Vol.2009-SLP-79 No.33
2009/12/22

Google N-gram を用いた 音声認識のタスク汎用性評価の試み

久保慶伍[†1] 三宅純平[†1,*1] 川波弘道[†1]
猿渡洋[†1] 鹿野清宏[†1]

近年，多様な発話に対応可能な音声対話システムの研究が行われている．その 1 つのアプローチにタスク外発話を検出し，Web 検索で処理する方法がある．しかし，一般に音声対話システムの言語モデルはタスク内の発話を認識できるようドメインを限定して構築されているため，多様性があるタスク外発話を精度良く認識できない．そこで，タスク外発話においてもある程度の認識性能を出せる汎用性の高い言語モデルが必要となる．本報告では，大規模テキストコーパスである Google N-gram（正式名称 : Web 日本語 N グラム第 1 版）を用いて言語モデルを構築し，その汎用性を 3 種類の音声データで評価した．読みは形態素解析器 *mecab* を用いて自動的に付与した．3 種類の音声データにおける単語正解率と単語正解精度を求めた結果，Google N-gram から構築した言語モデルは，音声データのドメインに合っている言語モデルよりも性能が劣るものの，新聞コーパスモデルと同等の単語正解率を得た．ただし，今回評価した Google N-gram の言語モデルはあくまでもベースラインであり，誤った読み付与を含んでいるなどの問題点がある．これらを改善すれば，より性能を向上できると考えられる．また，構築した Google N-gram の言語モデルは 3-gram であり，Google N-gram の最大の特徴であるデータ量を有効に活用して 4-gram や 5-gram のモデルを構築すれば，さらなる性能の向上が期待できる．

Evaluation of the Task Versatility of Google N-gram Models in Speech Recognition

Keigo KUBO,[†1] Jumpei MIYAKE,[†1,*1]
Hiromichi KAWANAMI,[†1] Hiroshi SARUWATARI[†1]
and Kiyohiro SHIKANO[†1]

In recent years, spoken dialogue systems capable of responding to various utterances have been studied. For example, there is an approach that detects out-of-task utterances and process them by the Web retrieval. However, in general, a language model in a spoken dialogue system is built to recognize in-task utterances. Therefore, it is difficult for a spoken dialogue system to recognize various out-of-task utterances with high accuracy. In this report, we constructed a tri-gram language model using the Google N-gram, which is a large text Corpus, and evaluated the versatility of the model with three types of speech data. As the Google N-gram does not include readings, they are automatically given by the morphological analyzer *mecab*. Results on word correct rate and word accuracy show that the language model built from Google N-gram is inferior to the models that customized for the domain. However, the model has equal performance to the JNAS, the Newspaper language model, on word correct rate. It should be mentioned that the evaluations contained in this report are the first trial and baseline results of the model. Because there are still several problems, such as wrong reading included in the Corpus, we can expect improvements in the performance by correcting them. In addition, as the language model built here is a tri-gram model, If 4-gram or 5-gram models are introduced, further improvement is also expected.

1. はじめに

近年，音声認識技術を用いたカーナビなどの製品や音声 Web 検索などのサービスが一般のユーザに提供されるようになった．音声認識を用いる利点としては，入力に手を使う必要がないこと，音声がほとんどの人間にとって扱いやすいインターフェースであることなどが挙げられる．

音声認識を用いたシステムの 1 つに，施設案内などを行う音声情報案内システムがある．実用的な音声情報案内システムを実現するためにはユーザの多様な発話とそれに対応するためのシステム拡張コストを検討する必要がある．この 2 つの問題に対応した応答手法の 1 つに，質問応答データベース（QADB）を用いる手法がある．この手法では質問例と適切な応答のペアをデータベース化した QADB を用いて，入力発話と質問例との類似度を計算し，最も類似度が高かった質問例に対応する応答を選択するものである．筆者らが開発・運用を行っている音声情報案内システム「たけまるくん」[1)] も QADB を用いた音声対話システムの 1 つである．

このような用例ベースの音声情報案内システムは，QADB 中にある想定内の発話（タス

†1 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of science and Technology
*1 現在，ヤフー株式会社
Presently with Yahoo Japan Corporation.

