2009－SLP－75 (11)
2009／2／7

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

佐々木 浩† 中野 鐵兵† 緒方 淳‡ 後藤 真孝‡ 小林 哲則†

†早稲田大学 ‡産業技術総合研究所

あらまし:
ポッドキャストの音声認識における言語モデルの適応手法を提案する．ポッドキャストは，幅広い話題，タスクの音声データが日々増え続けるという特徴を持っているため，言語モデルをいかにして学習，構築するかが認識性能を左右する大きなポイントとなる．本稿では，言語モデリングにおいて，あらかじめポッドキャストが持つメタ情報と「語彙情報サービス [9]」を活用することで，ポッドキャスト音声認識の性能向上をはかる．具体的には，あらかじめ用意された言語モデリング用学習テキストを各テキスト毎に特徴語を抽出し，ポッドキャストのタイトルや概要などに記載された語との共起を基準にテキストの選択を行い，ポッドキャスト毎に特化された言語モデルの学習を行う．加えて，学習テキストやポッドキャストのメタ情報上の語の不足から生じる，テキスト選択の精度低下の問題を解決するため，語彙情報サービスのタグ情報を活用する．本手法で適応された言語モデルを実際に用いて，その性能を単語パープレキシティと未知語率で評価した結果，単語パープレキシティがベースラインの 86 %，未知語率もベースラインの 80 %となり，言語モデルの性能が改善されたことが確認された．

Topic Dependent Language Modeling Using Collective Intelligence Based Vocabulary Information

Hiroshi SASAKI†, Teppei NAKANO†, Jun OGATA‡, Masataka GOTO‡, Tetsunori KOBAYASHI†

†Waseda University ‡National Institute of Advanced Industrial Science and Technology (AIST)

Abstract:
This paper presents a language model adaptation method for automatic transcription of podcasts. Since podcasts include speech data that contains a variety of topics and many newly created words, well designed language models are indispensable to achieve sufficient speech recognition rate. In this paper, we propose a new topic dependent language modeling method by using meta information of podcasts and vocabulary information service. In this method, a large amount of training data are collected from the Internet such as web news and blogs on a daily basis. By using RSS texts of podcasts, topic dependent texts are selected from these training data, and proper language models are created for each podcast. In addition, we utilize the tag information of the vocabulary information service to solve the problem of the precision fall of the text choice that the lack of the word in a learning text and a meta information of Podcast cause. The assessment result showed that the performance of the language model using this method is improved because the word perplexity of the result using this method is 86% of that of the baseline and the out-of-vocabraly rate of the result using this method is 80% of that of the baseline.

1 はじめに

音声情報検索は，音声認識技術のアプリケーションの１つとして重要視され，近年でも活発に研究が展開されている．しかしながら，現状の音声認識技術では，あらゆる音声データから検索に必要な索引情報を，精度よく抽出することが困難なこともあり，Google等の代表されるテキストの検索のように日常的に利用されるには至っていない．一方，最近では，音声版のブログ (Weblog) ともいえる「ポッドキャスト」や動画共有サービス等が普及したことで，音声データに対する検索の重要性がより一層増してきたといえる．そこで我々は，Web 上の日本語のポッドキャストを音声認識によって自動的にテキスト化することで，それらをユーザが全文検索できるだけではなく，詳細な閲覧，編集も可能な音声情報検索システム「PodCastle[1]-[3]」の開発を行っている．

ポッドキャストは，幅広いトピック，タスクの音声データが日々増え続けるという特徴を持っているため，言語モデルをいかにして学習，構築するかが認識性能を左右する大きなポイントとなる．そのために我々は，政治，経済，スポーツ，芸能など幅広い最新のトピックをカバーする Web ニュースを利

