Vol.2009-SLP-77 No.1
2009/7/17

集合知を利用した語彙情報サービスにおける自動語彙拡張の評価

佐々木 浩† 中野 鐵兵† 藤江 真也† 小林 哲則†
†早稲田大学

あらまし:
音声・言語アプリケーションにおける従来の語彙情報作成手法の問題点を解決するため，集合知を利用した語彙情報の収集・共有・管理システムを開発し，語彙情報を集中管理するためのオンラインデータベースシステムとして公開した．本システムでは，Web 資源からの語彙情報の自動収集の枠組みを備え，データの集約を図っている．そしてインタフェースを広く公開し，アプリケーション間の語彙定義の共有や，アプリケーションで使用する語彙の自動更新のサポートを図っている．本稿では，本システムの概要と応用例の紹介し，自動更新された語彙の評価を行う．その後，今後の応用や課題について検討をする．

Evaluation of Automatic Vocabrary Acquisition of Lexical Data Service using Collective Intelligence

Hiroshi SASAKI†, Teppei NAKANO†, Shinya FUJIE†, Tetsunori KOBAYASHI†
†Waseda University

Abstract:
In order to solve the problems of the conventional approach of designing lexicons, we developed the lexical data collection, sharing, and management system using collective intelligence. The system is designed as a data intensive system so that it can collect lexical information from all web-based resources. Also, the system interface is published so that lexical information are shared by many applications. In this paper, we describe an outline and examples of the application of this system and evaluate an accuracy of automatically acquired vocabrary using this system. Finally we examine a future problem and application.

1 はじめに

音声・言語アプリケーションで利用される語彙情報の整備をサポートするために開発されたウェブシステムの評価を行う．

音声認識アプリケーション開発における最も重要かつ困難な作業の一つとして，継続的な語彙情報のメンテナンスが挙げられる．アプリケーション用に設計した語彙は一度設計したら完成というものではなく，新規語彙の追加や既存語彙の修正などの持続的なメンテナンスが必要である．また，メンテナンスされた語彙情報を適切にアプリケーションへ適用する必要もある．これらを実現するため，語彙情報をサーバ上で分散管理し，ネットワークを通じてアプリケーションに配信するような枠組みが求められる．

そこで我々は，アプリケーションに用いる語彙情報作成の負荷を低減するため，“語彙情報サービス”を開発した [1][2]．このサービスは Web ベースのオンラインデータベースシステムとしてインターネット上に公開され，簡単な要求を投げるだけで必要な語彙情報を得ることができる．また，クローラによる Web 資源からの自動収集の枠組みと，利用者の集合知を利用した半自動的な語彙情報作成の枠組みによって，自由に利用可能な形式で語彙情報が集約され，日々データベースが増強されている．さらに，データベース上の語彙の情報が更新されたり，新規の語彙が追加された際に，それらの情報を利用者やアプリケーションへ反映する機構も用意されている．この枠組みを Proxy-Agent[3] と呼ぶ音声認識システム拡張の枠組みと組み合わせ，語彙情報の動的な更新を可能にする音声認識アプリケーション開発の新しい枠組みの実現を目指している．

こうした自動的な新規語彙獲得の枠組みをアプリケーションに適用する際，獲得語彙の高い充足性と正確性が求められる．あらゆる分野の語彙の要求に対し，システムがいかに恒常的に適切な新規語彙を供給し続けられるかがシステムの性能を大きく左右する．そこで本研究では，“語彙情報サービス”でいくつかの分野の語彙リストを生成し，それぞれの生成後の新規語彙獲得の状況を検証した．各条件でどれだけ適切な新規語彙を獲得できるかを評価する．

本稿では，次節で “語彙情報サービス” の概要について述べる．次に，3 節で本サービスの具体的な利用法と得られる語彙情報の例，自動的に増強される語彙情報の例を挙げ，その評価をする．4 節で，本サービスの実アプリケーションへの応用の紹介と今後の応用について検討する．

2 語彙情報サービスの概要

“語彙情報サービス” はオンラインサービスとして公開され，以下の特徴を持つ．

Data Intensive Systems 音声・言語アプリケーションで必要な語彙情報が集約され，単一の語の読

表 1: 語彙整備に用いている情報源 (タグ情報: タグに使用した情報．種類は以下の通り：(A) 話題性の高い語の読み, (B) 標準語彙の読み, (C) 地名の読み, (D) 飲食店名の読み, (E) 宿名と温泉名の読み, (F) 曲名の読み (G) 人名（ミュージシャン，タレント）の読み, (H) 経済・IT 用語の読み (I) 医学用語 (J) 英単語の読み）