Vol.2009-SLP-79 No.33
2009/12/22

ク内発話）に対しては応答可能だが，QADB にはない想定外の発話（タスク外発話）には対処できないという問題がある．この問題の解決法として，タスク外発話を検出し，Web 検索タスク[2] で処理することで，タスク外発話でも音声対話システムに何かしらの応答を行わせる方法が考えられる．しかし，音声対話システムの言語モデルはタスク内発話をうまく認識できるようドメインを限定して構築されている．このため，多様性があるタスク外発話において，誤った認識結果を出力する可能性が高くなる．そこで，タスク外発話の多様性に対応できる汎用性の高い言語モデルを構築する試みとして，大規模テキストコーパスである Google N-gram(正式名称 : Web 日本語 N グラム第 1 版[3]) を用いて言語モデルを構築し，その汎用性を評価した．

以下，2 節では音声情報案内システム「たけまるくん」とそのタスクの一つである Web 検索タスクについて説明し，3 節では Google N-gram とその言語モデルの構築方法を説明する．また，4 節ではその性能評価実験の結果を示し，5 節で結果をまとめる．

2. 音声情報案内システム「たけまるくん」

2.1 システムの概要

音声情報案内システム「たけまるくん」（図 1）は，2002 年 11 月より奈良県生駒市にある生駒市北コミュニティセンター ISTA はばたきに常設しているシステムである．このシステムでは施設に訪れた一般の人々に対して施設・観光情報案内を行っている．「たけまるくん」は対話戦略としては一問一答形式をとり，情報案内の他に時間や天気，エージェント自身に対する質問などの QA タスクを持っている．

「たけまるくん」の処理の流れを図 2 に示す．まず，マイクロホンより入力された音声には，Gaussian Mixture Model（GMM）による雑音棄却処理が行われる[4]．次に年齢層別に用意された音響モデルと言語モデルを用いて音声認識が行われ，音響尤度により年齢層を識別する[1]．この時，識別された年齢層の音声認識結果とその年齢層に合わせて用意した QADB の質問例を用いて式 (1) により類似度スコアが算出される[5]．式 (1) は音声認識結果と質問例との形態素単位での一致数を求め，それを質問例の形態素数か音声認識結果の平均形態素数の最大値で除算した値である．システムの応答としてはこの値を用いて最近傍法により類似度がもっとも高かった例に対応した応答が選択される．

$$\text{類似度スコア} = \frac{\text{形態素単位での一致数}}{\max\left(\text{質問例の形態素数}, \text{音声認識結果の平均形態素数}\right)} \quad (1)$$

図 1 音声情報案内システム「たけまるくん」
Fig. 1 Speech-oriented guidance system “Takemaru-kun.”

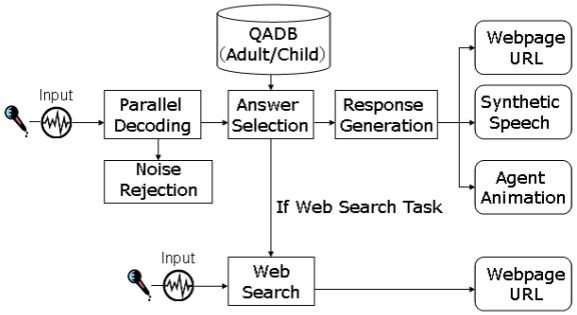

図 2 「たけまるくん」の応答処理の流れ
Fig. 2 Process flow of “Takemaru-kun.”

