Vol.2012-SLP-94 No.25
2012/12/21

トピック関連語推定とSTDによる未知語推定の評価

佐藤 壮一[†1]　伊藤 彰則[†1]

概要：本稿では，音声認識結果から関連する単語を推定するトピック関連語推定と，発話中にある単語が含まれているかどうかを見る検索語検出（Spoken Term Detection:STD）を用いて，音声認識における未知語を推定した．トピック関連語推定のみを用いた場合，STD のみを用いた場合，両方を用いた場合について，それぞれ比較し検討を行った．その結果，両方を用いた場合に推定語数が多い状況で，トピック関連語推定のみの場合に推定語数が少ない状況で最も良い再現率を得られることがわかった．また，トピック関連語推定の再現率が高い状態で STD を利用することで，トピック関連語推定のみの場合よりも高い適合率を得ることができることもわかった．

1. はじめに

近年，我々はインターネットの発展と共に，テキスト以外のマルチメディアコンテンツを利用する機会が増えている．しかし，音声を含む多くのコンテンツは，タイトルや説明文，タグなどといった限られたメタデータしか与えられておらず，検索やカテゴリ分けが困難とされている．これらに対して，最近実用化が進んでいる大語彙連続音声認識（ディクテーション）システムへの期待が高まっている．音声を含むコンテンツに対してディクテーションを行うことで，インデックスを自動付与することが可能になる．

しかし，現在の大語彙連続音声認識システムは 2 つの問題を抱えている．まずは認識誤りの問題である．音声認識結果には多くの認識誤りが存在する．そのため，音声にとって重要な単語の欠落や置換が起こりうる．もう 1 つの問題として語彙にない単語である未知語（out of vocabulary : OOV）の問題が挙げられる．今日の音声認識器の語彙の大きさは有限であるにもかかわらず，未知語は必ず認識できないという問題である．したがって，現在の音声認識では正確なインデックスの付与が難しい．

本研究では未知語の問題に着目している．この問題に対して Web 上の言語資源を用いて，未知語となる単語を補完することが注目されている．先行研究としては，認識結果から得られた単語を Web 検索のクエリとすること [1] や講義で使用されたスライドに含まれる特徴的な名詞をクエリとすること [2] が例として挙げられる．このような方法での未知語の獲得に関しては，Web から取得した言語データに偶然未知語が含まれることを期待するのみで含まれる未知語を積極的に推定する試みは行われてこなかった．

そこで，本稿では音声認識結果から話題に関連する単語を推定するトピック関連語推定 [3–5] や，発話の中にある単語が含まれているかどうかを見る検索語検出（Spoken Term Detection : STD）を用いて，形態素解析器の辞書に存在し，認識器の辞書に存在しない単語（本研究では未知語候補と呼ぶ）の中から未知語推定を行い，推定された未知語の再現率，適合率と推定語彙数に関して検討を行う．

2. 未知語推定概要

未知語推定の流れは以下の形をとる (**図 1**).

step1　音声認識器辞書と形態素解析器辞書を比較し未知語候補を作成.

step2　未知語候補のそれぞれの単語について，Web 検索結果から $tfidf$ を用いて文書ベクトルを生成.

step3　音声認識結果から $tfidf$ を用いた文書ベクトルを生成.

step4　**step2**，**step3** の文書ベクトルのコサイン類似度を求め，しきい値処理.

step5　音素認識結果と **step4** で得られた単語の音素列で，記号系列間の連続 DP を用いた STD を行い，しきい値処理.

これらの処理を行って，条件を満たすものを未知語推定結果とする.

なお，トピック関連語推定のみを用いた場合（**図 2**）と STD のみを用いた場合（**図 3**）は，それぞれ **step 5**，**step2** から **step4** の処理を抜かしたものとなっている.

