社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－DBS－141 (10)
2007－GN－ 62 (10)
2007／1／25

ＮｅｔＴｖ：ＮｅｔＮｅｗｓとテレビ放送のクロスプラットフォームにおける動画インデキシングと音声検索

田中克幸¶　滝口哲也¶¶　有木康雄¶¶
神戸大学工学部
E-mail: ¶katsutanaka@me.cs.scitec.kobe-u.ac.jp, ¶¶{takigu,ariki}@kobe-u.ac.jp

情報網・Ｗｅｂ２．０の発展や放送のデジタル化により、情報整理が困難なメディア、映像、画像、音などの普及が情報の無秩序な肥大化を促進し情報氾濫を招いている。情報量の爆発とプラットフォームの多様化により、ユーザーが欲しい情報が入手できない状況にあり、効率的にユーザーが欲しい情報だけを入手できる環境が必要とされてきている。そこで、本稿では、ＮｅｔＮｅｗｓとＴＶ映像のクロスプラットフォームの動画インデキシングと音声インタフェイスによる、検索システムを構築し、ユーザーが快適に動画観覧でき、疑問解決をできるＮｅｔＴｖシステムを構築し、情報の統合によるユーザーの検索軽減を目指した。

ＮｅｔＴｖ：Multimedia Cross-platform video indexing and retrieval with speech interface

Katsuyuki, TANAKA¶　Tetsuya, TAKIGUCHI¶¶　Yasuo, ARIKI ¶¶
Faculty of Engineering, Kobe University
E-mail: ¶katsutanaka@me.cs.scitec.kobe-u.ac.jp, ¶¶{takigu,ariki}@kobe-u.ac.jp

The advancement of information technology, which includes such developments as Web2.0, on digital TV and Broadband, enables anyone and everyone to access and participate to access any form of media, like documents, movies, images etc via the internet very easily. However, due to information growth and the decentralization of platforms, users are faced with increasing difficulty in finding the information that they really are interested in. Our research enables the searching of news on the internet (NetNews) and TV by speech interface, thereby offering users a better search of cross-platform videos.

１．はじめに

近年、ブロードバンド化、ユビキタス化、Ｗｅｂ２．０などインターネットの利用環境や情報技術の発展に伴い、多くの人々によって簡単に情報収集・提供が可能になっている。インターネットの速度が上がれば上がるほど、そこに流れる情報はより多くなり、近年では、テキストだけではなく、あらゆるメディアの情報（映像、画像、音）が普及し、WWWはあらゆる情報を提供している大規模なマルチメディア情報源となり、現代人の多くはＴＶの前で過ごす時間以上に、ネットの前での時間を費やす社会

となっている。

また、近年のテレビ放送のデジタル化計画により、放送とデータ通信の位置づけが変わり、テレビ放送も多チャンネル化し、色々な番組だけではなくデータも入手できるようになろうとしている。この結果テレビ放送は、１つの新しい multimedia 情報源としての位置づけを確立しようとしている。情報社会は情報提供という面で飛躍的に革新し、映像、画像、音（音楽）など、多種多様な情報が存在し、いまやあらゆる情報が簡単に入手できるようになった。

しかし、一見、便利そうに見えるインターネットや、多チャンネル化し色々な番組みを楽しめるはずのテレビも、流れている情報が急速に増え続けすぎて、無秩序な情報氾濫をまねいている。また、これらの多様化した情報を効果的に管理しきれていないことから情報の無法地帯といった状況にある。このため、便利なはずのインターネットが、ユーザーの欲しい情報を簡単に探し出せない空間となり、情報が巨大化していくなか、いかに効率的にユーザーが欲しい情報だけを入手できる環境が必要とされてきている。

そこで、本研究では、点在した media source の情報をユーザーが円滑入手できるように、Cross-platform の情報統合と情報整理、次世代検索インタフェイスの構築を最終目標に掲げ、より効果的な動画のインデキシングと音声インタフェイスによる検索について検討する。

２．関連研究と問題点

２．１検索とインデキシング

このような情報の氾濫に対処すべく、様々な研究がなされてきた。多くの検索エンジンは、テキスト検索を提供しており、YouTube[1]等は動画の検索、Google[2]は Text・イメージ・動画・マップ等コンテンツの検索を提供している。イメージの効果的なインデックス化の方法としては、周辺のテキストをメタデータとして使用している手法[3][4]が多く取られている。動画インデックス化では、Google や YouTube が採用している方法として、提供者が動画を投稿するときに付与するメタデータをもとに、テキストと連動させて行うのが主流となっている。

