Web上の類似記事自動収集による音声認識用言語モデルの適応と学習用ニュース記事コーパスの分析 *

高　橋　伸　弥 **
森　元　　　逞 **
入　江　由　紀 ***

Analysis of Training Corpus Collected by Web Crawling for Speech Recognition of News Documents

Shin-ya TAKAHASHI, Tsuyoshi MORIMOTO and Yuki IRIE

This paper shows some results of speech recognition experiments for broadcast news using the language model adaptation proposed in our previous report. The basic idea of this method is that a broadcast news has similar Web documents on the Internet news site, so the perfomance of speech recognition for the broadcast news can be improved with the language model adapted to the news documents collected by Web crawling. To show the effectiveness of this approach, some experimental results are demonstrated. In addtion, the analysis of text corpora collected from Web site and the relationship between the performance of this method and the tendency of the text corpora are shown.

Key Words: Information Retrieval, Speech Recognition, Broadcast News, Web Crawling

1. はじめに

1.1 研究目的および背景

放送媒体の多様化・多チャンネル化により，視聴者に提供される映像量は年々増加している．このため，映像を蓄積するだけでなく，視聴者自身による検索を容易にするための技術が求められている．特にニュース映像は，その内容の重要性と利用価値の観点から，索引付きのデータベースとして保存する価値が高いと考えられており，テレビ局を中心に既に多くの試みがなされている．しかし，ニュース映像は日々大量に作り出されているため，人手で索引付けを行うのは非常に膨大なコストを必要とする．

これに対し，ニュース映像の音声データを音声認識し，その認識結果から索引語として適切な語を抽出する方法が提案されている[(1),(2),(3)]．この手法は，高速に索引語を抽出できる点で実用的ではあるが，音声認識の精度が問題となる．

この問題に対し，我々はこれまで，ニュース映像のトピックに合わせて言語モデルを動的に更新させることを検討してきた[(4),(5)]．これは，配信されたニュース映像と同一の情報源から作成されたと考えられる World Wide Web（WWW：以下，Web）上のニュース記事から，トピックに適応した言語モデルを作成し，これを用いて音声認識を行うことで，信頼できる索引語を抽出し，更にこの一連の処理を繰り返し行うことで，音声認識の高精度化を実現しようというものである．本稿では，前報告(5)で提案した手法が，ニューストピックの性質によってどのように性能面での影響を受けるか，継続的な実験を行うことによって確認した．また処理の高速化を目的として，記事収集方法および記事選出方法の簡略化

* 平成18年7月31日受付
** 電子情報工学科
*** 電子情報工学専攻

を行った．さらに，収集した類似記事テキストから作成される学習用コーパスの質を確認するために，実際に収集されたテキストと正解記事との再現率および適合率を調べ，その傾向が性能とどのような関係があるかを分析した．

1.2　関連研究

上述のような，Web 上のテキストを利用してドメインに適応した言語モデルを作成する方法は，これまでにも多数提案されている[(6),(7),(8),(9)]．例えば(7)では，適応させたいドメインに関連した発話を入力すると，その認識結果を検索語として，類似テキストを検索し，検索結果からドメイン固有の単語を学習することにより，未知語に対応した言語モデルを構築することを行っている．また(8)では，会話文を対象として言語モデルの適応を行うために，会話文コーパス内の文を検索語として，Web 上のテキストを収集することを行っている．さらに(9)では，医療に関連するウェブページを自動収集し，フィルタ処理を行って有効なテキストを抜き出し，ドメインに適応した言語モデルを作成している．これらは，適応対象のドメインに関する検索語を与えることで，Web 上からテキストを収集するものであるが，本研究のようにドメイン（トピック）が未知である場合に，適切な検索語を自動的に選択することは行っていない．