[†1] 現在，東北大学大学院工学研究科
Presently with Graduate School of Engineering, Tohoku University

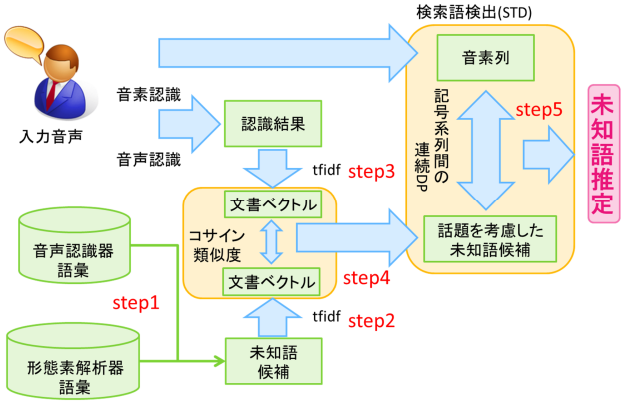

図 1 未知語推定概要
Fig. 1 Outline of OOV presumption

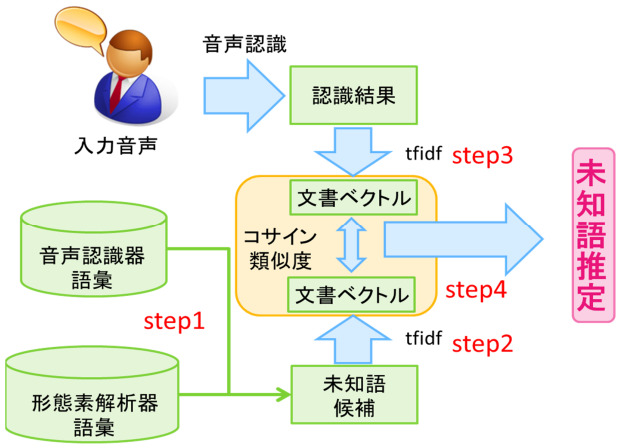

図 2 トピック関連語推定のみの場合
Fig. 2 Estimated topic related words only

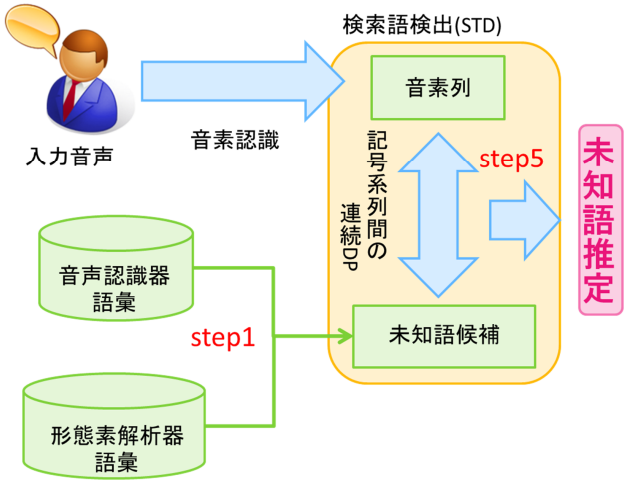

図 3 STD のみの場合
Fig. 3 STD only

3. トピック関連語推定

ある単語がどの程度発話のトピックに関連するかに関するスコアを推定する処理をトピック関連語推定と呼ぶ．このアプローチには一般的に $tfidf$ が用いられる．これにより文書の内容を表す要素である索引語（index term）がその文書及び検索要求の内容にどれだけ密接に関連しているかを索引語の重要度として索引語に付与することができる．

3.1 tfidf

ある文書 D 中に出現する索引語 w の頻度を索引語頻度（term frequency:tf）と呼び，$tf_D(w)$ で表す．これは同じ文書に多く表れる単語ほど検索の有力な手掛かりになりうることを示している．しかし，一般に頻度の高い単語は，文書を特徴づける上では役に立たないことが多い．そこで，索引語が全文書中のどの程度の文書に出現するかを表す尺度を逆文書出現頻度（inverse document frequency:idf）と呼び，

$$idf(w) = \log \frac{N}{df(w)} \tag{1}$$

で表す．ここで，N は検索対象となる文書集合中の全文書数，$df(w)$ は索引語 w が出現する文書数である．式（1）から，idf はある索引語が少数の文書にしか出現しない場合に大きくなり，どの文書にも出現する場合に最小の値を取る．N と $df(w)$ の比の対数を取るのは文書集合の規模に対して idf の値の変化を少なくするためである．つまり，こちらはさまざまな文書に現れる単語はあまり検索の有力な手掛かりになりえないことを表している．

これら 2 つの尺度を組み合わせて索引語の重みを計算することができ，式（2）のように求める．

$$tfidf_D(w) = tf_D(w) \cdot idf(w) = tf_D(w) \cdot \log \frac{N}{df(w)} \tag{2}$$