2.2 「たけまるくん」における Web 検索タスク

「たけまるくん」では汎用性のある情報提供を行うために音声認識による Web 検索タスクも提供している．Web 検索タスクは「検索開始」の発話をトリガーとして，QA タスクから Web 検索タスクへと 1 ターンだけ切り変わる．

この Web 検索タスクをタスク外発話にも用いることで，タスク外発話に対して何かしらの応答を行わせる方法が考えられる．このため，タスク外発話を検出する研究[6] が行われているが，タスク外発話を検出しても，そのタスク外発話をうまく認識できなければ，Web

Vol.2009-SLP-79 No.33
2009/12/22

検索タスクにおいてユーザのタスク外発話に無関係な Web ページを提示してしまう可能性がある．よって，タスク外発話の認識性能を向上するためにタスク外発話の多様性に対応できる汎用性の高い言語モデルを用いる必要がある．そこで今回，大規模テキストコーパスである Google N-gram に着目し，その汎用性の評価を試みた．

3. Google N-gram

3.1 Google N-gram とは

Google N-gram(正式名称 : Web 日本語 N グラム) とは Google 株式会社が提供している大規模テキストコーパスのことで，一般に公開されている Web ページから 1〜7-gram までの n-gram が抽出されたものである．基本的には日本語だが，他の言語が混ざっている n-gram もある．n-gram の抽出に用いられた Web ページのデータは 2007 年 7 月に取得されたもので，それ以降に作成された Web ページのデータは含まれていない．また，n-gram の抽出に用いられた Web ページの数は公開されていない．n-gram の抽出に関しては以下のような前処理が行われている[⋆1]．

(1) 文字コード変換
Web ページを UTF8 に変換して，n-gram を抽出する．

(2) 正規化
すべての文字列は Unicode が定める NFKC により正規化されている．表 1 は日本語に影響する主な正規化である．

(3) 文の分割
”.”, ”!”, ”?”, ”．”, ”。”, ” ! ”, ” ? ” を区切り文字として文を抽出している．そのため,「モーニング娘。」や「Yahoo!」 といった区切り文字を含む単語がある文は、文法上不適切な位置で分割される可能性がある。

(4) 対象文の同定，選別
Web ページは文章だけでなくタグも含まれるため，タグなどを抽出対象から除外する処理を行っている．具体的には 5 文字以下もしくは 1024 文字以上 (Byte 数ではなく、Unicode 文字数)，ひらがなが全体の 5%未満，日本語の code point(表 2 参照) の割合が 70%未満の場合は抽出対象から除外されている．

⋆1 GSK2007-C Web 日本語 N グラム第 1 版 README - http://www.gsk.or.jp/catalog/GSK2007-C/GSK2007C_README.utf8.txt

表 1 日本語に影響する主な正規化
Table 1 Normalization affect Japanese

全角英数字 → 半角英数字
半角カタカナ → 全角カタカナ
ローマ字数字 → アルファベット (Ⅲ→ III)
特殊記号 (㈱→ (株)，ﾄﾝ→トン)

表 2 判定に用いた日本語 code point
Table 2 The Japanese code points used for judgement of extraction sentence

U3040 .. U30FF
U31F0 .. U31FF
U3400 .. U34BF
U4E00 .. U0FFF
UF900 .. UFAFF

表 3 n-gram の詳細な情報
Table 3 Detailed information of n-gram

抽出 Web ページ数	非公開
取得期間	2007 年 7 月
総データサイズ	26GB
総単語数	255,198,240,937 (2550 億)
総文数	20,036,793,177 (200 億)
異なり 1-gram 数	2,565,424
異なり 2-gram 数	80,513,289
異なり 3-gram 数	394,482,216
異なり 4-gram 数	707,787,333
異なり 5-gram 数	776,378,943
異なり 6-gram 数	688,782,933
異なり 7-gram 数	570,204,252