一方，本研究と同様，ニュース音声を対象として Web 上の新聞記事を利用する研究がある[(10)]．認識結果中から選択された索引語を用いて，Web 上の類似記事を自動収集する点で本研究と類似しているが，検索対象をニュース音声と同一日時のニュース記事としている点で言語モデルの適応に有利な反面，収集テキストの量が限られるという問題がある．これに対し，過去の類似記事／関連記事も対象として検索・収集し，より多くのテキストを用いて言語モデルを適応させる方法が考えられるが，Web 上の情報は日々更新されるので，類似記事検索を行った時期によって結果が異なる可能性がある．

また，(10)では，ニュース番組中の１つのトピックを対象とした実験を行っているため，トピック適応という点で効果が明らかでない．ニュース番組内のトピックは，何らかの事件・事故のような一時的に注目を浴びる内容のものから，数ヵ月から数年にわたって継続的に報道されるもの，また天気予報や為替など常に放送されるものなど，それぞれ異なる性質を持つと考えられる．

本稿では，異なる検索時期における評価実験を複数のトピックに対して行い，トピックの性質による効果の有無を検討する．更に，言語モデル適応の繰り返し処理の効果を検討するための評価実験を行い，その有効性と問題点を示す．

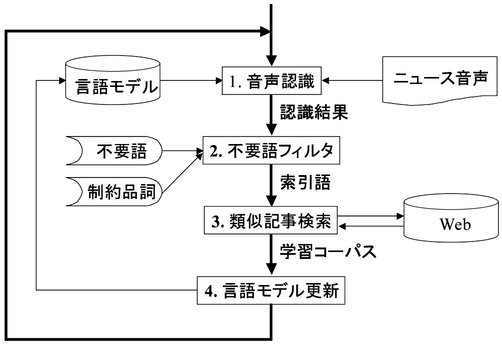

図1 索引語自動抽出システム

2.　索引語自動抽出システム

2.1　処理の流れ

索引語自動抽出システムの処理の流れを図１に示す．このシステムは，

１．索引語抽出対象となるニュース音声を汎用言語モデルを使用して音声認識器※で認識する

２．認識結果から索引語として適切でない語(不要語[(11)]や品詞（制約品詞）を不要語フィルタで除去し，索引語を得る

３．その索引語を検索質問として，ニューストピックに類似した記事を Web 上で検索し，収集する

４．収集した記事を学習コーパスとし，汎用言語モデルをトピックに適応した言語モデルへと更新させる

５．トピックに適応した言語モデルを用いて，再び同一のニュース音声を認識する

という処理を索引語が収束するまで繰り返し，収束後の索引語をニュース映像の索引語として抽出するものである．

2.2　類似記事の収集

類似記事を収集するための検索方法としては，音声認識結果の仮説の中から索引語候補単語を抜き出し，さらにその中から出現頻度上位５位までの単語を選択して，それらの論理和を既存の検索エンジンへの入力（検索条件）とした．検索語が多すぎると既存の検索エンジンでは十分な量の結果が得られず，また逆に検索語が少なすぎると無関係なページまで大量に検索してしまうことから，ここでは上位５位までの単語を使用している．また，収集記事と認識結果の間の類似度計算においては，記事および認識結果から抽出した索引語候補単語集合の間で，共通部分集合の要素数を和集合の要素数で除したものを類似度とした．ここで，単語の頻度による重みを考慮す

※日本語大語彙音声認識エンジン Julius(12)を使用

表1 放送されたニューストピック

トピック		発話時間	単語数 (索引語候補数)	話者数	地域 依存性	時期 依存性	
1	横田めぐみさん夫DNA 鑑定結果	77 秒	196（ 63）	3	全国	継続的	韓国語での電話インタビューを挟む
2	低気圧影響広範囲で激しい雨	78 秒	268（ 94）	1	全国	一時的	背景雑音（雨音）あり
3	テレビ局元社員横領無罪判決	110 秒	353（119）	1	全国	継続的	
4	原子炉流量計データ改ざん	73 秒	251（ 72）	1	全国	継続的	
5	気象情報（全国）	46 秒	126（ 34）	1	全国	—	
6	為替と株	24 秒	71（ 44）	1	全国	—	
7	鉄塔土台崩れ住民避難	90 秒	287（ 90）	2	地方	一時的	街頭インタビュー挟む
8	高速船衝突事故	100 秒	322（109）	1	地方	継続的	
9	気象情報（九州）	101 秒	339（101）	1	地方	—	