本研究では，Web 上の全文書を文書集合として $df(w)$ を求めるため，$df(w)$ は Web 検索ヒット数を利用する．また，N は Web 検索でヒット可能な Web ページ数であり，この値は求められないため，適当な定数を与えて計算を行った．

3.2 ベクトル空間モデル

文書を索引語の重みベクトルとして表現したものを文書ベクトル（document vector）と呼ぶ．各単語の重みに $tfidf$ を用いることで，ある文書 D の文書ベクトル $\boldsymbol{I}(D)$ は式（3）で表される．

$$I(D) = [tfidf_D(w_1), \cdots, tfidf_D(w_k)]^{\mathrm{T}} \tag{3}$$

このように，検索質問と文書を文書ベクトルで表現することで，検索質問に対する文書の適合度をベクトル間の類似度によって計算することができる．これをベクトル空間モデル（vector space model）と呼ぶ．

文書ベクトル間の類似度をコサイン類似度を使って求める．文書 D_1 の文書ベクトル $\boldsymbol{I}(D_1)$ と文書 D_2 の文書ベクトル $\boldsymbol{I}(D_2)$ 間のコサイン類似度は式（4）で求める．

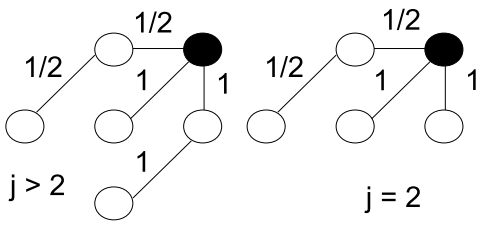

図 4 DP パスのアルゴリズム
Fig. 4 Algorithm of DP

$$sim(D_1, D_2) = \frac{\boldsymbol{I}(D_1)^{\mathrm{T}}\boldsymbol{I}(D_2)}{|\boldsymbol{I}(D_1)||\boldsymbol{I}(D_2)|} \tag{4}$$

これは両ベクトルの張る角度を表しているので，値が 1 の場合に両ベクトル，すなわち両文書が最も類似している．また，ベクトルの類似度を余弦で計算することによってベクトルのノルムは無視される．この性質が一概に良いとは言えないが，ベクトル空間でコサイン類似度を用いるとすぐれた検索性能が得られることが知られている．ベクトル空間モデルによる情報検索では，文書全体の類似性を評価できるため，音声認識における認識誤りが存在しても，ある程度の精度で関連文書検索が可能である．

3.3 単語 - 文書間の話題の関連性

単語と音声認識結果の話題の関連性を確認するためには，未知語候補 Q_i に関連する文書 D_{Q_i} が必要である．そこで，本研究では Web 検索を利用し，単語 Q_i をクエリとしたときに得られた文書を，関連する文書 D_{Q_i} とし，文書ベクトル $\boldsymbol{I}(D_{Q_i})$ をあらかじめ作成した [6]．音声認識結果 D_{SR} の文書ベクトル $\boldsymbol{I}(D_{SR})$ と未知語候補 Q_i 毎の文書ベクトル $\boldsymbol{I}(D_{Q_i})$ のコサイン類似度を式（5）で求めた．

$$sim(D_{SR}, D_{Q_i}) = \frac{\boldsymbol{I}(D_{SR})^{\mathrm{T}}\boldsymbol{I}(D_{Q_i})}{|\boldsymbol{I}(D_{SR})||\boldsymbol{I}(D_{Q_i})|} \tag{5}$$

このスコアに対してしきい値処理を行い，条件に合った単語をトピック関連語推定結果とする．

4. 検索語検出（STD）

検索語検出は音素記号系列における連続 DP を用いた検索方法を利用した [7]．連続 DP の計算には**図 4** に示す DP パスを用いた．

DP コストを求める手順は以下のとおりである．

(1) 初期条件 $j = 1, 2, \cdots, J$ について

$$g(0, j) = g(-1, j) = -\infty \tag{6}$$

(2) $i = 1, 2, \cdots, I$ について 3〜5 を実行

(3) $g(i, 1) = P_s(i, 1)$

$$g(i,2) = \max \begin{cases} g(i-2,1) \\ \quad + \frac{P_s(i,2)+P_a(i-1,2)}{2} \\ g(i-1,1) + P_s(i,2) \\ P_s(i,2) + P_o(i,1) \end{cases} \tag{7}$$

