記者原稿を利用した ニュース音声認識のための言語モデル

加藤　直人　　浦谷　則好　　江原　暉将
NHK放送技術研究所
E-mail: {katonao, uratani, eharate}@strl.nhk.or.jp

1　はじめに

聴覚障害者などからテレビ番組，特に「ニュース」の音声を字幕化してほしいという要望が強い．米国では人手でリアルタイムに入力しているが，日本語の場合，漢字に変換する作業等が必要なためリアルタイムで入力するのは熟練した人でも難しい．そこで，このような作業を自動的に行うことを目標に，音声認識技術を利用した字幕作成の研究を行っている[今井98]．システムとしては

○音声認識によるニュース音声の文字化
○人手によるチェック・修正

というプロセスが必要となる（図1）．

音声認識ための言語モデルを構築するには電子化されたコーパスが欠かせない．NHKのニュースではアナウンサーが読む原稿（アナウンス原稿）は，記者がワープロで書いた原稿（記者原稿）を元に，手書きの修正が放送の直前まで行われる．したがってアナウンス原稿は電子化されていない．しかし，手書きの修正は節の入れ換え程度である場合も多く，アナウンス原稿と記者原稿では一致している個所も少なくない．そこで現在，記者原稿を使って言語モデルを構築している．アナウンス原稿に出現する単語は直近の記者原稿に出現する可能性が高いので，直近の記者原稿に比重をかけ音声認識用の言語モデル（bi-gram, tri-gramモデル）を構築している[小林98]．

音声認識の精度向上のためには言語モデルのさらなる改良が必要である．1つの方法はアナウンス原稿と記者原稿とで一致している個所（20～30語の連続する単語に及ぶ場合もある）を利用することである．具体的には，n≧4以上のn-gramモデルや，可変n-gramモデル[政瀧98]を使うことが考えられる．しかし，前者ではn＝4，5，…，20，…のn-gramという莫大なデータを計算機上に記憶しておかなければならず現実的ではない．また，このようなn-gramは出現頻度が小さい（ほとんどの場合は出現頻度1回）ので，出現頻度が高い定型表現を対象としている後者の方法はあまり有効でない．

本稿ではそれほどデータ量を増やすことなく，出現頻度が1回のものにも対応する，n≧4のn-gramを利用する言語モデルについて述べる．ここで中心的役割をするのが**単語出現位置辞書**である．また，単語出現位置辞書は元となる記者原稿を特定することができるので，「人手によりチェック・修正」をする際の参考となる記者原稿を出力することも可能となる．