るために，索引語候補単語集合は単語の重複を許すものとした．

前報告(5)では，TF／IDF で重みづけされたベクトル空間モデル[13]を使用して，収集した記事の類似度を求めることを行ったが，対象とする全ての記事内の全単語に対してあらかじめ TF／IDF 値を計算しておく必要があることから，計算に非常に時間がかかるという問題点があった．そのため，ここでは簡易的な方法として上記のような類似度計算を採用することとした．また，記事の選択方法についても，前報告(5)では，類似度しきい値をあらかじめ実験的に求めておき，しきい値以上の記事を学習コーパスとしたが，ここでは認識結果との類似度が高い記事の上位100記事を学習コーパスとした．

2.3　言語モデルの更新

言語モデルの更新方法としては，既存のコーパスと適応対象のコーパスとを結合する方法や,既存の言語モデルと適応対象のコーパスから得られた言語モデルとを融合させる方法など様々な方法が提案されている[14],[15],[16]．提案システムにおいては，繰り返し処理を行うごとに多量のテキストコーパスが得られることを考慮して，コーパスを結合する方法で言語モデルを更新することとし，結合の際には単純に以前のコーパスに新しく収集したコーパスを追加（累積）していく方法をとる．なお繰り返し処理の１回目の更新においては，既存の言語モデルとの融合を行わず，収集したコーパスから新たに言語モデルを作成することとした．

3.　評価実験

3.1　実験条件

実験には，2006年４月12日の15時から15時15分に放送されたＮＨＫのニュース映像を用いた．このニュース映像の音声データをトピックごとに wav 形式で保存し，テストデータとした．表１に放送されたニューストピックを示す．表には，各トピックの放送時間と話者数，記事原稿の延べ単語数，延べ索引語候補数※※および地域依存性／時期依存性の分類を併せて示している．

このニュース番組では，前半（トピック１から６）は全国ニュース，後半は地方ニュースとなっている（地域依存性）．更に，各ニューストピックの内容から，継続的にある程度の期間にわたって報道がなされているもの，１日から数日程度の一時的なもの，どちらでもないものが含まれている（時期依存性）．

学習用テキストは，新聞社のウェブサイトを指定して，既存の検索エンジン※※※を用いて収集した．指定した新聞社は，朝日，毎日，読売，産業経済，日経，東京，西日本，京都，中日，中国新聞社，河北新報社の11社である．検索は放送翌日から２週間後までの毎日行い，更に１ヶ月後および２ヶ月後にも検索を行った．

学習開始時に用いる汎用言語モデルには，Julius ディクテーションキット Ver.3.1付属の Web から学習した６万語の言語モデルを用いた[12]．１回目の学習以降で作成するニュース適応言語モデルはバイグラム言語モデルとした．言語モデルの作成には，統計的言語モデルの作成キットである“CMU-Cambridge SLM Toolkit”[17] を用いた．なお音響モデルには Julius ディクテーションキット Ver.3.1付属の性別非依存モデルを用いている．

※※ここでは，一般名詞，固有名詞，サ変名詞のみを索引語候補としている．

※※※使用した検索エンジンは Google（http://www.google.com/）である．検索キーに site：を付加して新聞社ウェブサイトを指定した．なお収集の際には，検索エンジンのキャッシュデータも用いている．