(4) $j = 3, 4, \cdots, J$ について

$$g(i,j) = \max \begin{cases} g(i-2,j-1) \\ \quad + \frac{P_s(i,j)+P_a(i-1,j)}{2} \\ g(i-1,j-1) + P_s(i,j) \\ g(i-1,j-2) \\ \quad + P_s(i,j) + P_o(i,j-1) \end{cases} \tag{8}$$

(5) $d(i) = g(i, J)/J$

(6) $\max\{d(i)\} > \theta$ であるとき STD の結果とする．

ここで，$g(i, j)$ は状態量，I, J はそれぞれ入力列の全セグメント数，標準パターンとなる音素系列の全セグメント数，θ はしきい値とする．また，$P_a(i, j)$ は標準パターン側の $j-1, j$ に対応する音素間に，i に対応する音素が付加する確率の対数値，$P_o(i, j)$ は入力系列の $i-1, i$ に対応する音素間で，j に対応する音素が脱落する確率の対数値，$P_s(i, j)$ は j に対応する音素が i に対応する音素に置換する確率の対数値とする．各音素が挿入・脱落・置換する確率の推定は言語モデル学習データと同様の CSJ データベースの 2537 講演を利用している．

5. 実験条件

5.1 未知語候補

今回の実験で用いた音声認識システムの辞書は名詞約 4 万 8 千語で，形態素解析器の辞書（ipadic [8] と unidic [9] を統合して作成）は名詞約 32 万 2 千語であった．これらを比較し，得た未知語候補は名詞のみで約 29 万 5 千語であった．但し，複数読みを許した場合の未知語候補は約 32 万 7 千語となる．しかし，この中には Web 検索にふさわしくない単語も含まれる．例えば「すね」という単語が挙げられる．これは「そうですね」が「そう」「で」「すね」と分かれ，「すね」が名詞の「すね」（脚のすね）と誤って解析される可能性があるため検索にふさわしくないといえる．このような単語を除き，実際に未知語推定に用いた未知語候補は，Web 検索クエリと同様の約 28 万 7 千語である．

5.2 音素認識システム・音声認識システム

音素認識・音声認識のデコーダは Julius rev.4.2 を使用した．言語モデルは CSJ データベース [10] の 2537 講演から学習したものを利用した．音響モデルの詳細を**表 1** に示す．また，言語モデルの詳細を**表 2** に，音素・音声認識における言語モデルのエントリ数を**表 3** に示す．

学習セットを含まない CSJ データベースの 40 講演による音素認識結果を**表 4** に，音声認識結果を**表 5** に示す．表 4，表 5 における Acc，Sub，Del，Ins はそれぞれ，音

表 1 音響モデルの詳細
Table 1 Detail of acoustic model

音響モデル	混合連続分布 HMM（triphone モデル）
サンプリング周波数	16kHz
プリエンファシス	0.97
分析窓	Hamming 窓
分析窓長	25ms
窓間隔	10ms
特徴パラメータ	MFCC +Δ MFCC +Δ パワー（計 25 次）
周波数分析	等メル間隔フィルタバンク
フィルタバンク	24 チャンネル

表 2 言語モデルの詳細
Table 2 Detail of language model

言語モデル	音素 N-gram・単語 N-gram
バックオフスムージング	Witten-Bell discounting
語彙に含まない単語の扱い	OOV word = UNK

表 3 言語モデルのエントリ数
Table 3 Number of entries in language mode

	音素認識	音声認識
1-gram	44	40,686
2-gram	849	806,612
3-gram	15,342	2,436,584
4-gram	129,781	—
5-gram	608,658	—
6-gram	1,815,885	—

素・単語正解精度，置換誤り率，脱落誤り率，挿入誤り率である．

表 4 音素認識結果（%）
Table 4 Results of phoneme recognition（%）

Acc	Sub	Del	Ins
58.1	13.8	24.6	3.5

表 5 音声認識結果（%）
Table 5 Results of speech recognition（%）

Acc	Sub	Del	Ins
58.1	13.8	24.6	3.5

表 4 にある認識率で STD を行い，表 5 にある認識率でトピック関連語推定を行う．両者においてあまり高くない認識率であるため，スコア付けはあまり正確に行われない．

5.3 Web データ

Web 検索のクエリも未知語候補同様に ipadic と unidic を統合して作成した形態素解析器の辞書の名詞を利用した．その中でも**表 6** の chasen の辞書に基づく形態素のみをクエリになり得る単語としている．ただし，検索に悪影響を及ぼすと予想される単語をストップワード（stopword）とすることで，文書の内容を特徴付けるクエリのみを抽出した．ストップワードの例を**表 7** に示す．その結果，クエリの対象となった単語は名詞約 28 万 7 千語であり，クエリ毎のダウンロード URL 数は 50URL とした．ダウンロードを行った期間は 2010 年 2 月から同年 4 月までの約 3 ヵ月間である．Web 検索には Yahoo!JAPAN の提供する Web API を利用した [11].