用した言語モデリングを検討し，その有効性を示した [2][3]．この手法は，幅広いトピックを包含した大局的かつ大規模なモデルを構築し，より多くのトピックのポッドキャストに対応することを目的としたものである．しかし，一般的に，統計的言語モデルは対象音声のトピックのみにマッチしたコーパスから学習することが理想的であり，ポッドキャスト音声認識においてさらなる改善を得るためには，入力される各ポッドキャストに対して，トピックに関する動的な適応を行うことが有効と考えられる．言語モデルの適応手法，あるいはトピック依存言語モデルとしてはこれまで様々な検討がなされており，代表的なものとしては，各文書の単語出現確率に基づき文書自体の制約を導入する PLSA[4][5]，それを拡張させた LDA[6] に基づく方法がある．また講演音声認識においては，クラス N-gram を利用したオフライン教師なし適応手法 [7]，Web サーチエンジンを用いたトピック依存テキスト収集に基づく手法 [8] などが検討されている．言語モデル適応においては，対象となる音声に関するトピックの情報が事前に何らかの形で与えられる必要があり，上記の従来手法では適応前のベースラインモデルを用いて初期的な認識を行い，その結果のテキストをもとにトピックを推定している．しかしながら，ポッドキャストにおいては，前述のトピックに関する違いだけでなく，音響的な収録条件 (自由発話，雑音，背景音楽の有無等) も多様であるため誤認識が多く発生し，初期認識の段階でその音声のトピックを表す単語集合が得られるとは限らない．

本研究では，初期的な認識結果を利用するのではなく，ポッドキャストの各音声に付随するメタ情報を活用することでトピックを推定し，言語モデルの適応を行うことを考える．一般的に，ポッドキャストのメタ情報は，タイトルや概要などごく少量のテキストで与えられるため，それのみを用いてトピックを推定することは難しいと考えられる．そこで本手法では，我々が開発を進めている，集合知に基づく語彙情報サービス [9] を用いてメタ情報中の語の情報を補うことで，対象の音声に対するトピックを推定する．そして，推定されたトピックベクトルを利用することで，多様なトピックを包含する大規模なテキストコーパス中から，トピックに合ったテキストを選択し，トピック依存の言語モデルを生成する．

以下，次節でまず学習テキストの選択による言語モデル適応の基本的なアプローチを説明し，その際に想定される問題を述べる．次に，想定される問題

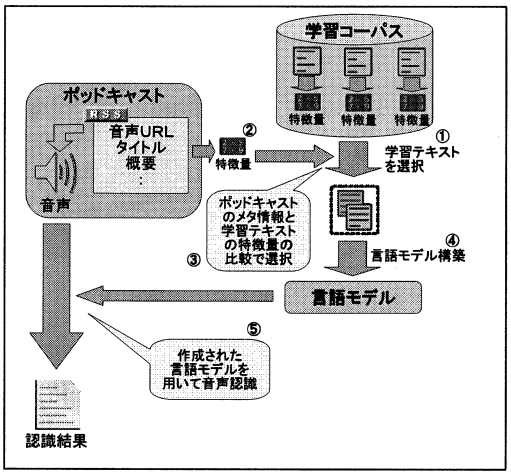

図 1: 提案手法の概要

に対する解決として，語彙情報サービスを用いる方法を 3 節で説明する．4 節で学習テキストを選択する際の本稿での学習テキストとポッドキャストのメタ情報の特徴量の抽出方法とその比較方法を説明し，5 節で，提案手法の有用性を評価する実験と，その結果について報告する．

2 テキスト選択による言語モデル適応

図 1 に提案する言語モデリング手法の概要を示す．学習コーパス中の各学習テキスト (Web ニュース記事) 毎にトピックとなる特徴量を抽出し (図 1①)，また，対象となる音声データ (ポッドキャスト) 毎にもそのメタ情報から特徴量を抽出する (図 1②)．抽出されたそれぞれの特徴量を比較し，対象音声のトピックに類似しているドキュメントを選択する (図 1③)．そして，選択された学習テキストを用いて，対象音声のトピックに依存した言語モデルを構築し，音声認識を行う (図 1④⑤)．

2.1 ポッドキャストの RSS 情報の利用

ポッドキャストはオンラインで配信される音声コンテンツの形式の一つである．1 回の配信の単位をエピソードと呼び，このエピソードは日々追加される．各エピソードの音声データ (mp3) の URL やタイトルなどのメタ情報（図 2）は RSS で配信される．RSS に記述されたタイトルや概要などの少量の文章から特徴量を抽出し，学習コーパス中の各学習

