索には推定される品詞が一意に決定されている．選択された類似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．類似度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（図 1）．今回の実験においては n は 1 から 10 の間で使用した．

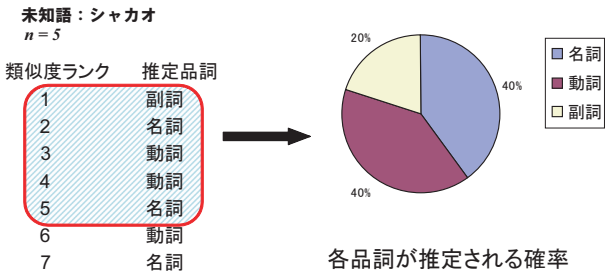

図 1: 類似用例の比率による確率的推定

3 評価実験

以下に実験概要とその結果及び考察を示す．

3.1 実験概要

Web 上の資料 [4] を参考に未知語を含んだ入力文を 28 個用意し，条件強化の有無による用例検索結果を比較し，2.2〜2.5 に記述した項目を変化させたそれぞれの品詞推定精度を比較する．実験条件は類似度計算式，条件強化の有無，用例データベースの規模，用例データベースに用いたコーパス，使用類似用例数 n である．

3.2 結果と考察

図 2 にパターン a，図 3 にパターン c における用例検索条件の強化の有無，用例データベースの規模による用例検索結果を示す．条件強化を行わない場合を強化なし，行った場合を強化ありとしている．

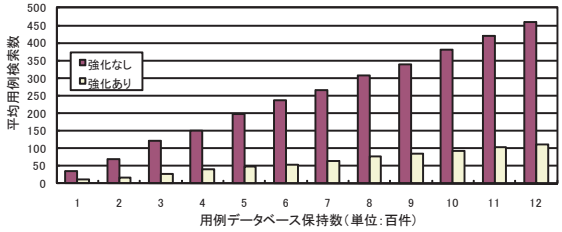

図 2: パターン a における条件強化による用例検索結果

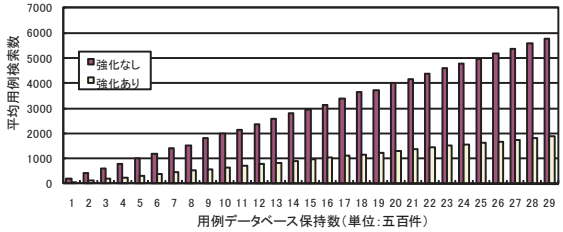

図 3: パターン c における条件強化による用例検索結果

検索条件を強化することにより，平均数は 1/3 以下まで減少している．また，データベースの規模による平均数の増加数はほぼ一定である．このため，品詞推定結果の偏りは小さいものと考えられる．

以下は各類似度計算式によるそれぞれの結果を示す．

品詞推定精度は正答数を品詞の推定可能であった入力文の個数で割った値である．正答数はそれぞれの入力文において，正しい品詞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．

3.2.1 ダイス係数

図 4，図 5，図 6 にそれぞれパターン a，b，c における条件強化の有無，用例データベースの規模の拡大による品詞推定精度の結果とそれぞれの近似直線を示す．これらの図で記されている平均推定精度とは品詞推定

に使用する類似用例数 n を 1 から 10 まで変化させた 10 セットの品詞推定結果の平均値を指す．

これらの結果から，用例検索条件の強化は品詞推定精度の向上に大きな影響を与えていないと言える．また，用例データベースの規模を拡大させることが品詞推定精度の向上に影響を与えている．用例データベースに用いたコーパスの違いは推定精度に大きな変化をもたらしてはいない．

また，図 7 に使用用例数 n の変化による品詞推定結果を条件強化の有無別に示す．この図における平均品詞推定精度は，パターン c における 29 セットの用例データベース規模の異なる品詞推定結果の平均値である．

この結果から品詞推定精度が最も大きくなるのは，類似度の大きい上位二件を用いた場合である．

3.2.2 Jaccard 係数

同様に条件強化，及び，用例データベース規模の拡大による品詞推定結果を図 8，図 9，図 10 に示し，条件強化，及び，使用用例数 n による品詞推定結果を図 11 に示す．

これらの結果から，3.2.1 と同様に，用例データベースの規模の拡大によって品詞推定精度が向上する一方，用例検索条件の強化の品詞推定精度の改善への影響が小さいことが言える．また，品詞推定に使用する類似用例数も類似度の大きい二件とすることが最も有効である．

3.2.3 コサイン類似度

同様に条件強化，及び，用例データベース規模の拡大による品詞推定結果を図 12，図 13，図 14 に示し，条件強化，及び，使用用例数 n による品詞推定結果を図 15 に示す．

