G_013

Web上の類似記事自動収集によるニューストピック適応言語モデル

Adaptation of Language Model with Iterative Web Crawling for Speech Recognition of Broadcast News

高橋 伸弥 †　森元 逞 †　入江由紀 ‡
Shin-ya Takahashi　Tsuyoshi Morimoto　Yuki Irie

1 はじめに

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。

このような、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]。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、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 [9]。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酷似しているが、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、[9] 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。

そこで本論文では、(1) 自動的に収集した Web 上の類似記事を用いて言語モデルを適応させる処理を繰り返し行うシステムを提案し、(2) 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。(3) 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。

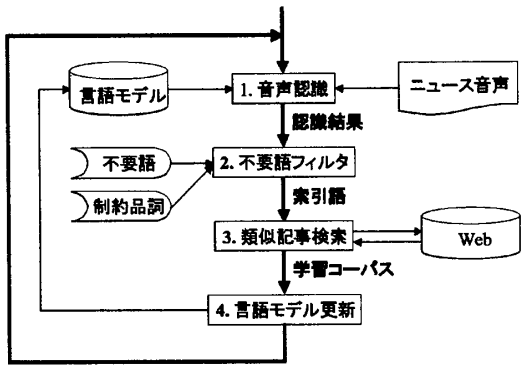

図1　索引語自動抽出システム

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*1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 [10]) 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

† 福岡大学工学部，Dept. EECS, Fukuoka Univ.
‡ 福岡大学大学院工学研究科，Grad. School, Fukuoka Univ.

*1 日本語大語彙音声認識エンジン Julius [11] 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	–	

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、出現頻度上位5位までの単語を選択し、それらの論理和を検索条件として既存の検索エンジンに入力する方法を用いた。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重複を許して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。情報検索等における類似度計算では、TF/IDFで重みづけされたベクトル空間モデル [12] がよく用いられているが、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している。また、記事の選択方法についても、類似度しきい値を適当に設定したり、クラスタリングを行って最大クラスタ内の記事を選択したりするなど様々な方法が考えられるが、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 [13],[14],[15]。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[*2]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1～6) 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～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[*3]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3週間後、1ヵ月後、2ヵ月後にも行った。

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 [11]。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"[16] 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。

実験結果の評価尺度としては、品詞制約および不要語

[*2] 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。

[*3] 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。

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

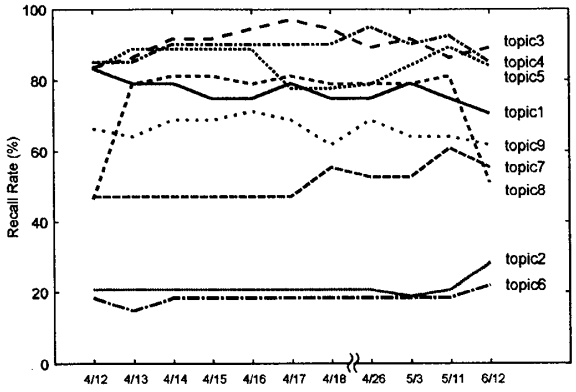

図２　検索時期による再現率の変化

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およびノイズ率を用いた。

$$\text{再現率} = \frac{\text{正しく抽出できた索引語延べ数}}{\text{正解の索引語延べ数}} \quad (1)$$

$$\text{ノイズ率} = \frac{\text{誤って抽出された索引語延べ数}}{\text{抽出された全索引語延べ数}} = 1-\text{適合率} \quad (2)$$

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に示した通りである。

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、数日から一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[*4]。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が、今後の課題としたい。

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、“○”は再現率、“△”は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、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。

トピック７で改善が見られなかった理由としては、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ためではないかと考えられる。また単語正解率に改善がなく、再現率も低いトピック２に関しては、「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が原因として考えられる。

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。

4　おわりに

本論文では、ニュース音声への高精度な索引語付けを自動的に行うことを目的として、音声認識結果から得られた索引語候補を用いてWeb上の類似記事を検索・収集し、ニューストピックに言語モデルを適応させる処理を繰り返し行うシステムを提案した。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。