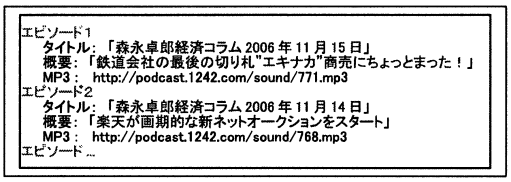

図 2: ポッドキャストのメタ情報 (RSS) の例

テキストの特徴量と比較し，エピソードの話題に沿うドキュメントを選択する．

2.2 テキスト選択の際に想定される問題

図 2 の例で，“タイトル：「森永卓郎経済コラム 2006 年 11 月 15 日」, 概要：「鉄道会社の最後の切り札 ”エキナカ ” 商売にちょっとまった！」” とあるように，RSS 上の文章量は非常に少ない上に新規性の高い表現が多い．そのため，テキスト選択を行う際，以下のような問題が想定される．

まず，ポッドキャストの RSS 上に出現した文章だけではそのポッドキャストの音声の話題が特定できない可能性がある．話題が特定できないと，適切に学習テキストを選択できないため，言語モデル適応の効果が表れない可能性がある．

また，ポッドキャストの RSS 上の文章が新規性の高い表現の場合，学習コーパス中に該当する表現がほとんど含まれず，適切に学習テキストが選択されない可能性がある．これも同様に言語モデル適応の精度を低下させてしまう要因となる．例えば，ポッドキャストの RSS 上に「年越し派遣村」という語があった場合，その語自体は学習テキストにはほとんど含まれなくとも，国内経済や派遣労働者問題に関した学習テキストが選択されることが望ましいが，RSS 上の語のみを基準に選択を行うと，適切に学習テキストを選択できない可能性がある．さらに，新規性の高い表現を含む場合，N-gram 学習の事前処理として行われる単語分割において，新規語彙に対する分割誤りに対処できない問題もある．

その上，言語モデリングにおいて各単語の音素情報が必要となるが，こうした新規語彙の読みの情報が無いことによって正常に言語モデリングを行うことができない可能性もある．

3 語彙情報データベースの活用

2.2 節の問題を解決するため，語彙情報データベース [9] の読み情報やメタ情報を利用する．語彙情報

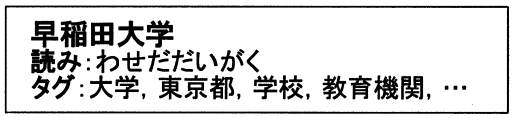

図 3: 語彙情報データベースの情報の例

データベースは Wikipedia[10] の記事やはてなキーワード [11] のキーワードなど，WWW 上で集合知として蓄積されている様々な情報から語を収集し，それを整理し，一元化された語彙資源として共有するオンラインサービスである．データベース上の語は日々更新され，新規性・正確性が維持されている．データベース上の語はそれぞれ読み情報，タグ情報，収集元の情報を持つ．タグ情報はその語の品詞やカテゴリを表す語への関連となっている．例えば「早稲田大学」には「大学」や「学校」などの語がタグとして関連付けられている (図 3)．ユーザはこのタグ情報を基に，大学名の語彙リストやレストラン名の語彙リストなど，適宜ユーザが必要とする語彙の集合の定義及び利用ができる．また，定義した語彙集合が更新されるとユーザに通知される枠組みも持つ．これらの機能は WEB アプリケーションとして動作するとともに，WEB API としても利用することができるので，WEB ページの解析・抽出作業などを行わなくても語彙情報を利用することができる．

3.1 語彙辞書としての利用

語彙情報データベースに登録されている膨大な語彙を用いて，単語分割の辞書構築や読み情報の取得を行う．これにより，未知語の数を最小限にすることができ，単語分割誤りや読み情報の欠落の影響を抑えることができる．

3.2 テキスト選択時のタグ情報の利用

テキスト選択の際，ポッドキャストの RSS 上の語や学習テキスト上の語のタグ情報を利用する．ポッドキャストの RSS 上の語や学習テキスト上の語のタグ情報によって，RSS 上の文章や学習テキスト上の文章に対するタグを求めることができるので，これらの情報も特徴量として利用する．

例えば，ポッドキャストの RSS 上に「年越し派遣村」という語があった場合,「年越し派遣村」に「労働問題」のタグが付与されているので，ポッドキャストの RSS の特徴量として「労働問題」も併せて学習テキストの選択を行う．こうすることにより，学習テキストに「年越し派遣村」が含まれなくても