(5) 単語分割
単語分割には形態素解析器 *mecab*-0.96 と *mecab-ipadic*-2.7.0-20070801 を用いている。分割結果は一切加工しておらず、解析エラーはそのままデータに反映されている。

また，Google N-gram は語彙と n-gram に対してカットオフが行われている．語彙は Web ページに 50 回以上出現した単語を基本語彙とし，50 回未満の単語は未知語 (⟨UNK⟩) としてカットオフされている．n-gram は出現頻度が 20 回以上の n-gram を抽出対象にし，頻度が 20 回未満の n-gram はデータに含んでいない．n-gram の詳細な情報を表 3 に示す．また，表 3 の総単語数は n-gram を作成した Web ページの全単語数であって，Google N-gram 中にある全単語数ではない．

表 4 Google N-gram における言語モデルの構築条件
Table 4 Building condition of language model on Google N-gram

CPU	64bit デュアルコア Intel(R)Xeon(R) プロセッサ 5160 2 基搭載
メモリ	32GB
形態素解析器	*mecab-0.96*
形態素解析辞書	*mecab-ipadic*-2.7.0-20070801
言語モデル構築ツール	SRILM 1.5.9
言語モデル平滑化手法	Witten-Bell 法

表 5 構築した逆向き 3-gram モデルの異なり n-gram 数
Table 5 reverse trigram language model

異なり 1-gram 数	1135528
異なり 2-gram 数	39722785
異なり 3-gram 数	249785039

3.2 Google N-gram の言語モデル構築

今回，Google N-gram を用いて，前向き 2-gram と逆向き 3-gram の言語モデルを構築した．表 4 に構築環境と使用したツールを示す．

まず，1-gram のデータを UTF-8 から EUC-JP への変換を行い，漢字，平仮名，片仮名，”.”, ”!”, ”?”, ”．”, ”。”, ”！”, ”？”から構成されている語彙だけを抽出し，語彙辞書を作成する．また，Google N-gram のデータには読みが付与されていないため，抽出した語彙に対して 1 語ずつ *mecab* を用いて，読み付与を行った．この際，*mecab* は 1 語ずつ読みを付与するので，文脈情報が使えず，誤った読みを付与してしまう確率が高い．今回は誤った読みを改善せずに，そのまま言語モデルを構築した．そして，抽出した語彙だけで構成されている 3-gram を Google N-gram から抽出し，その得られた 3-gram だけで前向き 2-gram と逆向き 3-gram の言語モデルを構築した．この方法により構築した逆向き 3-gram モデルの異なり n-gram 数を表 5 に示す．Google N-gram の評価実験ではこの言語モデルを使用する．

4. 評 価 実 験

評価実験では Google N-gram の汎用性を評価するため，3 種類の音声データにおいて Google N-gram の単語正解率と単語正解精度を求め，他の言語モデルと比較を行った．

4.1 評価した音声データ

評価した音声データは以下の 3 つである．

- JNAS テストセット
  日本音響学会新聞記事読み上げ音声データのテストセットで，毎日新聞の記事を読み上げた音声が収録されている．大人の話者 23 名の合計 200 文のテストセットである．また，この音声データの実験には音響モデルに JNAS の PTM モデルを用いる．
- 「たけまるくん」ユーザ発話
  音声情報案内システム「たけまるくん」は実発話を収録したものである．発話データの特徴として施設や観光案内，時間，天気，エージェント自身に対する質問などの発話が多い．評価に用いた発話データは 2003 年 8 月に収集したデータで，大人発話が 1053 文，子供発話が 6543 文ある．発話内容に重複があり，異なり発話数は大人が 584 文，子供が 3490 文である．また，この音声データの実験には大人発話は大人用のたけまる PTM モデルを使用し，子供発話は子供用のたけまる PTM モデルを使用する．
- 「たけまるくん」タスク外発話
  藤田らの研究[6] により作成されたデータで,「たけまるくん」の収集発話データ（2002/11〜2004/10）の中で，人手で「たけまるくん」のタスク外発話だと判断して，抽出したタスク外発話のデータである．よって,「たけまるくん」ユーザ発話の一部もデータとして含んでいる．大人発話が 894 文，子供発話が 7028 文ある．発話内容に重複があり，異なり発話数は大人が 736 文，子供が 6705 文である.「たけまるくん」ユーザ発話より発話内容に重複が少ないことから，多様性のある発話が多いということがわかる．また，この音声データの実験には大人発話は大人用のたけまる PTM モデルを使用し，子供発話は子供用のたけまる PTM モデルを使用する．