情報源	種類	タグ情報	情報源	種類	タグ情報
Wikipedia	(A)	カテゴリ	歌詞タイム[12]	(F)	アーティスト名
はてなキーワード API[3]	(A)	カテゴリ	DMM.com[13]	(G)	人名
Yahoo!辞書 - 新語探検[4]	(A)	カテゴリ	人名録 KEY PERSON[14]	(G)	人名
イザ語[5]	(A)	カテゴリ	生年月日データベース[15]	(G)	人名
FC2 キーワード[6]	(A)	カテゴリ	三菱電機 EPG データ	(G)	人名
ニコニコ大百科[7]	(A)		ASCII.jp[16]	(H)	
ipadic version 2.7.0	(B)	品詞	経済新語辞典[17]	(H)	経済
Yahoo!百科事典[8]	(B)	カテゴリ	iFinance[18]	(H)	カテゴリ
郵便番号データベース	(C)	市区町村 都道府県	e-Words[19]	(H)	カテゴリ
ホットペッパー Web サービス[9]	(D)	所在地 ジャンル	音訳の部屋[20] 医学用語の読み方	(I)	医学
ぐるなび Web サービス[10]	(D)	所在地と カテゴリ	カタカナ英単語[21]	(J)	
じゃらん Web サービス[11]	(E)	所在地			

み情報の取得から，アプリケーション用語彙の作成・管理まで，語彙情報に関連する全ての作業を提案システムで完結できるようになっている．またシステムが広く公開されることにより，自然と語彙に関連する情報が集約されるような仕組みを持っている．

Lexicon Lifecycle　アプリケーション用の語彙の新規作成から，その継続的な更新まで包括的な解法を提供する．

Cooperative Framework　語彙情報を必要とするアプリケーション同士のゆるやかな連携を可能にする．すなわち，アプリケーションで使用する語彙の定義と追加・修正された語彙情報の共有を可能にする．

2.1　語彙情報の集約

データベースは語の綴りの情報，読みの情報，収集元の情報を保持する．読みや収集元の情報は語の綴りの情報からの関連情報という形で保持している．語彙に対してのメタ情報として自由なタグ付けを許し，そのタグの情報も語単位で保持している．例えば，“早稲田大学” という語にはタグとして “名詞” や ‘大学” などが登録されている．タグを付与する際には，タグも語として登録し，語と語との関連としてタグを定義している．さらに，タグには役割を表す語を関連付けられている．例えば品詞を表すタグには “品詞” の語への関連を持つ．こうした情報を初期の段階で十分に確保するため，例えば ipadic[1] や Wikipedia[2] などの WWW 上で利用可能な語彙資源

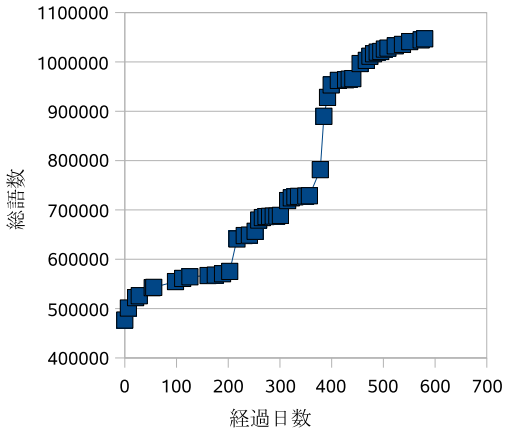

図 1: 総語数の遷移（急激に増加している点は新たな情報源を追加した時点．最後に情報源を追加した後で毎日 200 語以上増加している．）

[1] ipadic version 2.7.0,http://chasen.naist.jp/stable/ipadic/ipadic-2.7.0.tar.gz
[2] Wikipedia,http://ja.wikipedia.org/

(表 1) を活用している．また，システムが情報源を巡回し随時語彙情報を収集している．情報を引用する際には語彙情報の収集元の情報も保持，明記し，権利上の問題に配慮している．加えて，ユーザによる語彙の追加・修正の枠組みも設けている．データベースには 2009 年 6 月 11 日時点で 1,045,319 語が登録されており，毎日 200 語以上の語が新規に登録され続けている (図 1)．