「労働問題」によって比較的近い学習テキストを選択することができる．

4 学習テキストの特徴語の抽出とテキストの選択

本稿での学習テキストからの特徴語の抽出方法及びポッドキャストの RSS との比較による選択の方法を述べる．

4.1 学習テキストの特徴語の抽出

各学習テキストに対し，形態素解析を行い，各形態素の学習テキスト内での tf-idf を求め，これを学習テキストの特徴量とする．tf-idf は，文章中の特徴的な単語を抽出するためのアルゴリズムであり，以下のような tf と idf の二つの指標で計算される。

$$\begin{aligned} tfidf(w) &= tf(w) \cdot idf(w) \\ tf(w) &= \frac{n_w}{\sum_{alli} n_i} \\ idf(w) &= \log \frac{D}{d_w} \end{aligned}$$

ここで，n_w は文章中の単語 w の出現回数，D は全文章数，d_w は単語 w を含む文章数である．tf-idf が高いほど特徴的な語とみなせる．学習テキスト内の全ての語に対し，tf-idf を算出するが，3 の手法でタグ情報を統合する場合，学習テキスト内の語に付与された全てのタグとなる語を併せて tf-idf を算出する．

4.2 ポッドキャストの RSS を利用したテキスト選択と言語モデル適応

ポッドキャストの RSS は学習テキストに比べ非常に量が少ないため，4.1 節の様に特徴量を求めることが困難である．そこで，RSS 内の全ての語を特徴量とし，各学習テキスト内の語との共起関係を調べ，共起した語の学習テキスト内での tf-idf を用いて学習テキストの類似度を定義し，その類似度に基づいて学習テキストを選択する．

4.2.1 学習テキストの語の tf-idf に基づく類似度の定義

あるポッドキャストの RSS 上の文章を形態素解析した結果を $M = \{m_1, m_2, \cdots, m_n\}$，ある 1 つの学習テキストを形態素解析した結果を $D = \{d_1, d_2, \cdots, d_m\}$，ある形態素 w の学習テキスト D での tf-idf を $tfidf(D, w)$ としたとき，あるポッドキャストとある学習テキストの類似度 $S(M, D)$ を以下のように定義する．

$$S(M, D) = \sum_{i=1}^{n} tfidf(D, m_i)$$

例えば，RSS 上の文章が “郵政造反組の復党に批判の声” で，その形態素解析の結果が { 郵政造反組/の/復党/に/批判/の/声 } であるとする．これらの形態素のある 1 つの学習テキストでの tf-idf が “郵政造反組” = 12, “の” = 1, “復党” = 7, “に” = 1, “批判” = 3, “声” = 2 とすると，この学習テキストとの類似度は $12+1+7+1+3+1+2 = 27$ となる．

4.2.2 学習テキストの語とタグ情報の tf-idf に基づく類似度の定義

3 節のようにタグ情報を利用する場合は，語のタグとなる語を併せた tf-idf で類似度を算出する．つまり，ある単語 w のタグとなる語を $T_w = \{t_{w1}, t_{w2}, \cdots, t_{wu}\}$ とするとき，4.2.1 節で定義した M, D を $M = \{m_1, m_2, \cdots, m_n, t_{m_1 1}, t_{m_1 2}, \cdots, t_{m_n u}\}$, $D = \{d_1, d_2, \cdots, d_m, t_{d_1 1}, t_{d_1 2}, \cdots, t_{d_m v}\}$ として類似度を算出する．例えば，形態素解析の結果が { 郵政造反組/の/復党/に/批判/の/声 } で，“郵政造反組” “復党” に “政治” というタグ，“の” “に” に “助詞” というタグ，“声” に “名詞” というタグが付与されていた場合，{ 郵政造反組, の, 復党, に, 批判, 声, 政治, 助詞, 名詞 } を基に特徴量を構成し，類似度を算出する．

4.2.3 学習テキストとの類似度に基づいたテキスト選択

あるポッドキャストに対して各学習テキストの類似度を求めたら，その類似度の上位 N テキストを選択し，言語モデルを構築する．これによりポッドキャストに適応された言語モデルが構築できる．