実験結果の評価尺度としては，品詞制約および不要語フィルタを施したあとの索引語候補に対し，以下の式で計算される再現率，適合率およびノイズ率を用いた．

$$再現率=\frac{正しく抽出できた上位索引語の延べ数}{正解文中の全索引語候補延べ数} \qquad (1)$$

$$適合率=\frac{正しく抽出できた全索引語候補延べ数}{抽出された全索引語候補延べ数} \qquad (2)$$

$$ノイズ率=1-適合率 \qquad (3)$$

ここで，再現率の計算においては，認識結果中の索引語候補のうち出現頻度上位10位内の単語のみを対象とした．実験で用いた不要語および制約品詞は表２に示す通りである．

表2 不要語と制約品詞

不 要 語	こと，人，話，他，発表
制約品詞	一般名詞，固有名詞，サ変接続名詞以外の品詞

3.2　実験結果

3.2.1　検索時期による性能評価

放送翌日から２ヵ月後までの間に類似記事検索を行い，各トピックに適応させた言語モデルの性能がどのように変化するかを調べた．図２に，繰り返し１回目の適応処理で得られた言語モデルによる索引語の再現率の変化を示す．

グラフから，ほとんどのトピックに対し，数日から１週間程度では再現率にほぼ変化が無いことが分かる．この理由としては，今回の実験で使用した既存の検索エンジン自体に記事収集の遅延が生じていることが挙げられる[※※※※]．また他の理由として，一時的なニュースや地方版のニュースの場合には，もともと記事が少ないため時間的変化の影響を受けにくいことが考えられる．

日数が経つにつれ再現率が若干向上しているものは，継続的に続報が Web 上に掲載され，時間が経つにつれ類似記事が増加しているケースであると推測される．また，１ヵ月以上経つと再現率が低下しているトピックが多く見られたことから，数日から数週間の範囲でニュース記事検索を行うのが効果的であると考えられる．長期間にわたり継続的に関連したニューストピックとして表れるケースもあると予想されるので，数ヵ月もしくは年単位での変化についても調査する必要がある．

3.2.2　トピックごとの性能評価

次に，言語モデル適応処理を繰り返し行った場合の実験結果を図３に示す．ここで，繰り返し回数は５回とし，類似記事検索はニュース番組放送の２週間後に行った．“□”は音声認識結果の単語正解率，“○”は索引語再現率，“△”は索引語ノイズ率を表している．また棒グラフでコーパスサイズも併せて示した．グラフから見て分かるように，トピック７を除いた全てのトピックで索引語再現率を向上させることができている．また単語正解率の変化が横ばいであっても，索引語再現率が向上してるケース（トピック1,2,5）が見られた．このことから，ニュースを特徴付ける高頻出の単語を含むようなニュース記事をうまく収集出来ていると考えられる．

※※※※Google では，最新のニュース記事が反映されるまでに短くて１日，長くても２，３日ほどの遅延が見られた．

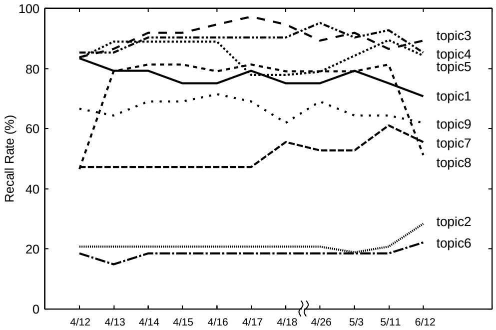

図2 検索時期による索引語再現率の変化

トピック７で改善が見られなかった理由としては，音声の品質の悪さが考えられる．実際に，トピック７に含まれている街頭インタビューの区間（18秒間）を取り除いて認識実験を行ってみると単語正解率が 68.4%となり，インタビュー区間を含む場合の59.9%と比べ，約８%の性能劣化を引き起こしていることが分かった．また，その内容が特に地方色の強いニュースであり，全国区の新聞社サイトでは対象の記事が見つからなかった可能性も原因として考えられるが，繰り返し２回目以降で若干改善されていることから，本手法によりある程度，音声品質の悪さから来る性能劣化をカバーできていると言える．

トピック２に関しても同様に，音声データ全体を通して風雨の音などの雑音が混入していることから，品質の悪さが性能改善の見られなかった原因として考えられる．さらに「低気圧の影響による広範囲かつ非常に激しい降雨」という全国的なニュースであったにも関わらず，その日の午後の状況という一時的（瞬間的）なニュースであったため，該当するニュース記事が検索できなかったこと可能性も原因として挙げられる．