表 6 クエリとして利用する品詞
Table 6 POS used as queries

名詞-一般，名詞-固有名詞，名詞-固有名詞-一般， 名詞-固有名詞-人名，名詞-固有名詞-人名-姓， 名詞-固有名詞-人名-名，名詞-固有名詞-組織， 名詞-固有名詞-地域，名詞-固有名詞-地域-一般， 名詞-固有名詞-地域-国，名詞-副詞可能， 名詞-形容動詞語幹，名詞-サ変接続

表 7 ストップワードの例
Table 7 Example of stopword

ぁ，ぃ，ぅ，あ，い，う，ア，イ，ウ， 一回，十回，一日，十日，十分， から，ところ，かく，した，ます，かしら， 普通，最初，きょう，結局，今度，全体， 主，上，以上，以下，方，殆ど，例

Web 文書は，一般的に HTML で書かれているため，Web 文書を整形して利用した．具体的には，HTML などのタグを消去した後に，行ごとに以下のルールを満たすかどうかで Web 文書を整形する．

- 句点（！，？，♪を含む）で終了する．
- アルファベット，数字，記号の割合が 20%以下．
- 行の長さが 10 文字以上．

以後，Web データはこのような整形を行った後のテキスト文書である．

6. 実験結果と考察

テストデータとして言語モデルの学習データを含まない 40 講演で未知語推定実験を行った．評価には推定語数と，以下の再現率（recall）・適合率（precision）を利用した．

$$recall = \frac{\text{正しく推定した未知語数}}{\text{未知語数}} \tag{9}$$

$$precision = \frac{\text{正しく推定した未知語数}}{\text{推定語数}} \tag{10}$$

図 5 に推定語数と再現率の関係のグラフを示す．トピック関連語推定における再現率がそれぞれ最大で約 97%，92%，87%，82% となるようしきい値を固定し，STD のしきい値を変化させた．トピック関連語推定においても STD においても，誤認識が多い状況下でもある程度の未知語推定が可能であることが見てとれる．さらにそれらを組み合わせる