4.2 実 験 方 法

評価実験では Google N-gram の汎用性を評価するため，3 種類の音声データにおいて Google N-gram の単語正解率と単語正解精度を求め，以下の言語モデルと比較を行った．

- MNP45: 1991 年 1 月〜1994 年 9 月の 45ヶ月分の毎日新聞記事コーパス（JNAS）から構築した言語モデル
- たけまるモデル: 2003 年 8 月以外の 2002 年 11 月〜2004 年 10 月における「たけまるくん」の書き起しデータから構築した言語モデル
- たけまるタスク内モデル: タスク外発話を除外した 2002 年 11 月〜2004 年 10 月における「たけまるくん」の書き起しデータから構築した言語モデル

これらは全て Google N-gram と形態素解析の条件を同じにするため形態素解析には *mecab-*

表 6 実験条件
Table 6 experiment condition

JNAS テストセット	言語モデル	Google N-gram MNP45 たけまるモデル（大人）
	音響モデル	JNAS PTM モデル
	音声認識エンジン	Julius Ver. 4.1.2
「たけまるくん」ユーザ発話	言語モデル	Google N-gram MNP45 たけまるモデル（大人・子供別）
	音響モデル	たけまる PTM モデル (大人・子供別)
	音声認識エンジン	Julius Ver. 4.1.2
「たけまるくん」タスク外発話	言語モデル	Google N-gram MNP45 たけまるタスク内モデル（大人・子供別）
	音響モデル	たけまる PTM モデル（大人・子供別）
	音声認識エンジン	Julius Ver. 4.1.2

0.96 と *mecab-ipadic*-2.7.0-20070801 を用いている．また，言語モデル構築ツールと言語モデル平滑化手法も Google N-gram と同じ SRILM 1.5.9 と Witten-Bell 法を使用し，前向き 2-gram と逆向き 3-gram の言語モデルを構築した．さらに，たけまるモデルとたけまるタスク内モデルは大人と子供の発話をそれぞれ分けて，大人用言語モデルと子供用言語モデルを構築している．音声データが大人の場合は大人用言語モデルを子供の場合は子供用言語モデルを使用した．表 6 が実験条件である．

4.3 実験結果

実験により得られた各音声データの評価の結果を表 7 に示す．Google N-gram は JNAS テストセットと「たけまるくん」ユーザ発話において単語正解率，単語正解精度が共にドメインに適応している MNP45 やたけまるモデル（大人・子供別）よりも低いことがわかる．このことから，あらゆる言語表現を含む巨大なテキストコーパスを用いて言語モデルを構築しても，ドメインに適応した言語モデルの性能には到達しないことがわかる．これは，言語モデルがあらゆる言語表現を含んで言語的制約が弱まるためである．

さらに，「たけまるくん」タスク外発話においても単語正解率，単語正解精度ともにタスク外発話を考慮していないたけまるタスク内モデル（大人・子供別）よりも低いことがわかる．Google N-gram がたけまるタスク内モデル（大人・子供別）よりも単語正解率，単語正解精度ともに低いのは，たけまるタスク内モデル（大人・子供別）は学習にタスク外発話