トピック６に関しては，もともとの音声データが短いものであり，かつ数詞を多く含む内容であったために，初回認識時の誤認識が類似記事収集時に悪影響を及ぼし

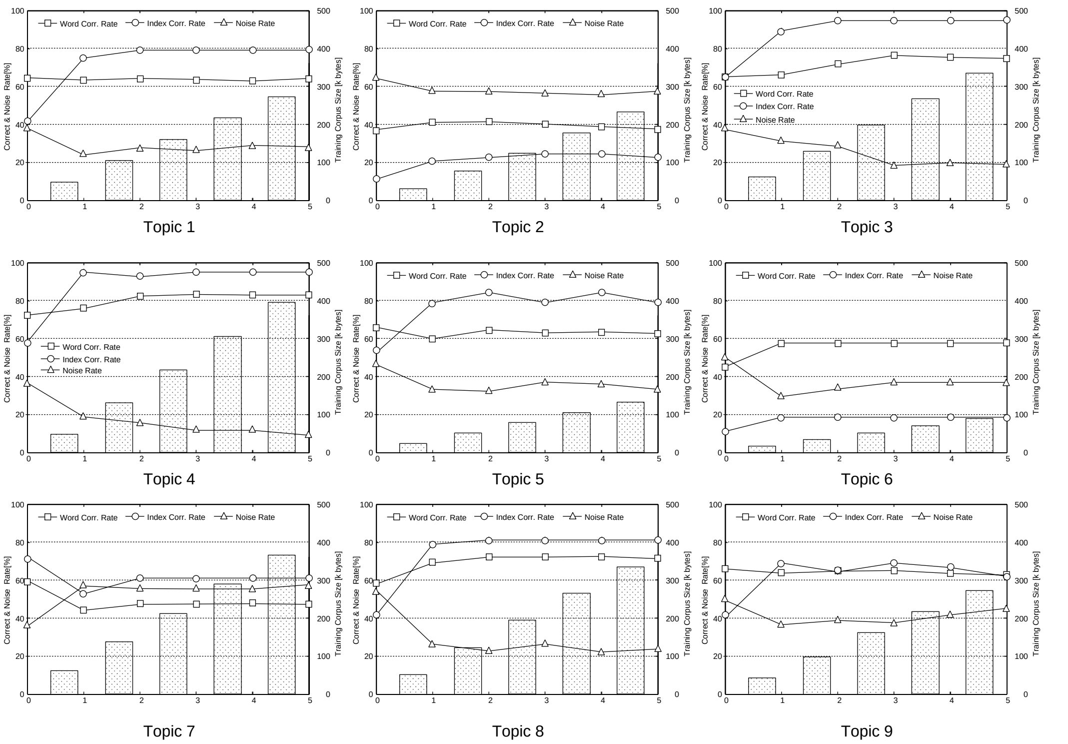

図3 実験結果（２週間後）

たと思われる．

言語モデルの適応を繰り返し行うことにより，再現率を向上させることができたのは，トピック全体のうち半数程度であった．全体として３，４回の繰り返しで再現率がほぼ収束していることから，今回の記事収集において記事数を固定にしたために，十分な量の記事を収集できなかった可能性が考えられる．これに対し，類似度がしきい値以上の記事を学習対象として，記事数を制限しない方法が考えられるが，記事の選定方法，類似度の計算方法と併せて今後の課題である．

3.2.3　収集したテキストコーパスの分析

次に，前述のようなトピックごとの性能の違いが，収集された類似テキストの質に依存していることを考慮し，それぞれのテキストコーパスに対して分析を行った．対象としたテキストコーパスは，繰り返し処理１回目すなわち既存の言語モデルによる認識結果を用いて収集した100記事である（収集時期は２週間後）．