これらの結果から，用例データベースの規模の拡大による品詞推定精度の向上と，類似用例検索条件の強化の推定精度の向上に大きく影響していることが言える．また，品詞推定に使用用例数は四件以上，十件以下ならば，結果に大差がなく有効である．

3.3 総合的な考察

上記の結果から，各条件に対する考察を述べる．

3.3.1 検索条件の強化

今回の実験において，類似用例検索条件の強化は助詞の細分化によって行われた．その結果，コサイン類似度を用いた場合にその品詞推定精度の向上に大きな影響を与えた．しかし，コサイン類似度を用いた結果は他の結果と比較し，その品詞推定精度は低い．

今回用いたコサイン類似度の計算式は用例データベース内の単語の出現率を用いた重み付け処理を省略しているため，短い用例が類似度の上位に来やすくなっている．少なくとも，今回用いた用例データベースの中では短い用例の中では，正答を持つ用例が少なかったためコサイン類似度の品詞推定精度が低いものとなったと考えられる．用例が短くなれば，その品詞並びのパターンも少なくなるため，助詞を細分化することで，より厳密に用例の類似性を測ることとなり，品詞推定精度が向上したと考えられる．

また，品詞推定精度の向上には大きく影響はしていないが，品詞推定精度を維持しつつ，検索用例数を 1/3 から 1/4 に圧縮できている．コーパス [3]，用例データベースの保持数 14500，コサイン類似度における 28 の入力文セットそれぞれの step3 における類似度計算，及び，類似度上位一件を取得するための計算時間が，強化なしの場合は平均 23.4 秒（標準偏差 14.9 秒）であるのに対し，強化ありの場合は平均 9.76 秒（標準偏差 8.21 秒）と 1/2 以上短縮することができ，処理速度の向上には有効である．ただし，MacBookAir（CPU:1.86GHz, Memory:2GB）を用いて Ruby 言語で実装した．

3.3.2 用例データベース

今回の実験の結果，用例データベースの保持用例数の拡大は品詞推定精度の向上に有効であるという知見を得た．これは，用例データベースの保持用例が増加することで，より多くの正しい品詞推定結果が得られる用例が増加するためであると考えられる．正しくない結果を得る用例も増加するが，用例の類似度を用いることで，そういった用例は除くことが可能である．

しかし，保持用例数の増加による品詞推定精度の向上は，本研究の目的においては必ずしも有効ではない．本研究の目的は円滑な対話であるため，用例データベースの増加は処理時間の増加を招き，対話応答の遅延に繋がると考えられるためである．

そのため，対話応答に支障が無い範囲で最大の効果が得られるデータベースの規模の設定が重要である．

コサイン類似度を用いた場合，他の計算式に比べ明らかに用例データベースの規模による推定精度の向上率が小さいのは，類似度の上位に上ってくる用例数が絶対的に少ないためと考えられる．

3.3.3 使用類似用例数 n

コサイン類似度を用いた場合は使用用例数は $n = 4$ 以上から品詞推定精度に大差がなくなっている．これはコサイン類似度においては，短い用例が上位に来やすいことを考えると，より多くの類似用例を用いることで，正答を含む用例を利用することができるためと考えられる．

ダイス係数，Jaccard 係数を用いた場合は，使用する類似用例数は二件とすることが有効である．使用用例数を増やすとその精度は下がっており，表層文字列の類似度のみに依存した手法では，間違った結果を導く用例を多く残していると考えられる．