5 言語モデルの作成と評価

提案手法で実際に言語モデルを構築し，その性能を評価した．

表 1: ポッドキャスト毎の実験結果

	ALL		RANDOM		WORD		TAG	
評価データ	Perplexity	OOV[%]	Perplexity	OOV	Perplexity	OOV	Perplexity	OOV
A	276.078	2.26	248.846	1.81	224.477	1.52	228.070	1.52
B	215.054	2.54	172.716	2.43	175.489	2.12	174.742	2.01
C	223.861	2.35	198.222	1.84	207.353	1.80	208.021	1.80
D	233.297	3.89	194.218	3.41	210.720	3.41	208.698	3.41
E	134.324	1.51	119.496	1.51	110.475	1.36	111.983	1.36
F	357.267	3.07	293.048	2.62	316.895	2.39	315.744	2.39
G	252.564	2.53	230.956	2.53	206.039	2.39	215.212	2.25
H	242.688	3.47	224.482	2.61	224.499	2.66	235.043	2.50

5.1 評価方法

以下の条件で書き起こし済みの各ポッドキャストに対する単語パープレキシティと未知語率を比較した．また，話し言葉への対応を行うため，日本語話し言葉コーパス [12] を全ての言語モデルに融合させた．

- CSJ + 全テキスト (ALL): ベースラインとして収集された全てのテキストを CSJ と融合し，言語モデルを構築
- CSJ + ランダム選択 (RANDOM): 収集されたテキストをランダムに選択し，CSJ と融合して言語モデルを構築
- CSJ + 単語ベース選択 (WORD): 収集されたテキストを 4.1〜4.2.1 節の手法で選択し，CSJ と融合して言語モデルを構築
- CSJ + 単語・タグベース選択 (TAG): 収集されたテキストを 4.2.2 節による語のタグ情報も利用した手法で選択し，CSJ と融合して言語モデルを構築

5.2 実験

事前に用意する学習テキストとして，Yahoo!ニュース [13] の 2007 年〜2008 年 12 月 24 日までの記事 1405348 記事を利用し，テキストの選択を行う場合は，その中から 300000 記事を選択し，日本語話し言葉コーパスと融合して言語モデルを構築した．言語モデルの構築には Palmkit[14] を用いた．言語モデル構築時の語彙リストは選択された学習テキスト上の頻度の上位 60000 語の語彙を用い，bi-gram や tri-gram のカットオフをそれぞれ出現回数 4 回以下に設定し，ディスカウントには Witten-Bell discounting を用いた．語彙情報データベースは 2008 年 12 月 12 日時点の総語数 927829 語のものを利用した．また，形態素解析の辞書には語彙情報データベース内の全ての語を用いた．評価対象となるポッドキャストの RSS はタイトルと概要のみを用い，これらの項目が適切に記述されているものを 8 エピソード用意した．

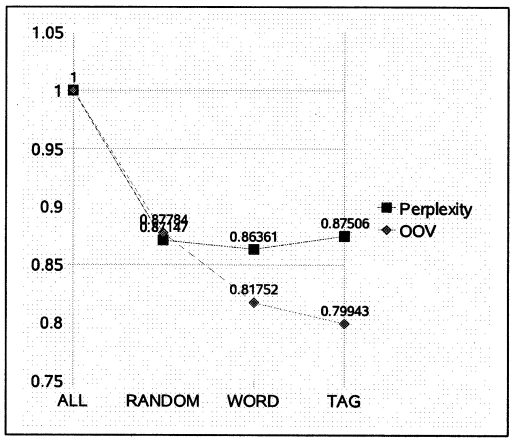

図 4: 各実験条件での比較結果

5.3 結果と考察

表 1 にポッドキャスト毎の音声の形式をニュースや対談などの形で示し，併せて単語パープレキシティと未知語率を示す．また，図 4 に各条件での単語パープレキシティと未知語率をベースラインの値で正規化した結果の平均を示す．全体の傾向として，テキストを選択した条件では単語パープレキシティ，未知語率が共に改善し，提案手法ではランダムに選択するよりも未知語率が改善する結果となった．語のタグ情報を利用した特徴量を基準にテキストを選択すると，語のみを基準とした手法よりもパープレキシティが少し増加し，未知語率が低下する結果となった．