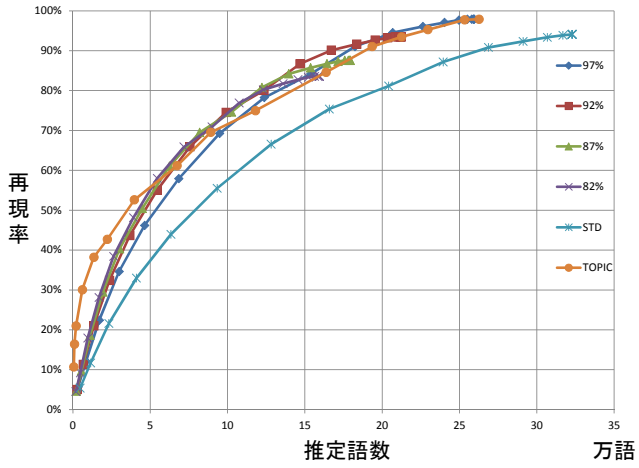

図 5 推定語数と再現率の関係

Fig. 5 Relationship between number of estimated words and recall

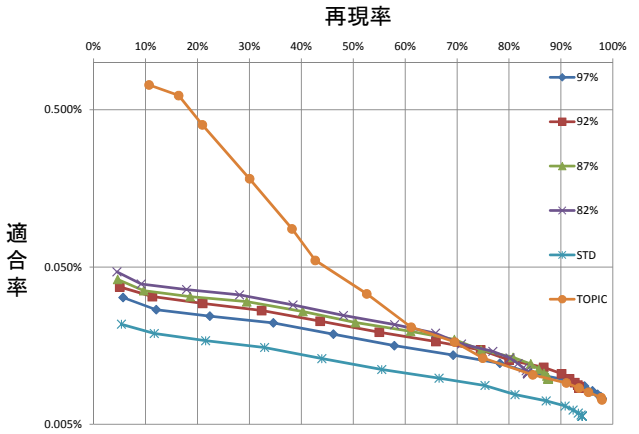

図 6 再現率と適合率の関係

Fig. 6 Relationship between recall and precision

ことによって，推定語数が多い場合はトピック関連語推定と STD を組み合わせる方が高い再現率を得ることができたが，推定語数が少ない場合はトピック関連語推定のみの方が良い結果であった．これは誤認識を多く含む音素認識結果を用いた STD では，発話中に含まれている単語を正しく推定できていないことが理由として挙げられる．よって厳しすぎる STD の制約は未知語推定の妨げになることがわかった．

また，**図 6** に再現率と適合率の関係のグラフを示す．こちらも同様にトピック関連語推定における再現率がそれぞれ最大で約 97%，92%，87%，82% となるようしきい値を固定し，STD のしきい値を変化させた．

この結果から，トピック関連語推定において高い再現率のしきい値で固定し，STD のしきい値を変化させた場合，STD の制約が緩い条件でトピック関連語推定のみの場合よりも適合率の改善が見られた．本稿では高い再現率の場合のみしか検討を行わなかったが，低い再現率でも同様の傾向が得られ，発話における重要な未知語の推定が可能なのであれば，未知語推定を用いた言語モデル適応にこの手法が利用できる可能性がある．

7. まとめと今後の課題

本稿では，言語的情報に音響的情報を加えることにより，未知語推定の精度向上を期待し，トピック関連語推定と STD を用いた未知語推定を行い，トピック関連語推定のみの場合や STD のみを用いた場合と比較し，再現率と推定語数について検討した．未知語推定にある程度の効果が期待できる STD をトピック関連語推定に組み合わせることで，推定語数が多い場合に精度が上がることがわかった．しかし，誤認識を多く含む音素認識結果を STD に利用しているため，STD による制限を大きくすることで推定精度の低下が見られた．

また，トピック関連語推定と STD を組み合わせることによって，トピック関連語推定の再現率が高い場合，トピック関連語推定のみの場合よりも適合率を改善できることがわかった．同様の傾向がトピック関連語推定の再現率が低い場合でも得られるか確認することが今後の課題ある．さらに，推定語数が十数万語と多いため，音声認識結果からクエリを構成し Web 検索を行い，検索結果から得られた単語を未知語候補とする手法で未知語推定実験を行う予定である．

参考文献

[1] 増村亮，咸聖俊，伊藤彰則：Web 上の言語資源を利用した大規模話し言葉データからの言語モデル作成，音講論 (春)，pp75-78(2011).
[2] 根本雄介，秋田祐哉，河原達也：講義音声認識のためのスライド情報を用いた言語モデル適応，言語処理学会年次大発表論文集，pp131-134(2007).
[3] T.Misu, *et al.*: A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts, In Proc.Interspeech, pp.9-12(2006).
[4] A.Ito, *et al.*: Unsupervised language model adaptation based on keyword clustering and query availability estimation, In Proc.Conf.on Audio, Language and Image Processing, pp1412-1418(2008).
[5] 増村亮，伊藤仁，伊藤彰則，牧野正三：Web 検索結果を利用したトピック関連語推定に基づく言語モデルの教師なし適応，音講論 (春)，pp57-58(2010).
[6] R.Masumura, S. Hahm and A.Ito : Language Model Expansion Using Webdata for Spoken Document Retrieval, In Proc.Interspeech, pp.2133-2136(2011).
[7] 岡田他：構文駆動型連続 DP 法による連続音声中からの活用語スポッティング，電子情報通信学会論文誌，Vol.J70-D, pp.2479-2490(1987).
[8] ipadic-2.7.0, 入手先 ⟨http://chasen.aist-nara.ac.jp/stable/ipadic/⟩.
[9] 形態素解析辞書 UniDic, 入手先 ⟨http://www.tokuteicorpus.jp/dist/⟩.
[10] K.Maekawa *et al.*: Spontaneous speech corpus of Japanese, In proc.LREC, pp.947-952(2000).
[11] Yahoo! developer's network，入手先 ⟨http://developer.yahoo.com/⟩.