表 7 実験結果
Table 7 experiment result

		大人		子供	
		単語正解率	単語正解精度	単語正解率	単語正解精度
JNAS テストセット	Google N-gram	**59.38**	**35.44**	-	-
	MNP45	83.94	81.95	-	-
	たけまるモデル（大人）	34.65	24.9	-	-
「たけまるくん」ユーザ発話	Google N-gram	**58.7**	**39.81**	**47.83**	**20.48**
	MNP45	56.18	36.77	46.83	28.92
	たけまるモデル（大人・子供別）	78.9	64.06	75.3	59.64
「たけまるくん」タスク外発話	Google N-gram	**50.93**	**34.88**	**34.57**	**11.42**
	MNP45	50.88	40.16	31.48	24.52
	たけまるタスク内モデル（大人・子供別）	55.43	43.25	48.2	41.21

を含んでいないものの，ユーザ発話のフレーズや語彙において類似性があるためと考えられる．よって，Google N-gram とタスク内モデル（大人・子供別）の融合を行い，タスク外発話の発話に類似しているユーザ発話のフレーズや語彙に対応し，尚且つ多様な発話にも対応することで「たけまるくん」タスク外発話の認識性能の向上が期待できる．

また，Google N-gram は JNAS テストセットにおいてドメインに適応していないたけまるモデル（大人）よりも単語正解率が高く，「たけまるくん」ユーザ発話と「たけまるくん」タスク外発話においても MNP45 と同等の単語正解率を得ることができている．このことから，ドメインに適応した言語モデルには適わないものの，Google N-gram は MNP45 と同等の性能を持っていることがわかった．

さらに，大人と子供の「たけまるくん」タスク外発話における単語正解率と単語正解精度を図 3 と図 4 に示す．図 3 と図 4 から，Google N-gram は他の言語モデルと比べて，単語正解精度が単語正解率よりも低い値を取りやすいことがわかった．このことから，Google N-gram は挿入誤りが他の言語モデルと比べて起きやすいことがわかった．

5. おわりに

タスク外発話の多様性に対応できる汎用性の高い言語モデルを構築する試みとして，大規模テキストコーパスである Google N-gram に，自動的に読みを付与して言語モデルを構築し，その汎用性を評価した．Google N-gram はドメインに適応した言語モデルより性能が劣るものの，MNP45 と同等の性能を持っていることがわかった．

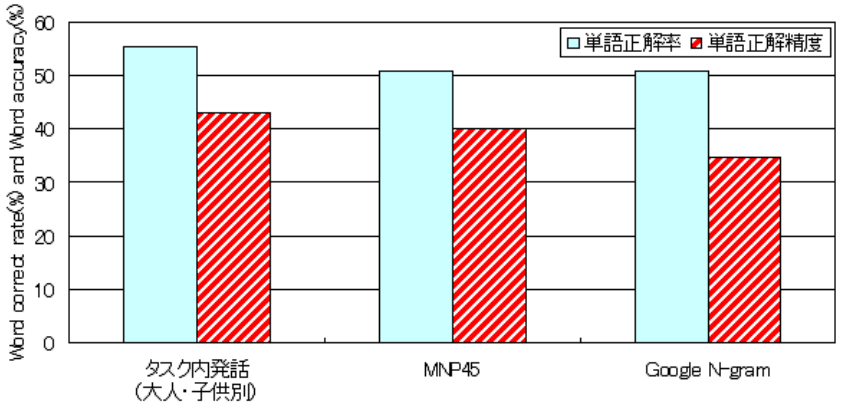

図 3 大人の「たけまるくん」タスク外発話における単語正解率と単語正解精度
Fig. 3 word correct rate and word accuracy for out-of-task utterances of “Takemaru-kun.” in adult