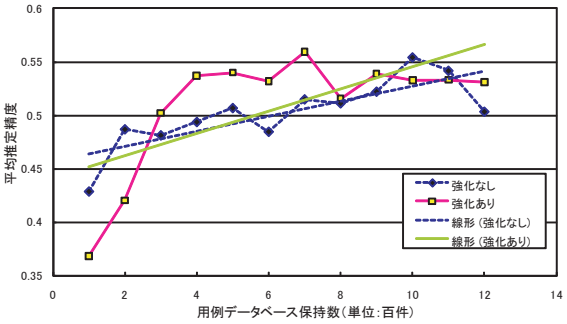

図 4: パターン a におけるダイス係数を用いた条件強化，用例データベース毎の平均品詞推定精度

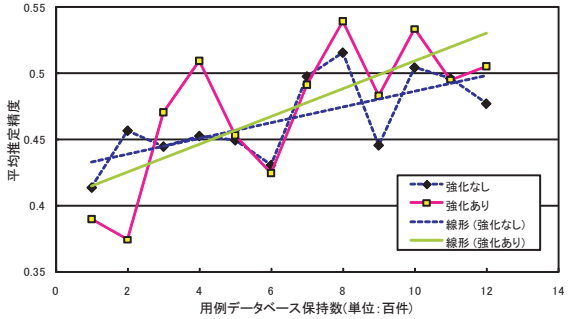

図 5: パターン b におけるダイス係数を用いた条件強化，用例データベース毎の平均品詞推定精度

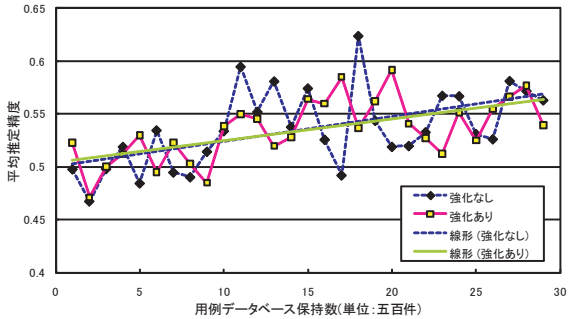

図 6: パターン c におけるダイス係数を用いた条件強化，用例データベース毎の平均品詞推定精度

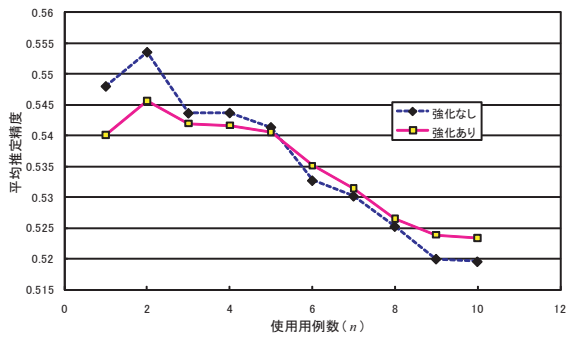

図 7: パターン c におけるダイス係数を用いた場合の使用用例数 n による平均品詞推定精度

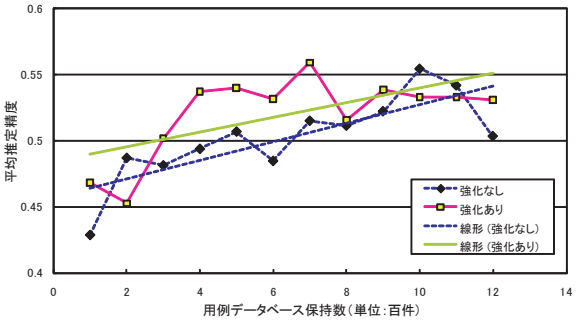

図 8: パターン a における Jaccard 係数を用いた条件強化，用例データベース毎の平均品詞推定精度

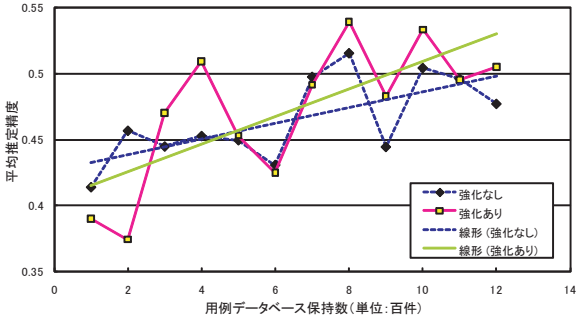

図 9: パターン b における Jaccard 係数を用いた条件強化，用例データベース毎の平均品詞推定精度

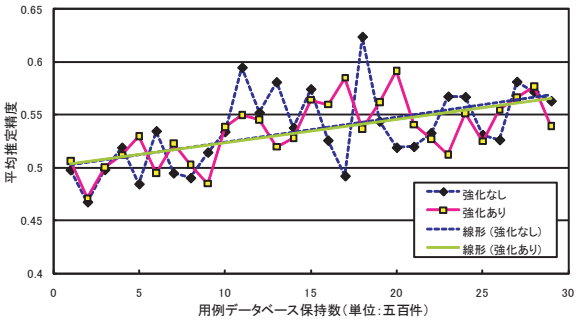

図 10: パターン c における Jaccard 係数を用いた条件強化，用例データベース毎の平均品詞推定精度

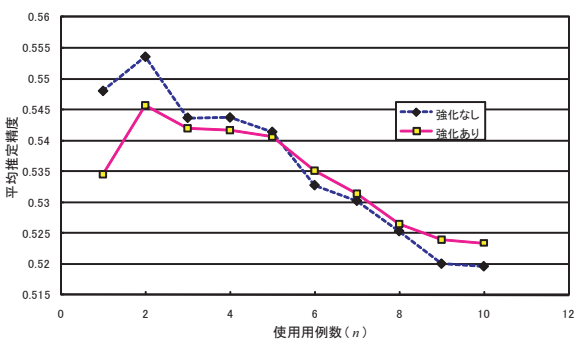

図 11: パターン c における Jaccard 係数を用いた場合の使用用例数 n による平均品詞推定精度

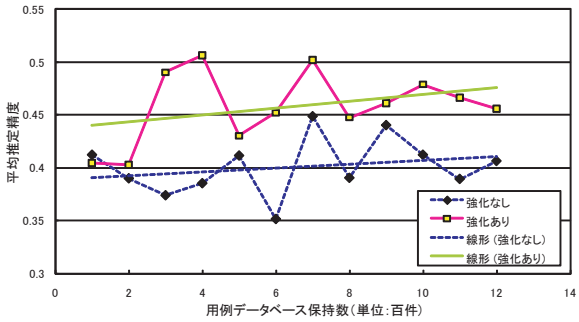

図 12: パターン a におけるコサイン類似度を用いた条件強化，用例データベース毎の平均品詞推定精度

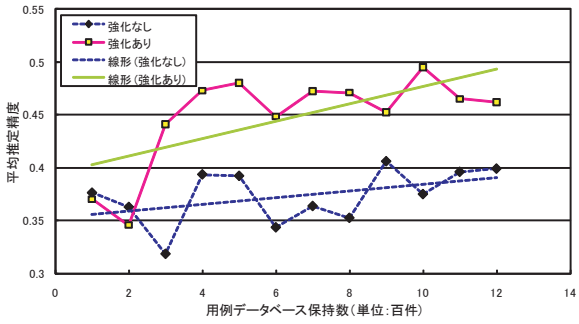

図 13: パターン b におけるコサイン類似度を用いた条件強化，用例データベース毎の平均品詞推定精度