しかし、情報整理が困難なメディア（動画、画像、音・・・etc）の情報が肥大化を辿る一方、情報提供の発展のわりに、“情報整理”といった技術はあまり発展していない現状がある。各 Modality における解析技術の限界などにより、よりリッチなインデックス化がなされていないことが問題としてあげられる。ＣＢＩＲなどを使って画像 query 型の検索も研究されているが[8]、画像描写技術の問題により、その精度に限界がある。特に、映像検索を提供している検索エンジンの多くは、ユーザーによるマニュアル付与された情報を動画のメタデータとして、インデックス化をしており、情報不足や正確性・詳細性に欠けるところがある。特にリッチなコンテンツである動画を手動のデータ付随で表現しきれないのが現実で、多くの動画を投稿しようとしたときの手間は多大なものがある。

２．２放送をインターネットにおける情報統合

放送コンテンツにおいても、多チャンネル化により、番組数が増えすぎて、自分の好きなタレントや歌手の出演番組、好きな話題（eg.明治維新）について、提供された番組などを探すのは不可能な状況である。加えて、最近ではオンデマンド放送の普及や放送のデジタル化により、放送とデータ通信の位置づけが変わり、インターネットと放送事業の垣根は崩れ、さらに過剰な情報氾濫を招くのは明らかであると予想できる。

現在、Text、イメージ、動画の検索を問わず、

ユーザーがキーワードを羅列した Query 型の検索が主流で、何万ページもの検索結果の中からユーザーのナビゲートにより情報をフィルタリングしていく手法がとられている。さらに、コンテンツ内でもっと知りたい情報を見つけるたびに、検索エンジンを使って検索し、膨大な量の情報からフィルタリングする処理を繰り返さなければならない状況である。

また、情報がネットドメイン、ＴＶドメインなど異なったプラットフォームに分散し点在している現状では、どこに何があるのか、どこから何を見つけたらいいのか、欲しいものを手に入れるのが困難な状況になっている。インターネットと放送は、双方情報を扱うメディアであるにもかかわらず、“情報統合”がなされていないといっても過言でない。

２．３音声対話検索

音声対話システムにおいては、従来、発話認識を行うドメインを設定し、そのドメインに存在する語彙で辞書を作り、それを基に発話を認識するスタイルが多い[9]。しかし、この方法は発話認識領域が限られてしまうので発話に対する柔軟性がない。ニュースやＷｅｂなどの日々内容が変わっていくようなダイナミックなドメインでは実用的ではない。近年では、Ｗｅｂなどを使い大語彙のコーパスを生成する方法も提案されているが[11]、ユーザーの自由発話を認識できるようになるには、辞書のメンテナンスに莫大な労力と時間がかかるため、これも実用的ではないと考えられる。

このように、増え続けるマルチメディア情報に対して、検索エンジンなどにより情報コントロールは円滑に行われているように見えるが、ユーザーの負担は大きく、情報量はあるものの、提供状況の質は良いものとはいえない。多くの情報を扱うのではなく、より効率的に各ユーザーのニーズに特化した、ユーザー嗜好型情報コントロールの必要性が求められていると言える。

３．解決方法とＮｅｔＴｖシステム概要

３．１解決方法

多くの場合、ユーザーが求めている情報は、WWWやＴＶ情報のほんの一部の情報で、情報全てを必要としていない。そして、通信と放送の融合が重要となりつつある現状では、ＴＶ放送とネットを１つのプラットフォームとして考えるべきである。また、ＷＥＢやＴＶなど日々流動するコンテンツを対象とて音声認識システムを構築する場合、決まりきったことしか言えないドメインに限定した、音声対話方法は向いていない。ユーザーのニーズが反映される状況であれば、大語彙である必要はなく、ユーザーがその時々に発話しそうな単語を認識できることにある。

つまり、前章であげられたマルチメディアの問題点を克服するには、ユーザーがより効果的にそのときに必要な情報をプラットフフォームフリーに入手できることが重要で、情報ドメインをユーザーが欲しい嗜好情報ドメインにマッピングすることが大切であると考えられる。