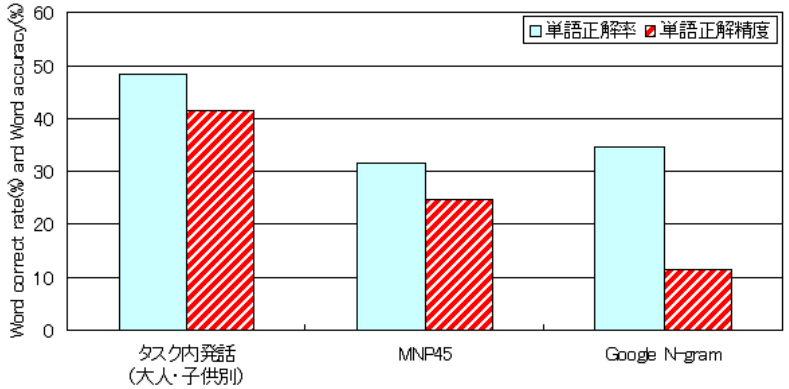

図 4 子供の「たけまるくん」タスク外発話における単語正解率と単語正解精度
Fig. 4 word correct rate and word accuracy for out-of-task utterances of “Takemaru-kun.” in child

また,「たけまるくん」タスク外発話においてたけまるタスク内モデル（大人・子供別）が他のモデルよりも認識性能が高かった．これは，たけまるタスク内モデル（大人・子供別）は学習にタスク外発話を含んでいないものの，ユーザ発話のフレーズや語彙において類似性があるためと考えられる．よって，Google N-gram とたけまるタスク内モデル（大人・子供別）の融合を行うことで,「たけまるくん」タスク外発話の認識性能の向上が期待できる．

また，今回の実験結果である図 3 と図 4 から Google Ngram は他の言語モデルと比べて，単語正解精度が単語正解率よりも低い値を取りやすいことがわかった．これは挿入誤りが他の言語モデルと比べて起きやすいことを示している．この問題は挿入ペナルティの調整や 3-gram より上の 4-gram，5-gram を用いて言語的制約を強めることで改善が期待できる．

さらに，今回の実験により，Google N-gram に対して正確な読みを付与できないという問題点も見つかった．これは，*mecab* を用いて 1 語ずつ読みを付与していることから，解析の際に文脈情報が使えず，正確な読みが付与できないからである．これを改善するには，n-gram の各単語を一度統合して，形態素解析しなおすなどの処理が考えられる．

今回評価した Google N-gram の言語モデルはあくまでもベースラインであり，今後これをもとに言語モデルの融合，挿入ペナルティなどの認識エンジンのパラメータ調整，Google N-gram の逆向き 4〜5-gram の構築，読み付与の改善などを行い認識性能の向上を試みる．

参 考 文 献

1) R. Nisimura, A. Lee, H. Saruwatari, K. Shikano: Public Speech-oriented Guidance System with Adult and Child Discrimination Capability, *In Proc. ICASSP* 2004, pp.433-436, 2004.
2) 三宅純平，竹内翔大，川波弘道，猿渡洋，鹿野清宏: 音声対話システムにおける Web 検索タスクの発話分析と Web 検索のための大規模単語コーパスの検討 (言語モデル), 情報処理学会研究報告. SLP, 音声言語情報処理, Vol.2008, No.68, pp.19-24, 2008.
3) 工藤拓, 賀沢秀人著: Web 日本語 N グラム第 1 版, 言語資源協会発行, 2007.
4) A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari, K. Shikano: ”Noice Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs,” *In Proc. ICSLP* 2004, TuA1302p-2, Vol.I, pp.173–176, Oct. 2004.
5) S. Takeuchi, T. Cincarek, H. Kawanami, H. Saruwatari, K. Shikano: Question and Answer Database Optimization Using Speech Recognition Results , INTERSPEECH 2008, pp.451-454, Sep, 2008.
6) 藤田洋子，竹内翔大，川波弘道，松井知子，猿渡洋，鹿野清宏: SVM を用いたタスク外発話検出における特徴量の組み合わせに関する検討, 日本音響学会講演論文集, 3-1-2, pp. 89-92, Sep. 2009.