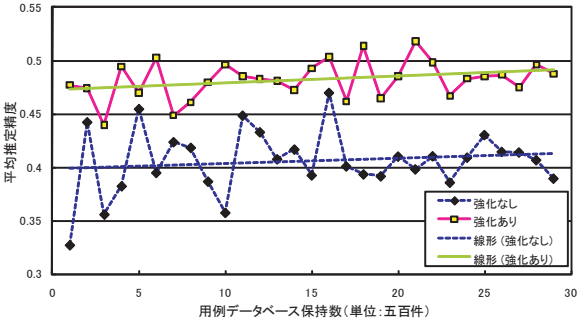

図 14: パターン c におけるコサイン類似度を用いた条件強化，用例データベース毎の平均品詞推定精度

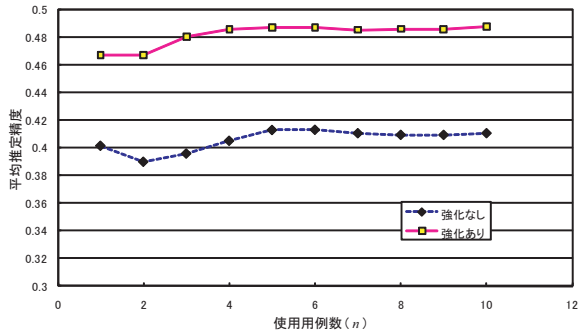

図 15: パターン c におけるコサイン類似度を用いた場合の使用用例数 n による平均品詞推定精度

4 まとめ

今回の実験結果から，以下の知見を得た．

1. 類似度の計算式では，類似要素数を少なくしたほうが効果が高くなり，また計算手法による結果の差異はほとんど無い．
2. 検索条件の強化は，品詞推定精度の向上という点では効果は小さいが，検索用例数を少なくするため，処理速度の向上には有効である．
3. 用例データベースの規模の拡大は品詞推定精度の向上に有効ではあるが，処理速度を考慮すると，対話システムにおいては有効ではない．
4. 本提案においては，コーパスに依存した品詞推定精度の差異は小さい．
5. 類似度の大きい少数の類似用例を使用することが有効であるが，最適な数値はその他の要素により変化するため一定ではない．
6. 表層文字列のみに依存した類似度では未知語の品詞推定には有効ではなく，類似用例が多く残ってしまう．

5 おわりに

本稿では，品詞並びの検索条件を強化した場合，用例データベースの規模を拡大した場合，用例データベースの作成に異なるコーパスを使用した場合，使用用例数を変化させた場合の品詞推定精度の比較を行った．

今後は表層文字列以外に基づく類似度の指標を用いる手法についても研究を行っていく．

参考文献

[1] 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 (2010/10).

[2] 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).

[3] 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（研究上の都合により，原著者の了解を得てテキストデータの加工を行った）．

[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).