2.2 語彙情報の利用

本サービスは Web アプリケーションとして動作し，Web ブラウザ上または WEB API 経由で利用する．データベースの語彙の利用は，直接語彙情報を参照する方法と事前に語彙リストを定義してその語彙リストを利用する方法がある．

2.2.1 直接語彙情報を参照する方法

情報を取得したい語彙をクエリとして，その語彙の情報を直接参照する (図 2 (上))．ユーザは Web ブラウザまたは WEB API を用いて本サービスに希望の語を送信すると，その語の読みやタグの情報を得ることができる．利用する語の集合が既に分かっており，その語の詳細な情報を得たいというケースに有用である．また，クエリにデータベースに存在しない語が含まれた場合は形態素解析結果などから読みを推定し，何らかの読み情報を返す機能を併せ持つ．こうした情報を蓄積することで，データベースに存在しない，多くの要求がある語をデータベースに追加することを可能にしている．

2.2.2 語彙リストを定義して利用する方法

Web ブラウザ上で語の集合を検索・編集し，その語の集合を語彙リストとして利用する (図 2 (下))．本サービスは語の集合を定義する方法として，各語のタグの付与条件を用いる方法を採用している．

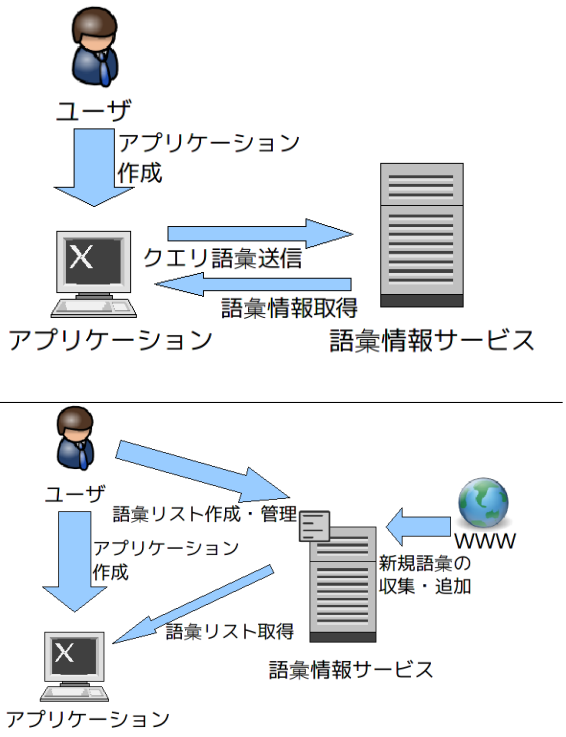

図 2: 語彙情報サービスの利用法．(上) 直接語彙の情報を参照．(下) タグ条件を指定して語彙リストを取得

ユーザは本サービスに対して利用したい語の集合のタグ条件を送信すると，データベース内でのそのタグ条件を満たす語の集合を得ることができる．その結果を基に必要な語の追加や不必要な語の削除を行うと，よりユーザの希望に沿ったタグ条件が推奨される．もし適切なタグ条件が得られなければ，編集した結果を表すようなタグ条件の登録を促す．このようにして，ユーザの希望する語の集合をタグの付与条件という明確な形で定義し，それによって目的の語の集合を利用できるようにする．ユーザはこのタグの付与条件をクエリとして，語彙リストのファイルの形式でのダウンロードや WEB API 経由での参照を行うことができる．利用する語の集合を用意できないケースや語の集合に関して継続的な管理が必要なケースに有用である．

2.3 語彙リストの管理・共有

ユーザの定義した語の集合は本サービスに蓄積され，いつでも再利用することができる上に、Web ブラウザ上で自由に修正・編集ができる．さらに，語の集合を定義するタグ条件を基に，データベースに新しく登録された語 (新着語と呼ぶ) がユーザに通知される仕組みを持つ．ユーザは新着語を確認することで語の集合の新規性の維持が可能となる．また，これらの語の集合はタグ条件として本サービスを利用するユーザ間で広く共有することができる．

