ニュース音声の自動要約法とその評価法に関する検討 *

◎堀 智織　　古井 貞熙　　(東工大)

1. はじめに

これまで我々は，単語重要度と言語尤度を用いて，放送ニュース音声を発話毎に要約する音声自動要約手法を提案してきた [1][2][3]．本報告では，提案した音声自動要約手法を用いて生成された自動要約文に対し，要約文としての適性を評価するための評価尺度について検討する．被験者が手動で作成した要約文を正解要約文として評価を行う．評価尺度として，自動要約文における重要単語の抽出率，および正解要約文中に含まれる単語連鎖に対する適合率を用いる．NHK のニュース音声を大語彙連続音声認識システムを用いて音声認識した結果を 60–70% の文字数に自動要約し，提案する評価尺度を用いて自動要約文を評価した結果を報告する．

2. 音声要約のアプローチ

本手法は，認識された各発話文から相対的に重要な単語を，文字数を基準とした特定の割合 (要約率) の範囲で抽出し，それらを接合することにより要約文を生成する．要約文として抽出された各単語の単語重要度 (重要度スコア) と言語尤度 (言語スコア) の累積スコアを，要約文の尤もらしさを示す要約スコアと定義した．この要約スコアを最大とする単語列を動的計画法を用いて抽出し，最適な要約文とする [1][2]．

要約スコア

要約スコアは，各単語の単語重要度スコアと単語列の連接に対する言語スコアによって次式のように定義される．N 個の単語からなる認識単語列 $W = w_1, w_2, \ldots, w_N$ から要約文として M $(M < N)$ 個の単語を抽出し接合した単語列 $V = v_1, v_2, \ldots, v_M$ の要約スコアは次式によって示される．

$$S(V) = \sum_{m=1}^{M} \{\log P(v_m | v_{m-2} v_{m-1}) + \lambda I(v_m)\} \quad (1)$$

言語スコアには単語 trigram $P(v_m \mid v_{m-2}\ v_{m-1})$ を用いている．単語重要度スコア $I\ (v_m)$ には話題語らしさを表す話題語スコアとして重み付き TF・IDF 尺度を用いる．単語列より抽出された部分単語列を $V = v_1, v_2, \ldots, v_M (M < N)$ とするとき，要約処理は (1) 式を最大にする $\hat{V}$ を求める問題となり，動的計画法を用いて解くことができる．

さらに，同一の認識結果から生成された要約率の異なる要約文を比較するため，単語数に基づく正規化要約スコアを定義した [3]．

3. 自動生成要約文に対する評価尺度

これまで，音声認識結果より自動生成された要約文は，書き起こし文中における「重要」な単語の要約文への抽出率 (適合率) と書き起こし文に対する要約文の文意の適性を評価してきた．ただし,「重要」な単語とは，被験者により重要な情報を担う単語として選択された単語である．また，文意は被験者により「同意」「包含」「異なる」の 3 段階で評価を行った．

しかしながら,「重要」な単語の抽出率では，重要度の低い単語の抽出率は考慮されていない．さらに，文意に関する評価は，基準が曖昧で自動要約文の精度を詳細に評価するには至らなかった．そこで，書き起こしテキストに対し，被験者が手動で作成した要約文を正解要約文として，自動要約文に対しさらに詳細な評価を行う．

正解要約文は 10 人の被験者により,「日本語として意味が通じる範囲で，原文の文意を保持しつつ，60–70%の文字数になるよう単語を削減」という条件下で作成された．作成される正解要約文は，同一の書き起こしテキストから同条件で作成した場合でも，被験者により変動し一意には定まらない．このため，正解要約文の多様性を考慮した評価尺度が求められる．ここでは，正解要約文に対する自動要約文における全単語の重要度を考慮した重要単語抽出率，および全正解要約文中に含まれる部分単語列の単語連鎖に対する適合率を評価尺度として用いる．この尺度は，重要単語の抽出率が高く，単語連鎖の適合率が高い自動要約文が正解要約文に近く，要約文としても尤もらしいことに基づいている．