３．２ＮｅｔＴｖ概要

本論文では、まず、プラットフォームの統合と（これを、″マルチセマンティックメディア空間″と呼ぶ）、日々変化するダイナミックな環境での音声認識を実現すべくその環境基盤づくりに重点を置き、インターネットネット、ＴＶ放送などの映像メディアを、以下の点で情報コントロールが円滑に行えるシステムを構築した：

１．　ＮｅｔＮｅｗｓとＴＶ映像のクロスプラットフォームにおける動画インデキ

シング

2．　映像のメタ情報の自動付与

3．　音声認識のための辞書を自動生成した上で、情報ドメインに流動性をもたせ、音声インタフェイスによる検索と質問応答機能を実現する

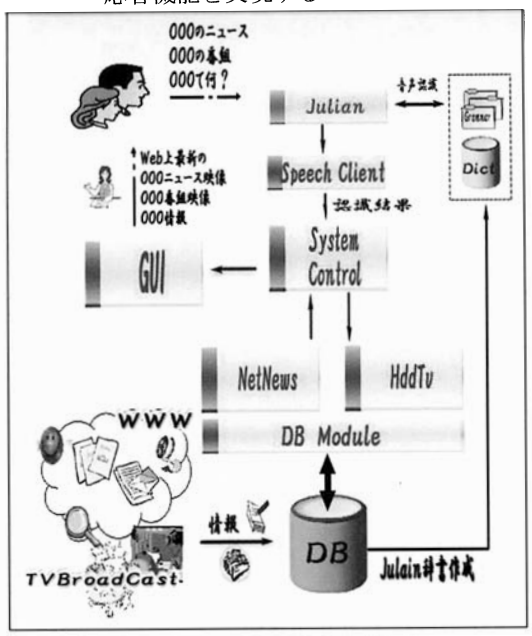

Figure 1 ＮｅｔＴｖ 概要

このシステムは、Figure1 に示すように、２つのモジュール、ＮｅｔＮｅｗｓとＨｄｄＴｖからなる。ＮｅｔＮｅｗｓはネット上に流れるニュースを自動的に集めてきて、周りのテキスト情報をメタ情報として付与しニュースをインデックス化することにより、ニュースのダイナミックなコンテンツ変化に対応できる。

ＨｄｄＴｖはユーザーが見たいＴＶ番組をＨＤＤ型の映像録画環境で視聴できる環境を想定し、ネット上にあるＥＰＧを自動的に収集し、それをメタ情報として、映像に付与する。こうして集められた情報を元に、映像をキーワードのよりインデックス化し、音声インタフェイス用の辞書を自動的に作成して、検索可能にする。また、キーワードに関してユーザーが疑問に思うことを質問できる機能も装備している。

これにより手動によるインデックス化の手間を省き、ＴＶとネットのクロスプラットフォームの情報を扱うことができ、ドメインに限度されることが無い音声認識システムを構築できると考えられる。これをＮｅｔＴｖと呼び、ユーザーがクロスプラットフォーム間に点在する動画を快適に観覧でき、疑問に思うことを質問可能にし、現在のトピックにそった自然発想による対話動画検索システムの構築を試みた。

４．ＮｅｔＴｖシステム詳細

ＮｅｔＮｅｗｓ、ＨｄｄＴｖ、音声インタフェイスの詳細について述べる。

４．１ＮｅｔＮｅｗｓ

Figure2 にＮｅｔＮｅｗｓの構成を示す。ネット上にあるニュース動画を自動的に入手するとともに、時事のトピックをキーワード音声検索により観覧することが可能である。