今後の課題としては、(1)類似記事検索アルゴリズムの改良、(2)言語モデル更新方法の検討が挙げられる。

[*4] 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。

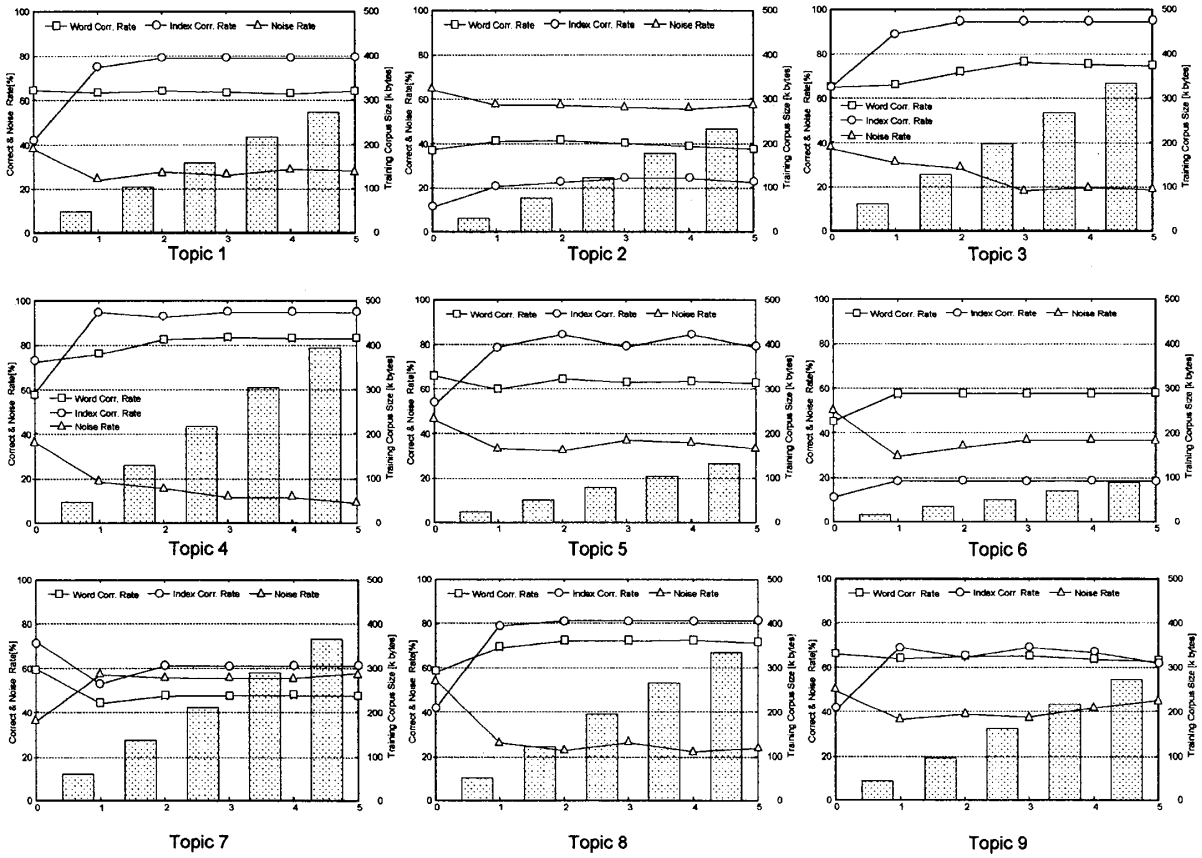

図3　実験結果

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, No.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] A. Berger and R. Miller, "Just-in-time language modeling", Proc. of ICASSP'98 (1998).

[7] 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.

[8] 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).

[9] 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).

[10] 徳永："情報検索と言語処理", 東京大学出版会 (1999).

[11] http://julius.sourceforge.jp/

[12] G. Salton *et. al.*: "A vector space model for automatic indexing", Communications of the ACM, Vol.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).

[13] 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).

[14] 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).

[15] 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).

[16] P.R.Clarkson and R.Rosenfeld: "Statistical Language Modeling Using the CMU-Cambridge Toolkit", *Proc. ESCA Eurospeech*, pp.2707-2710 (1997).