表３にトピックごとに適応した言語モデルの比較結果を示す．表に示したテストセットパープレキシティは，正解文をテストセットとして CMU Toolkit の言語モデル評価ツール evallm で計算した．また認識性能の比較のために，図３で示した索引語候補再現率と，既存の言語モデルを使用した結果に対する改善率を併せて示している．表からトピック６に関しては，十分な量の記事が収集出来てないこと，また作成された言語モデルの性能が良くないことが確認できるが，同じく再現率が低かったトピック２および性能改善が見られなかったトピック７に関しては，言語モデルの評価からだけでは違いがはっきりしない．

そこで次に，どれくらい正解文と類似した記事が収集出来ているかを評価するために，正解文に対する再現率および適合率を計算した．結果を，表４に示す．ここで再現率および適合率の計算は，正解文および収集記事内の索引語候補単語全てを対象とした．表から分かるように，初期認識時に誤認識単語を多く含んでいたトピック２とトピック６は収集した記事の再現率，適合率が非常に低くなっていることが分かる．特にトピック６では分散が非常に大きいことから，収集した記事がトピックに特化したものとなっていない．また，再現率，適合率ともに最小値 0.0%の記事が含まれていることから，認識結果の文章で検索した場合に，誤認識の影響により全く異なるトピックの記事も収集してしまっていることが分かる．再現率，適合率ともに0.0%の記事は，トピック２で11%，トピック６で20%，トピック９で23%であっ

表3 トピック適応言語モデルの比較

トピック番号	語彙数	コーパスサイズ(Kbytes)	テストセットパープレキシティ	索引語再現率(改善率)
1	1510	48.3	9.8	75.0%（+33.3%）
2	2037	31.2	7.4	20.8%（+ 9.4%）
3	2949	62.6	7.8	89.2%（+24.3%）
4	2483	47.4	8.0	95.1%（+36.6%）
5	1717	24.5	6.5	79.0%（+26.3%）
6	897	16.7	10.7	18.5%（+ 7.4%）
7	2760	62.7	7.5	52.8%（−19.4%）
8	2547	52.1	8.3	79.1%（+37.2%）
9	2668	42.9	6.2	69.1%（+28.6%）

表4 収集したテキストの正解文に対する再現率・適合率

トピック番号	索引語候補平均数	再現率				適合率			
		平均	分散	最大値	最小値	平均	分散	最大値	最小値
1	46	21.5%	157.8	58.7%	3.2%	33.5%	225.6	83.3%	7.3%
2	25	2.9%	9.5	19.1%	0.0%	12.2%	91.5	42.9%	0.0%
3	59	7.0%	9.6	16.8%	2.5%	16.1%	40.8	36.4%	6.6%
4	40	8.9%	66.6	55.7%	2.9%	18.1%	123.3	60.0%	4.2%
5	19	8.7%	34.3	29.4%	0.0%	18.2%	92.7	50.0%	0.0%
6	13	9.6%	104.4	29.5%	0.0%	31.4%	925.4	100.0%	0.0%
7	57	11.3%	18.4	24.4%	4.4%	21.2%	60.9	41.7%	8.1%
8	45	6.1%	29.6	37.0%	1.9%	17.2%	136.6	60.0%	3.8%
9	32	3.2%	9.3	13.9%	0.0%	12.3%	121.7	50.0%	0.0%

た．このうちトピック９に関しては，初期認識時の認識率がトピック２や６と比べて若干高く分散が小さいことから，全く適合しない記事が含まれていたにも関わらず，内容的に似た記事が収集できたことで性能が改善できたと思われる．

さらに収集記事の分布傾向を見るために，再現率(recall)，適合率（precision）のヒストグラムを求めた(図４)．ヒストグラムを見ると，改善率が高いグループ（トピック3,4,5など）は似たような分布になっていることがわかる．またトピック1,4については，再現率５%以下の記事が少ないこと，50%超の記事があることなどが共通している．

以上より，収集した記事群に表れる索引語候補の出現分布から，どの程度トピックに特化した記事を収集できたかを評価することで，さらにトピック適応の精度を向上させることができると思われる．実際の適応処理中には，正解文との類似度を計算することは出来ないため，収集したテキスト群に対してクラスタリングを行うなどの処理を行う必要があるだろう．