最も効果が表れたAのポッドキャストのRSSは以下のようなものである．

> タイトル:ニュース・ズームアップ+α 1/10
> 嶌信彦
> 概要：ますます迷走するイラク

このポッドキャストの結果ではタグ情報を利用した条件の値が，単語パープレキシティ，未知語共に他の条件よりも低い．これは“イラク”という語に“中東”,“アジア”,“西アジア”などのタグが付与され，より話題に沿ったテキストが選択できたことによる効果だと思われる．しかしながら，いくつかのポッドキャストに関してはタグによる効果が表れず，パープレキシティがかえって増加してしまう結果となった．その原因の1つとして，複数の観点を持つ語のタグの利用によるトピック推定誤りがある．1つの表記で複数の意味を持つ単語がある．例えば“行列”は人の作る行列と数学用語の行列の意味を持つ．そうした語が利用された場合，文章中の用途に関わらず，全ての意味でのタグ情報が利用されてしまう．そのためトピックの推定を誤ってしまうケースがあった．例えば，Hのポッドキャストの RSS は以下のようなものである．

> タイトル：第18回「コンピューターウォーズ　「ユビキタス」」
> 概要：ICタグや携帯、SUICAなどを活用し、欲しい情報や支払いが直ちにできるコンピューターネットワーク型社会=ユビキタス社会…

このRSSの概要の“携帯”という語に“医学”というタグが付与されていた．このようなことが原因でトピック推定を誤り，言語モデルの性能を低下させてしまった．また，形態素解析誤りも致命的となった．Hのポッドキャストの概要の“利便性”という語は“利便/性”と形態素解析された．そして，“性”という語に“医学”というタグが付与されていた．このように，誤った形態素解析結果にタグが付与されたため，トピック推定を誤り，言語モデルの性能を低下させてしまったものがあった．

6　まとめ

ポッドキャストの音声認識率改善のための言語モデル適応手法として，ポッドキャストのRSSと学習テキストを比較し，ポッドキャストの話題に沿うような学習テキストを選択し，言語モデルを構築する手法を提案した．また，ポッドキャストのRSSの新規性や量の不足から生じるテキスト選択の精度低下の問題を解決するため，語彙情報サービスを用いた．実験を行い，ポッドキャストの話題に沿うような言語モデルを構築するためにRSSの情報を用いることの有用性を確認した．

謝辞　本研究は，経済産業省，平成18-20年度戦略的技術開発委託費「音声認識基盤技術の開発」の一部として実施されたものである．

参考文献

[1] 後藤真孝, 緒方淳, 江渡浩一郎, “PodCastleの提案: 音声認識研究2.0を目指して” 情処研報, 2007-SLP-65-7, 2007.

[2] 緒方淳, 後藤真孝, 江渡浩一郎,“PodCastleの実現:Web2.0に基づく音声認識性能の向上について” 情報処理学会研究報告, 2007-SLP-65-8, 2007.

[3] 緒方淳, 松原勇介, 後藤真孝, “PodCastle: 集合知に基づくWebキーワードを考慮した言語モデリング，” 日本音響学会2008年秋季研究発表会講演論文集, Sep 2008.

[4] T. Hofmann, “Probabilistic Latent Semantic Indexing”, In Proc. SIG-IR, 1999.

[5] Y.Akita and T.Kawahara, “Language Model Adaptation Based on PLSA of Topics and Speakers for Automatic Transcription of Panel Discussions”, IEICE Trans., Vol.E88-D, No.3, pp.439-445, 2005.

[6] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol3, pp.993-1022, 2003.

[7] 横山忠介, 篠崎隆宏, 岩野公司, 古井貞熙, “言語モデルのバッチ型教師なし適応化法”, 信学技報, SP2002-151, p.19-24, 2002.

[8] M.Suzuki, Y.Kajiura, A.Ito and S.Makino, “Unsupervised Language Model Adaptation Based on Automatic Text Collection from WWW”, In Proc. Interspeech, 2006.

[9] 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.

[10] Wikipedia,`http://ja.wikipedia.org/wiki/`.

[11] はてなキーワード, `http://d.hatena.ne.jp/keyword`.

[12] “日本語話し言葉コーパス，” `http://www.kokken.go.jp/katsudo/seika/corpus/`.

[13] “Yahoo!ニュース，” `http://headlines.yahoo.co.jp/`.

[14] “Palmkit，” `http://palmkit.sourceforge.net/`.