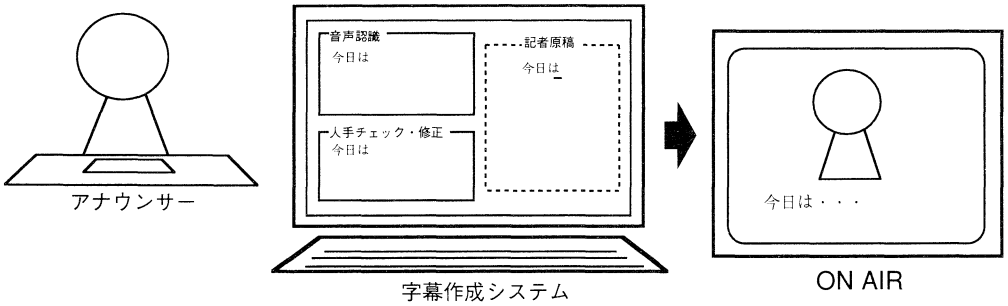

図1　字幕作成システムの一例

日付：1996年07月02日
番組名：７時のニュース
－－－－－－－－－－－－－－－－－－－－－－－－－
アジア太平洋地域で子どもの人権を守る為、メディアがどのような役割を果たしていくのかを話し合う、初めての国際会議がフィリピンのマニラで始まりました。
えこの会議はＡＢＵアジア太平洋放送連合やＵＮＩＣＥＦ国連児童基金などが主催して開いたもので、アジア太平洋地域の二十二の国と地域から、テレビや新聞広告などのメディア関係者らおよそ三百人が一堂に会して、子どもの人権とメディアの役割を話し合う初めての試みです。
初日のきょうは子ども達の歌や踊りを交えた開会式のあと、フィリピンのリナライゴ社会福祉相が、子ども達にとって必要なメディアのあり方を探らなければならないとあいさつし、続いて各国の子ども達のメディアへのさまざまな意見が会場のモニター画面で紹介されました。
会議では、子どもに対する虐待など人権侵害の問題や、アジア太平洋地域のテレビメディア間の番組の制作協力などについても討議し、最終日の五日には、子どもの為のアジアメディア宣言を採択して閉幕することになっています。
＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

書き起こし原稿

タイトル：子供人権メディアサミット
日付：1996年07月02日
作成部：国際
－－－－－－－－－－－－－－－－－－－－－－－－－
アジア、太平洋地域で子供の人権を守るためメディアがどのような役割を果たして行くかを話し合う初めての国際会議がフィリピンのマニラで始まりました。
「子供の人権とメディアに関するアジアサミット」と名付けられたこの会議はＡＢＵ・アジア太平洋放送連合やユニセフ・国連児童基金などが主催してアジア・太平洋地域の二十二の国と地域からテレビや新聞、広告などのメディア関係者らおよそ三百人が一堂に会して子供の人権とメディアの役割を話し合う初めての試みです。
初日のきょうは子供達の歌や踊りを交えた開会式のあとフィリピンのリナ・ライゴ社会福祉相が「子供達にとって必要なメディアのあり方を探らなければならない」と挨拶し、続いて各国の子供達のメディアへの様々な意見が会場のモニター画面で紹介されました。
会議では子供に対する虐待など人権侵害の問題やアジア・太平洋地域のテレビメディア間の番組の制作協力などについても討議し最終日の五日には「子供のためのアジアメディア宣言」を採択して閉幕する事になっています。
＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

記者原稿

図２　書き起こし原稿と記者原稿の例

2　書き起こし原稿と記者原稿

書き起こし原稿と，その元となっていると思われる記者原稿の例を図２示す．ここで，「書き起こし原稿」とは，実際のニュースを人手で書き起こしたものを指す．したがって，アナウンス原稿とほとんどかわらないが，アナウンサーが話した言いよどみ（「え」）等も含まれる．一方，記者原稿はその記事を作成する部署によって７つのジャンル（政治，経済，社会，国際，スポーツ，ネット，首都圏）に分かれている（図２の場合は国際）．１日の記事数は日によって異なるが，１ジャンルあたり２０～３０記事であり，１記事あたりの単語数は３００～４００語程である．

図２を見ると，書き起こし原稿も記者原稿も１文あたり５０単語程含まれており，１文が非常に長いことがわかる．書き起こし原稿と記者原稿を比較すると，記号（「,」,「・」等）を除けば４単語連続以上で一致している単語列がかなりあり，２０単語を超える場合もある．

3　記者原稿を利用した言語モデル

記者原稿を利用した言語モデルでは，コーパスから言語モデルを構築する際に，従来の tri-gram（または bi-gram）とともに，**単語出現位置辞書**を自動的に生成する．音声認識の際には単語出現位置を使うことにより，出現頻度が１回，$n \geqq 4$ の n-gram の言語的制約をすることが可能となる．

3．1　単語出現位置辞書

単語出現位置辞書を自動生成するコーパスとしては，直近（前日，当日）の記者原稿を使う．この記者原稿を形態素解析して単語に分割し，それぞれの単語に対して出現した位置（単語出現位置）を収集する（図３）．単語出現位置の番号はジャンル，記事の出現位置，単語の出現位置の３つ情報によって次のように７桁の数で定義している．