4. おわりに

本論文では，Web 上の類似記事を利用してニューストピック適応言語モデルを得る処理を再帰的に行うシステムの性能評価を行った．異なる検索時期における性能評価を行った結果，ニュース放映直後ではあまり大きな効果は得られなかったが，数日後から１ヵ月の間では，性能の改善が見られた．月単位もしくは年単位の検索時期のずれに関しては今後引続き検証して行く必要があるが，継続的なニュースに関しては同等の性能を得ることが出来ると考えている．更に，繰り返し言語モデル適応を行う評価実験を行った結果，再現率に対し，既存の言語モデルを使用する場合に比して，最大で約 40%，繰り返しを行わない場合に比して，最大で約５%の改善を得ることが出来た．また，トピックの性質，すなわち一時的なニュースか継続的なニュースか，また全国ニュースか地方ニュースかによって，索引語抽出精度の改善率に違いが表れることを示した．次に，収集したテキストコーパスの質を評価するため，正解文と収集記事との間で再現率および適合率を求め，ヒストグラムにより分布傾向を調べた．その結果，性能改善を実現したコーパスに関しては，高い再現率の記事が含まれること，再現率の低過ぎる記事すなわちトピックに無関係な記事が少ないことを確認した．また性能改善が見られなかったコーパスに関しては，正解文に全く適合しない記事が多く含まれていたこと，また分散が非常に大きいことが分かった．今後は，本報告で示した問題点，結論を考慮して，