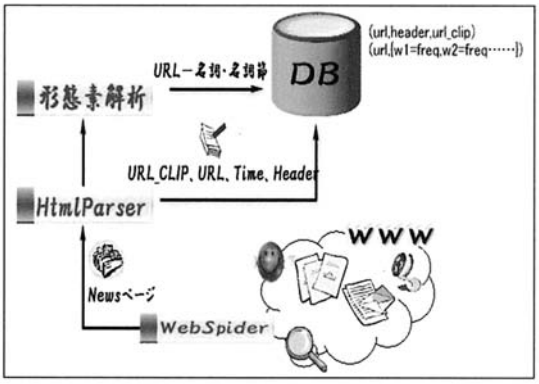

Figure 2 ＮｅｔＮｅｗｓ

Ｗｅｂ上のニュースを人が閲覧する典型的な方法は、Figure3 に示すようにニュースのヘッダーリストページからリンクされた詳細記

事ページを参照する方法である。

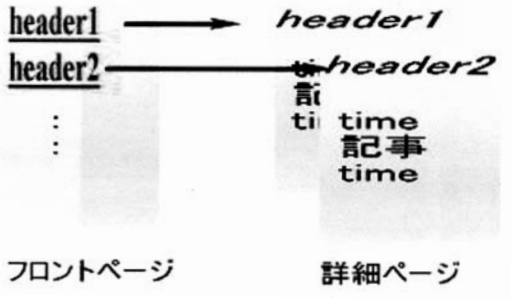

Figure 3 News Site

これを利用して、自動的にニュースサイトから、リンクと詳細記事を次の手順で集めてくる。

1. インターネット内を巡回し、ニュースサイトからヘッダーリンクを探して（url,header）のペアリストを作る。
2. そのリンクを辿って詳細記事ページを集めてくる。

集めた各詳細記事ページにｈｔｍｌパーザをかけ、記事部分と動画部分を切り出してくる。

3. html タグ内から動画ＵＲＬを探し出して抜き出してくる。
4. リンクで集めたヘッダーを探してそれを記事の始まりとし、時間の表示などのタイムスタンプをもとに記事の終わりと判断して記事を切り抜く。

url と詳細記事を用いて動画（url_clip）のインデックス化を行う。

5. 抜き出してきた詳細記事とヘッダーを文単位に分解して、茶筅[10]に掛け形態素解析を行う。
6. 解析結果から、名詞、未知語を取り出してインデックスキーとする。また、名詞の連続とカタカナの連続は、１つのフレーズとしてキーとみなす。

この結果（url,header,url_clip）、(url,{w1=freq1,w2=freq2……})のインデックステーブルが作成され、キーワードによる検索が可能となる。映像中のシーンを“Video Grammar”を使ってシーン別に摘出し、メタデータを付与する研究[5]もされており、シーンインデックス化という方法を行うことも可能であると考えられるが、オンラインのコンテンツを解析するのは難しいため、的確に動画コンテンツ内容を示しているニュース記事は動画のインデックスに最適であると考える。

４．２ＨｄｄＴｖ

ＴＶ放送で、ユーザーが見たいと思うＴＶ番組を録画した環境において、ＥＰＧなどを利用することにより、情報付与を行い、音声インタフェイスによる閲覧と、関連情報の検索を可能にする。Figure4 にＨｄｄＴｖの概要を示す。

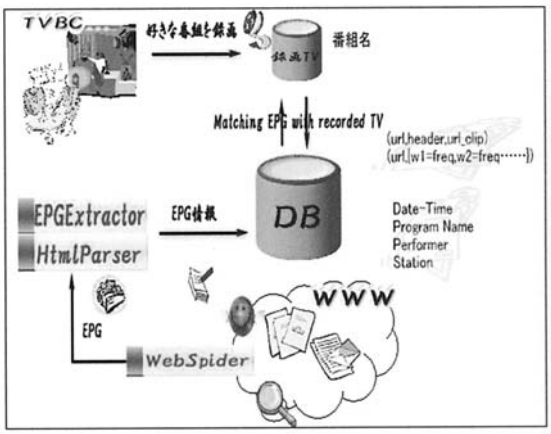

Figure 4 ＨｄｄＴｖ

ネット上の様々なサイトでＥＰＧが公開されているので、録画された番組とＥＰＧとのマッチングを行い以下の手順で実現する。

1. 通常、先１週間のｉＥＰＧが提供されているので,これを定期的に自動的に入手してくる。（地上波、BS アナログ、BS デジタル、CS）
2. 入手したｉＥＰＧより、人名、番組名、