【単語出現位置の番号の定義】

単語出現位置の番号 XYYYZZZ（７桁）

X　（上位１桁目）：ジャンル番号

（1：政治，2：経済，3：社会，4：国際，5：スポーツ，6：ネット，7：首都圏）

YYY（２～４桁目）：そのジャンルにおける記事位置

ZZZ（５～７桁目）：その記事における単語出現位置

例えば，単語出現位置の番号 6002021 は，ジャンルがネット（6），記事の出現位置が２番目（002），単語の出現位置２１番目（021）ということを表す．

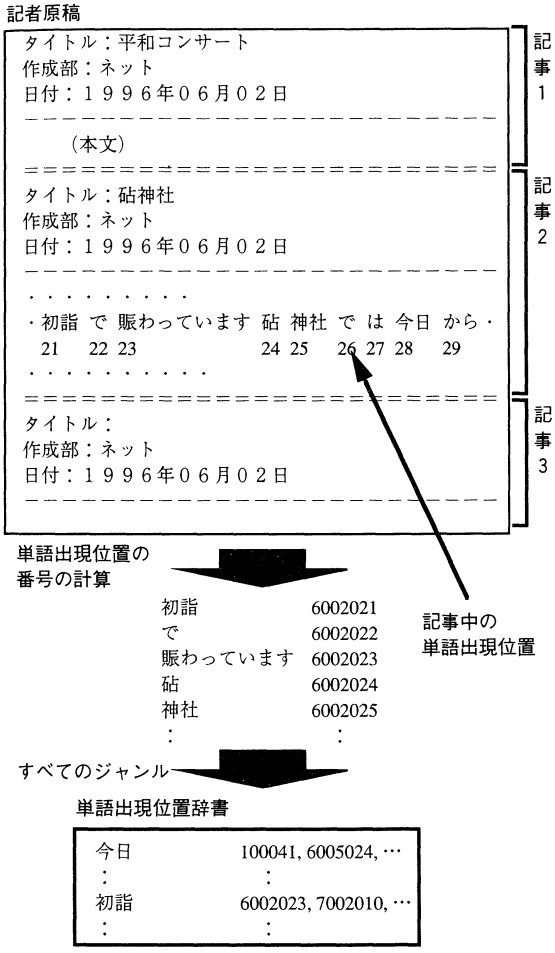

図３　単語出現位置辞書の作成

３．２　音声認識での利用

音声認識では，単語出現位置辞書によって各認識候補の単語出現位置の番号を求め，この数字が連続している単語列を優先する．

具体的には次の３つの利用方法を考えている．

(a) tri-gram による言語制約後の利用

(b) tri-gram との併用による利用

(c) デコード上で縦型探索としての利用

以下でそれぞれの方法を簡単に説明する．

(a) tri-gram による言語制約後の利用

tri-gram による言語制約の後に，スコアの再計算として単語位置辞書を使う．

マルチパスサーチの第１パスにおいてtri-gramを使い，ある程度認識候補を絞る．ただし，ここで用いる tri-gram は直近の記者原稿はもとより，それ以外の記者原稿も使って構築する．

第２パスにおいて，単語出現位置辞書を使い，単語出現位置の番号が連続する単語列にはスコアを加える．例えば，４単語が連続して場合に４点加算するとし，認識候補の単語が「初詣で賑わっています砧」であったとする．図３の単語出現位置辞書を使うと「初詣（6002021）」，「で（602022）」，「賑わっています（602023）」，「砧（602024）」と単語出現位置の番号が４つ連続しているので，この単語列には４点加算される．この際，「で」はコーパス中に何度も出現するので，単語出現位置の番号がさまざま得られるであろうが，前後の単語と連続している番号を選ぶ．

最後に最もスコアが高い経路を求める．

(b) tri-gram との併用による利用

tri-gram による言語制約をする際に，単語位置辞書を併用する．

それまでに得られている経路の最後の単語を w^i，それまでに連続している n-gram の長さを N（すなわち N-gram を利用），単語出現位置を c とおく．次単語の認識候補 w^{i+1}_k $(k = 1, 2, \cdots, k_0)$ の確率を p'_k を，tri-gram による確率 p_k に基づいて式（１）で動的に再計算する．