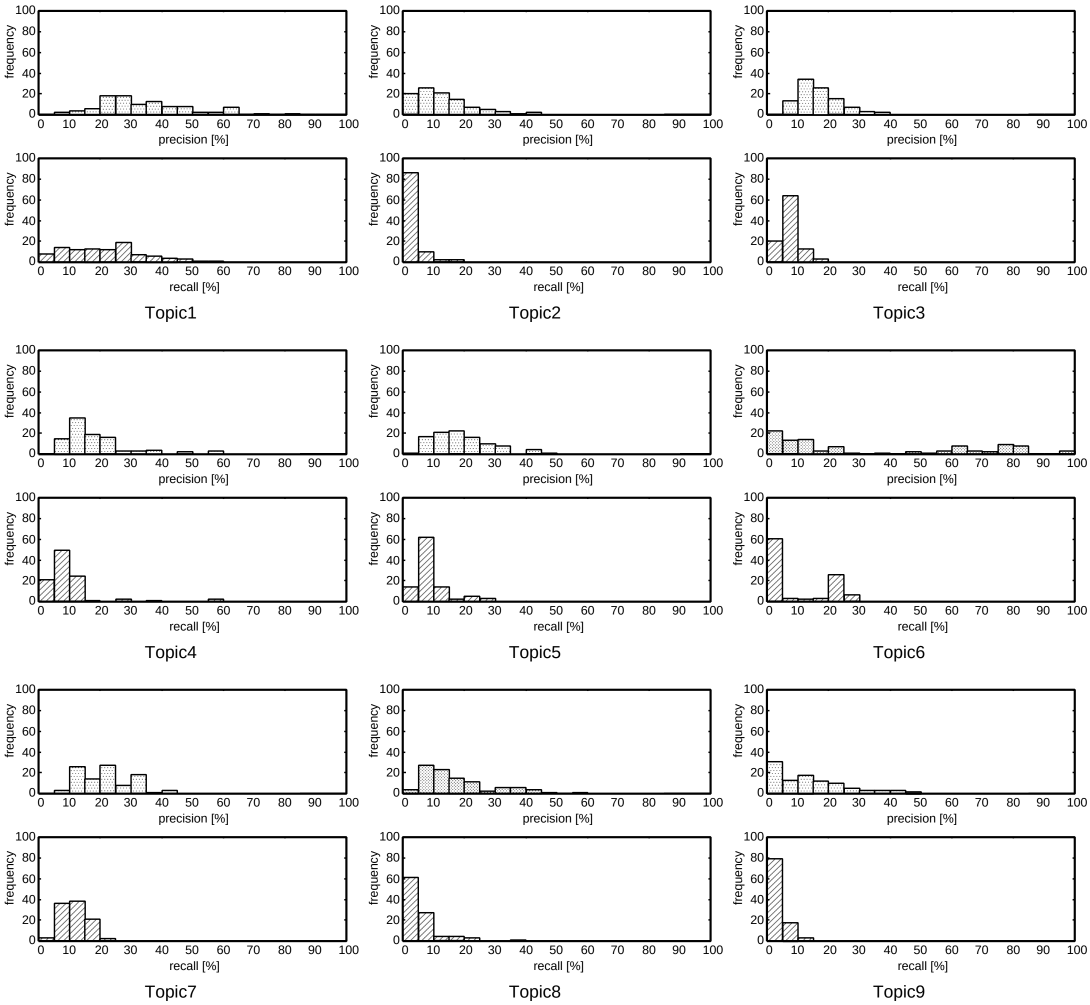

図4 正解文に対する再現率・適合率の分布

更に高精度な言語モデル適応手法について検討したい．

参 考 文 献

(1) D. Abberley, S.Renalas, and G.Cook, “Retrieval of broadcast news documents with the THISL system,” Proc. ICASSP'98, pp.3781-3784(1998).

(2) 西崎，中川：“音声キーワードによるニュース音声データベース検索手法”，情処学論，Vol.42，No.12, pp.3173-3184(2001).

(3) 西崎，中川：“音声認識誤りと未知語に頑健な音声文書検索手法”，信学論 D-Ⅱ，Vol.J86-D-Ⅱ，N0.10, pp.1369-1381(2003).

(4) 高井，森元，高橋：“Web 上の動画ニュース検索のための索引語抽出”，電気関係学会九州支部第56回連合大会講演論文集(2003)

(5) 高橋，高井，森元：“ニュース映像検索システムのための索引語の自動抽出”，福岡大学工学集報，No.76, pp.15-22(2006).

(6) X. Zhu and R. Rosenfield, “Improving trigram language modeling with the world wide web”, Proc. of ICASSP’01(2001).

(7) A. Berger and R. Miller，“Just-in-time language modeling”, Proc. of ICASSP’98(1998).

(8) I. Bulyko，M.Ostendorf and A.Stolcke，“Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures”, Proc of HLT-ACL, 2003.

(9) 西村 他：“Web からの音声認識用言語モデル自動

生成ツールの開発”，情処研報　SLP-35-8，pp.49-54(2005)．
(10) 伊藤，西崎，関口：“Web 上の類似記事を利用した音声文書の認識性能の改善”，情処研報 SLP-59-9, pp.49-54(2005)．
(11) 徳永：“情報検索と言語処理”，東京大学出版会(1999)．
(12) httP://julius.sourceforge.jp/
(13) G. Salton *et. al.*: “A vector space model for automatic indexing”, Communications of the ACM, Vo1.18, No.11, pp.613-620, 1975. Reprinted in Readings in Information Retrieval, Jones, K.S. and Willett, P.(Eds.), Morgan Kaufmann Publishers, pp.273-280(1997).
(14) 政瀧 他：“最大事後確率推定による N-gram 言語モデルのタスク適応”，信学論 D-Ⅱ，Vol.J81-D-Ⅱ, No.11, pp.2519-2525(1998)．
(15) 長友 他：“相補的バックオフを用いた言語モデル融合ツールの構築”，情処研報 SLP-35-9，pp.49-54, (2001)．
(16) 広瀬，嶺松，森谷：“単語間の関連性を利用した音声認識用言語モデルのドメイン適応”，情処論 Vol.43, No.7，pp.2065-2074(2002)．
(17) P. R. Clarkson and R. Rosenfeld: “Statistical Language Modeling Using the CMU-Cambridge Toolkit”, *Proc. ESCA Eurospeech*, pp.2707-2710 (1997).