時間、日のテーブルを別々に作り、それらをリンクするプログラムテーブルを作成する。

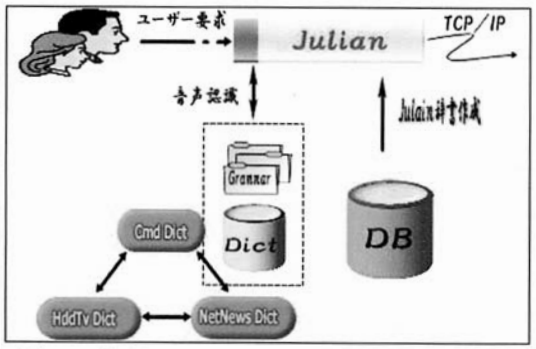

Figure 5 Ｓｐｅｅｃｈ Ｉｎｔｅｒｆａｃｅ

3．ＴＶ番組を録画した際に付与されている番組名と日時から、ＥＰＧとの照合をとり、ＥＰＧ情報を動画に付与する。

4．ＴＶ番組のタイトルを形態素解析にかけ、名詞と未知語を検出し、記号以外の品詞の連続を動画検索のキー単語とする。

以上の処理により、ＥＰＧを持って、出演者、日時、番組名などのキーをもとに検索することを可能にした。また、映像に付与されたＥＰＧと番組タイトルをもとに、ＮｅｔＮｅｗｓ同様に(url, header, url_clip)[1]、(url, {w1=freq1, w2=freq2……})のインデックステーブルを作成することができる。付与されたＥＰＧ情報と先１週間のＥＰＧとをマッチングすることで、その出演者が出演する他の番組などの情報を録画番組に自動的に付与することができるため、ＰＣやＴＶの前に座ってユーザーが検索をかけて探す負担を軽減できる。

４．３Ｓｐｅｅｃｈ Ｉｎｔｅｒｆａｃｅ

情報が日々更新される環境では、音声認識用の辞書をドメインごとに自動生成し、ドメインを柔軟に切り替えられる方法がより効果的である。そこで、ＮｅｔＴｖでは、トピックに関する音声認識用の辞書をドメインごとに日々自動的にしている。従って、ユーザーがドメインを変えるたびに、また、ドメインの内容が日々更新されるたびに、音声認識用の辞書をダイナミックに変化さることができ、多様なドメインの最新情報に関する音声認識を実現している。

Figure5 に対話の概要を示す。Julius-3.5.2 リリース[6]を Julian モードでコンパイルして検索キーワードを可変できるように文法を生成し、これをモジュールモードで起動させて、ＴＣＰ／ＩＰによりシステム Speech Client 部分と直結する。各ドメインのデータが集められてくると、ＮｅｔＮｅｗｓとＨｄｄＴｖで作成したインデックステーブル中にあるインデックスが自動更新されるので、この更新されたインデックスを音声認識用辞書として用いる。

Table. 1 発話方法リスト

NetNews cmd
最新情報リスト
ＯＯＯのニュース
詳細記事がみたい
ＯＯＯて何・誰？
HddTv cmd
最新情報リスト
ＯＯＯの番組
ＥＰＧが見たい
ＯＯＯて何・誰？
ＯＯＯの出演番組

Julian からＸｍｌで送られてきたユーザー

[1] header=title,url=epg

の発話内容をＸｍｌパーザーにより解析し、必要な情報(WORD, CLASSID, CM)を取り出して、各モジュールの protocol Handler に送る。protocol Handler は Julian に定義されたグラマーパターンを元に、アクション候補（Protocol）が定義されており、それを参照して、発話内容に対するアクションを決定する。各モジュールのシステム要求パターンをTable1 に表記する。

ＯＯＯは何？などの質問はキーワード部分を切り出したのにＷｉｋｉによりその説明を検索し、ユーザーに提示する。

５．実験

10 人（男 9 女1）にシステムを実際に使ってもらい、ＮｅｔＴｖの検証を行った。

５．１評価方法

ユーザーのシステムへの全ての発話をユーザー発話、ユーザー発話のうち検索を行うものを検索要求発話と定義し、評価方法と定義を以下に述べる。

タスク成功率:ユーザー発話を認識して希望したタスクが成功した割合を表す。

検索成功率:検索要求発話に対して検索が成功した割合を表す

トピックにそった内容度（precision）：検索結果が検索要求に対してどれだけ関係あるかを表す。#relevant docs retrieved/#docs retrieved

満足度:５段階でシステムの使いやすさや満足度を評価した。１．不満、２．やや不満、３．普通、４．やや満足、５．満足

５．２実験結果

１１７ニュース（６９１８語彙）と１８番組（１４０語彙）の環境下において実験を行った。

Table. 2 発話回数に対するタスク成功率と検索性効率

発話回数	タスク成功率	検索成功率
1回	６３％	４８％
2回	６９％	５７％
3回	７０％	５７％

Table. 3 ユーザーの満足度

満足度	1	2	3	4	5
人数	1	3	4	1	1

１回の発話に対する音声の正確な認識は、語彙の多いＮｅｔＮｅｗｓでは難しかった。認識困難な単語（eg. タイ）などは何度も発話を要した。タスクの２回までの言い直しにより成功率は約７０％の精度で行うことができた。