[3] はてなキーワード, http://d.hatena.ne.jp/keyword
[4] 新語探検, http://dic.yahoo.co.jp/newword
[5] イザ語, http://www.iza.ne.jp/izaword/
[6] FC2 キーワード, http://keyword.fc2.com/
[7] ニコニコ大百科, http://dic.nicovideo.jp/
[8] Yahoo!百科事典, http://100.yahoo.co.jp/
[9] ホットペッパー, http://www.hotpepper.jp/
[10] ぐるなび, http://www.gnavi.co.jp/
[11] じゃらん, http://www.jalan.net/
[12] 歌詞タイム, http://www.kasi-time.com/
[13] DMM.com, http://www.dmm.com/
[14] 人名録 KEY PERSON, http://www.person.cbr-j.com/
[15] 生年月日データベース, http://www.d4.dion.ne.jp/~warapon/data00/
[16] ASCII.jp, http://ascii.jp/
[17] 経済新語辞典, http://bizplus.nikkei.co.jp/shingo/
[18] iFinance, http://www.ifinance.ne.jp/
[19] e-Words, http://e-words.jp/
[20] 音訳の部屋, http://hiramatu-hifuka.com/onyak/onyindx.html
[21] カタカナ英単語, http://homepage2.nifty.com/katakanaEnglish/

図 3: 語彙情報サービス．（左上）クエリの入力画面.（右上）検索結果表示画面.（左下）クエリの再生成.（右下）ファイル出力.

3 語彙情報サービスの利用例と自動語彙拡張の検証

3.1 語彙情報サービスの利用例

本サービスを用いて実際に語彙リストの作成・更新をする．例として人名以外の映画関連の語彙リストの整備を行なう．

まず，ユーザは検索画面上部のフォームに想定されるタグ条件をクエリとして入力する（図 3（左上））．クエリは and,or,not や括弧によって複数の条件を指定できる．ここでは “映画” とした．クエリの入力後，フォームの下の検索結果出力画面にそのクエリでの検索結果が表示される（図 3（右上））．各語の右隣には語の詳細情報画面へのリンクと修正ページへのリンクがあり，ユーザはここから語彙のタグ情報や情報源へ参照や，1 語単位での修正を行うことができる．ここで，語彙の絞り込みを行うため，より適切なタグ条件を求める．ユーザは検索結果のうち数語の採用・削除指定を行う．検索結果中で適切・不適切だと思われる語に対して，その名前の左にある “採用” “削除” のボタンを押し，採用・削除の指定を行う．システムはその結果を分析してより適切なタグ条件をユーザに提示する．これにより結果の全てを確認しなくても，数語の編集で語彙を絞り込むことができる．今回は人名以外の映画関連の語彙が必要なのでここでは人名を数語削除した．そして，画面上部の “クエリ再生成” を押すとより適切なタグ条件をシステムが推奨する（図 3（左下））．推奨されたタグ条件で再度検索を行うと，人名が除かれた映画関連語の語彙が得られる．条件 “映画” での結果と削除した語彙，“映画 not 人物” での結果を表 2 に示す．

画面上部には結果を登録するフォームがある．ユーザはここにリストの名前を入力して登録すると各形式へ出力するための保存画面が現れる．希望する形式のボタンを押すと，その形式でリストがダウンロードできる（図 3（右下））．出力形式は CSV ファイルや Julius 孤立単語認識用辞書形式などが用意されている．ユーザが作成した語彙リストはデータベースに保持され，語彙リスト管理画面へのリンクが表示される．ユーザはここから過去の語彙リストの作成・管理ができる．WEB API を用いることにより，語彙リストを利用したいアプリケーションから直接本サービスに保持された語彙リストを利用することができる．

語彙リスト作成後，データベースに “映画 not 人物” に該当する語が新しく追加された場合は，追加語彙候補として語彙リスト管理画面に通知がなされる．これによりユーザは過去に作成した語彙リストの更新を効率的に行うことができる．また，WEB API を用いる際は追加語彙候補を含めた語彙リストを利用することもできる．

表 2: “映画” “映画 not 人物” での結果

条件	語	語数	削除した語
‘映画 ‘’	007, 香港国際警察, 市原隼人, ダイ・ハード 2, ペネロペ・クルス 等	18345	市原隼人, ペネロペ・クルス, おすぎ 等
“映画 not 人物”	007, 香港国際警察, ダイ・ハード 2, 華氏９１１ 等	4249	