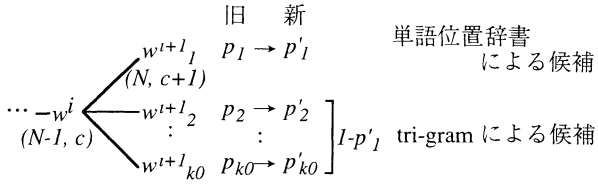

図４　確率値の動的再計算

確率の動的再計算

$$\begin{cases} p'_1 = q_0 \times (1 - \exp^{-\lambda \times N}) & (1\,a) \\ p'_k = (1 - p'_1) \times \dfrac{p_k}{\sum_{k=2}^{k0} p_k} \quad (2 \leqq k \leqq k0) & (1\,b) \end{cases}$$

（q_0，λは定数，$0 \leqq q_0 \leqq 1$，$\lambda > 0$）

式（１ａ）は，前単語の単語出現位置cから予測される単語w^{i+1}_1（$k=1$のときとする）の場合であり，その新たな確率p'_1（確率値）をNに応じて計算する．式（１ｂ）は，tri-gramから得られる認識候補w^{i+1}_k（それを$2 \leqq k \leqq k0$のときとする）の場合であり，tri-gramの確率p_kに応じて残りの$1-p'_1$を分配する．

ただし，これらの計算はあらかじめ決めておいたN_0に対して，$N \geqq N_0$のときのみ実行する．例えば，$N_0 = 4$とすれば，このような計算が行われるのは4-gram以上の場合である．

(c) デコードの縦型探索の利用

縦型探索のトリガーとして利用する．

通常はデコードは横型探索をし，前述の(c)の処理が始まった後は縦型探索をする．縦型探索を続け，単語出現位置から予測された単語の音響スコアが，あるしきい値より悪くなった場合には横型探索に復帰する．この利用法では縦型探索も行うので処理の高速化が期待できる．

4　おわりに

記者原稿から作成された単語出現位置辞書を使うことにより，n≧4であるn-gramを利用した音声認識のための言語モデルについて述べた．今後は実際に計算機上にインプリメントし，perplexityの計算や認識実験によって本言語モデルの検証を行う予定である．

今回は１文中で連続する単語列への利用のみについて説明したが，単語出現位置は以下のように言語制約として利用することも可能である．

○文間にまたがった制約

前文の最後の単語出現位置を記録しておき，次の文の先頭の単語の制約に利用する．例えば，記者原稿で「…しました(4001025)．一方(4001026)，…」というのが存在し，認識結果が「しました(4001025)」と１文の認識を終了したとする．すると，次の文の先頭では認識候補では「一方」を優先することが可能となる．

○ジャンルの制約

単語（主に内容語）が出現するジャンルは上位１桁目をみればわかる．そこで，ジャンルがかけ離れている候補（例えば，経済とスポーツ）の場合にはペナルティを与えたり，ジャンルが近い候補（政治と社会）の場合には優先したりすることが可能である．単語のジャンル別出現数は単語位置辞書から計算できる．

○文（節）の構造変化があった場合の制約

例えば，認識候補として「砧神社は初詣で」と得られたとする．図３の記者原稿では「砧(6002024) 神社 (6002025)」と「初詣 (6002021)」は文構造の変化により出現順が変わったが，単語位置が25 (or 24) と21と近いので，この認識候補を優先することが可能となる．

謝辞

本研究を進めるにあたって適切な助言をいただいた当研究所音声認識・言語情報処理グループに感謝する．

【参考文献】

[今井98]今井ほか「ニュース番組自動字幕化のための音声認識システム」情報処理学会研究会，HI80-11(SLP-23-11)，pp.59-64, 1998.

[小林98]小林ほか「ニュース音声認識のための時期依存言語モデル」日本音響学会講演論文集，2-1-17, pp.71-72, 1998.

[政瀧95]政瀧ほか「連続音声認識のための可変長連鎖統計言語モデル」電子情報通信学会研究会報告，SP95-73, pp.1-6, 1995.