検索の成功率は、データベース上にないキーワードの発話を除くと、約６５％となる。音声認識して検索した結果は検索トピックにそぐわないものは出てこなかった。

認識率の精度の問題点としては、マイクと口元の距離が離れ過ぎるため認識率が低下する、文法の規則通りに決まった言い方をしないと認識できない、辞書にない未知語を発話すると誤った単語に認識されてしまい誤作動を起こすなどが考えられる。

満足度については、ユーザーの発話の繰り返し回数に満足度が左右される傾向にあった。ユーザーは３回以上の言い直しに難色を示すようである。

発話訓練されたユーザーがシステムを使用すると Table4 のような結果になる。検索成功率はキーワードがデータベースになかった発話を除くと約８０％になる。これより、ある程度の訓練をうければ、システムをより快適に使

用できると考えられる。

Table. 4 発話訓練者による成功率

発話回数	タスク成功率	検索成功率
1回	８５％	６７％

６．おわりに

本研究では、スピーチインターフェイスによりユーザーが動画を検索することができるシステムを構築した。ＮｅｔＮｅｗｓとＴＶ番組のクロスプラットフォームを１つのマルチメディア空間とみなして、情報の統合を試み、自動的にタグを付与した。音声インタフェイスは、日々変化するコンテンツに対応するように、ドメインを固定するのではなく、ドメインを可変することにより、ダイナミック性のあるシステムとした。

これにより、ユーザーはキーボード＋マウス主体のインタフェイスから、ソファに座ってリラックスした状態で動画の観覧することができる。さらに、クリックを繰り返して見たいものを探す必要もなくなり、より検索・閲覧環境が快適になる。通信と放送の融合が促進される中、ネットとＴＶを１つのプラットフォームとして考えるのは重要である。

現在ＴＶコンテンツのキーワードはタイトルに限定されているが、ＥＰＧ内のprogram-subtitleなどの情報より、コンテンツの拡張を検討して、より効果的なインデッキシングを模索する予定である。またニュース記事やprogram-subtitleの内容から、ユーザーが疑問に思うような事象を予測して、予めネット上から答えを探し出せるようなインテリジェンスの構築も検討している。さらに、動画のコンテンツ解析による情報付与、また、Ｂｌｏｇなどを用いた口コミ情報や評価によるインデキシングの可能性も検討する予定である。

参考文献

[1]グーグル：http://www.google.com/

[2] YouTube：http://www.youtube.com/

[3]是津耕司，田中克己：画像の文脈情報のWeb からの抽出と提示，日本データベース学会Letters，Vol.2，No.1，pp.99-102(2003)

[4]出原博、藤本典幸，竹野浩，荻原兼一：WWW画像検索における画像周辺のＨＴＭＬ構文構造を考慮した画像説明文の抽出方法，電子情報通信学会記述研究報告書，DE2005-136，pp.19--24，(2005).

[5]天野美紀，上原邦昭，熊野雅仁，有木康雄，下條真司，春藤憲司，塚田清志:”映像文法に基づく映像編集支援システム”情報処理学会論文誌，Vol.44，No.03，pp.915-924 (2003)

[6]河原達也，李晃伸:「連続音声認識ソフトウエア Julius」人工知能学会誌，Vol.20，No.1，pp.41--49(2005)

[8]Smeulders，A.W.M. et al.：Content-Based Image Retrieval at the End of the Early Years，IEEE Trans. on Pattern Analysis and Machine Intelligence，Vol. 22，No. 12，pp. 1349-1380 (2000).

[9]翠 輝久，河原 達也：質問応答技術を利用したインタラクティブな音声対話システムSIG-SLUD-A602-06 (2006)

[10]松本裕治：形態素解析システム「茶筅」，情報処理，Vol.41，No.11，pp1208-1214，(2000)

[11]Atsushi Fujii，Katunobu Itou：Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)，pp.1153-1156(2003)

[12]工藤拓，松本裕治:チャンキングの段階適用による係り受け解析，情報処理学会論文誌Vol.2002，Vol.43，No.6 (2002)