3.2 タグ条件を用いた自動語彙拡張の検証

3.1 節の語彙の検索には 2009 年 5 月 1 日のデータベースが用いられている．2009 年 6 月 12 日でのデータベースを用いて同じ条件で検索を行い，それらを比較することにより，追加語彙候補の自動拡張の効果を検証する．用いるタグ条件は 3.1 節で用いた “映画 not 人物” の他，コンピュータ関連用語を得るために “コンピュータ” を，人名以外のテレビ番組関連の語を得るために “テレビ番組 not 人物” を，事件に関する語を得るために “事件” を用いる．表 3 に追加語彙候補の例を示す．また，追加語彙候補が適切なものかを評価するため，各条件での追加語彙候補のうち，適切でない語として音声入力に適さない語を人手で判断し，それ以外の語の割合で正確性を評価した．その結果を表 4 に示す．

各条件で比較的新しい語が追加されていることがわかる．また，“映画 not 人物” 以外の条件では 9 割以上の正確性で新規語彙が獲得できていることがわかる．このような新しい語彙を適用していくことにより，ユーザは作成したアプリケーションで用いる語彙リストの新規性の維持を効率的に行うことができる．不正解語を発生させた原因については次節で述べる．

3.3 タグ条件を用いた自動語彙拡張の現状の課題

新規語彙のうち，充分にタグが付与されていないものが存在する．そうした語彙はタグ条件にマッチしないため，追加することができない．例えば，データベースには “映画” のタグを持っていない映画関連用語が追加されている可能性がある．こうした語は “映画” のタグ条件にマッチしないため，表 4 の “映画 not 人物” の条件で追加することができない．また，not 条件にマッチしなければ，語彙リストの正確性が低下する．例えば新しく追加された人名に “人物” というタグが付与されなければ，“not 人物” の語彙リストに人名が含まれてしまう．これは表 4 において “映画 not 人物” “テレビ番組 not 人物” の不正解語を発生させてしまった大きな原因の 1 つである．“長澤まさみ” などの語は “人物” のタグが付与されていなかったため，“映画 not 人物” のリストに含まれてしまった．こうした人名が多く含まれてしまったため，“映画 not 人物” での正解率は低くなった．さらに，除外したい語にのみ含まれるようなタグが存在しない場合も不正解語を発生させてしまう．例えば表 4 において “コンピュータ” の結果にある “計画法” などの広い意味を持つ語や “事故” の結果にある “日本の航空事故” などの音声入力に適さない語は今回は除外対象となるが，そうした語のみに含まれるタグが存在しないため，除外できるようなタグの not 条件を作ることができなかった．加えて，1 つの情報源のみからしか得られないような利用頻度の低い語に関しては，長い時間が経過したとしても充分なタグが付与されないため，正確な語彙リストを作成することができない．表 4 において “コンピュータ” の条件に “MS09-018”[22] という記号列が含まれていた．こうした記号列は音声入力に適さないため今回の除外対象となるが，“コンピュータ” 以外のタグが付与されていないため，こうした語を省きたい場合にその条件を作ることができない．これらの問題に対処するため，タグ情報を自動的に補完していくような仕組みも検討していく必要がある．

4 語彙情報サービスの応用

本サービスの適用例として，2.2.1 節の直接語彙情報を参照する方法を用いたアプリケーションを紹介する．音声による項目選択を利用した Web ブラウザ [4] において，Web サイトの項目に用いられる語彙を音声認識させるため，その読みの情報を本サービスを用いて取得し，音声認識に用いている．新規性が高く，分野も特定されない語彙の読みを適切に取得するために本サービス活用している．また，音声コマンドの操作を用いた車載情報端末 [5] において，地名や施設名などの音声認識辞書を構築する際にも本サービスを用いている．都道府県名＞市区町村名＞地域名などの階層構造を持つ項目を構築するために本サービスのタグ情報も用いている．さらに，音声コンテンツのメタ情報のトピックを推定し，そのトピックに沿ったコーパスを選択肢し，音声認識用の言語モデルの適応を行う手法 [6] を用いる際にも本サービスのタグ情報を用いている．音声コンテンツのメタ情報やコーパスのテキスト情報を増やすため，それらに含まれる語彙のタグ情報を本サービスを用いて抽出し，活用した．これらの利用法に関し，本サービスの語彙が効果的に活用されていることを確認している．

今後は 2.2.2 節の語彙リストを作成する方法を応用したアプリケーションについても検討していく．

[22] Windows のセキュリティパッチのコード番号の 1 つであるが，これははてなキーワードからのみ収集されている．

表 3: 各条件での語彙の増加

条件	語	語数 (2009/5/1)	語数 (2009/6/12)	追加された語彙
“映画 not 人物”	007, 香港国際警察, ダイ・ハード 2, 華氏９１１ 等	4249	4629	サブウェイ 123, ポー川のひかり, Last Blood, 携帯彼氏 等
“コンピュータ”	電子マネー, 電子掲示板, P2P,Ubuntu 等	1918	2073	GENO ウィルス, 牧場系サイト, iPhone 3G S,MS08-070 等
“テレビ番組 not 人物”	サザエさん, 必殺仕事人 2007, MR.BRAIN 2, ドクターフー 等	2404	2448	バスカッシュ, 水曜シアター 9, 不毛地帯,7 万人探偵ニトベ 等
“事件”	よど号ハイジャック事件, コロンバイン高校銃撃事件 等	1836	1873	中央大学教授刺殺事件, 足利事件 等

表 4: 各条件での追加語彙正確性

条件	追加語数	正解語数	不正解語数	正解率	不正解語の例
“映画 not 人物”	380	219	161	0.576	長澤まさみ, ジョエル・マクレイ 等
“コンピュータ”	155	146	9	0.942	NOT FOUND, 計画法 等
“テレビ番組 not 人物”	44	42	2	0.955	壱岐正, シェリル・ノーム
“事件”	43	42	1	0.977	日本の航空事故

その例として音声での Web 検索システムが考えられる．音声を用いた Web 検索システムの開発において，広範囲かつ新規性の高い語を認識できることが求められる．しかし，あらゆる分野の語を網羅的に用いた認識辞書を用いると認識率が低下してしまう問題がある．そこで，例えば入力語の分野を指定し，その分野の語彙リストを本サービスで動的に生成し，その語彙リストを認識辞書として用い，認識対象を限定した音声認識を行うといった仕組みを導入することで問題の解決を図る．今回行った評価により，分野を指定して適切かつ新規性の高い語彙リストを本サービスを用いて生成・維持できることを確認したため，こうした音声での自由な Web 検索を実用レベルで可能にすることが期待できる．

5 まとめと今後の予定

音声・言語アプリケーションにおける語彙情報作成手法の問題点の解決を目指すオンラインデータベースサービスの紹介を行った．そして，その機能の１つである自動語彙拡張の評価を行い，新規性の高い語が適切に獲得できていることを確認した．今後は応用アプリケーションや利用者の増加を目指し，アプリケーションを越えた語彙情報の管理・共有の効果をより発揮させていくことを目指す．また，本サービスのデータベースの中の語彙には，現状では読み情報やタグ情報を持っていないものも少なくない．そうした語彙を適切に利用できるようにするため，読み情報やタグ情報を自動的に補完していくような仕組みも検討していく必要がある．

謝辞 本研究は，早稲田大学理工学研究所・プロジェクト研究「音声認識基盤技術」の一部として実施されたものである．

参考文献

[1] 中野 鐵兵, 佐々木 浩, 藤江 真也, 小林 哲則, “WWW を用いた語彙情報の収集・共有・管理システム,” 情報処理学会音声言語情報処理研究会, SIG-SLP-71-12, May 2008.

[2] 佐々木浩, 中野鐵兵，藤江真也，小林哲則，“音声認識アプリケーション開発のための語彙情報サービス，” 日本音響学会秋季研究発表会講演論文集, 2008.

[3] Teppei Nakano, Shinya Fujie, and Tetsunori Kobayashi. EXTENSIBLE SPEECH RECOGNITION SYSTEM USING PROXY-AGENT. Proc. of ASRU2007, pp.601-606, December 2007.

[4] 秋元啓孝，中野鐵兵，小林哲則，“音声による Web リンク選択インタフェースの検討,” 情報処理学会全国大会講演論文集，2009.

[5] Teppei Nakano, Tomoyuki Kumai, Tetsunori Kobayashi, Yasushi Ishikawa, “Design and Formulation for Speech Interface Based on Flexible Shortcuts,” Proc. Interspeech 2008, pp.2474-2477, Sept. 2008.

[6] 佐々木 浩，中野 鐵兵，緒方 淳，後藤 真孝，小林 哲則，“集合知に基づく語彙情報を用いたトピック依存言語モデリング，” 情処研報, SIG-SLP-075，pp.57-62，Feb. 2009.