3.1 重要単語の抽出率

原文の情報の保持という点から要約文を評価するため，単語の主観的重要度に基づき認識文中の単語の自動要約文への抽出率を評価する．多くの被験者により正解要約文に抽出された単語は重要度が高く，抽出されにくい単語は重要度が低いと仮定できる．正解要約文中における各単語の被験者による選択率を各単語の主観的重要度スコアとし，自動要約文中の全単語の主観的重要度スコアの平均を重要単語の抽出率とした．この抽出率は，各単語の適合率の期待値と考えることができる．要約文 $V = v_1, v_2, \ldots, v_M$ における，適合率 R は，次式のように求められる．

$$R = \frac{\sum_{m=1}^{M} \frac{c(v_m)}{a}}{M} \quad (2)$$

a :　正解要約文を作成した被験者数
M :　要約文の総単語数
v_m :　要約文の m 番目の単語
$c(v_m)$:　v_m を選択した被験者の数

3.2 単語連鎖の適合率

式 (2) で示される重要単語の抽出率は，自動要約文中の単語単位の評価はできるが，要約文が日本語として適性であるか，文意を保持できているのかまでは評価できない．これまで，自動要約文中に出現する不適切な単語の連鎖が原文の文意を損ねる例が多く見られたことから，単語の連鎖に関する評価が求められる．そこで，自動要約文中に含まれる特定の長さの単語列が正解要約文に含まれるか否かを割合によって評価する．自動要約文に含まれる任意長の全ての単語列が，複数の正解要約文において少なくとも一つに含まれる割合を以下のように計算する．

要約文 $V = v_1, v_2, \ldots, v_M$ に含まれる長さ D の単語列の抽出率 C_D を，次式のように求める．

$$C_D = \frac{\sum_{m=D}^{M} \delta(v_{m-D+1}, \ldots, v_{m-1}, v_m)}{M - D + 1} \quad (3)$$

ここで,

$$\delta(u_D) = \begin{cases} 1 & if \ \ u_D \in U_D \\ 0 & if \ \ u_D \notin U_D \end{cases} \quad (4)$$

---

*A Study on Automatic Summarization of Broadcast News Speech and Evaluation Methods
By Chiori Hori and Sadaoki Furui (Tokyo Institute of Technology)

u_D : 長さ D の各単語連鎖
U_D : 正解要約文中に含まれる長さ D の単語列の集合

但し，原文の異なる位置に出現する単語は，たとえ表記が同一であっても別の単語として扱う．C_D は $D = 1$ の場合は単語単位の抽出率 (C_1)，要約文の系列長 M に等しければ文単位の正解率 (C_M) を示す値となる．$D \geq 2$ のときは要約文の文頭と文末の適性も考慮するため，要約文には文頭記号と文末記号を付加して評価する．

4. 音声自動要約実験と評価

実験方法

放送ニュース音声を音声認識し，単語正解精度が 90% 以上の文 21 発話について音声自動要約手法を用いて要約文を生成した [3]．要約用言語モデルとして毎日新聞の trigram(MAI) とニュース原稿に基づく trigram(NHK) を用いた．音声認識結果 (REC) と書き起こし文 (ANS) に対し要約文を自動生成した．生成された自動要約文を主観的重要単語の抽出率と単語連鎖適合率を用いて評価する．比較として無作為に単語抽出を行い要約文を生成した (RDM)．

評価結果

図 1 に，単語連鎖の適合率と重要単語の抽出率の関係を示す．ただし，単語連鎖は 3 単語とする．重要単語の抽出率において手法間の有意な差は無い．一方，単語連鎖の適合率は，無作為に抽出した要約文に対し自動要約文が有意に高い．毎日新聞で平均 0.74%，NHK のニュース原稿で平均 0.68%と，新聞に基づく言語モデルを用いて作成した自動要約文の単語連鎖適合率が高い．書き起こし文に対する自動要約文に対し，認識結果に対する自動要約文で単語連鎖適合率がやや低い．これは，音声認識誤りによるものである．

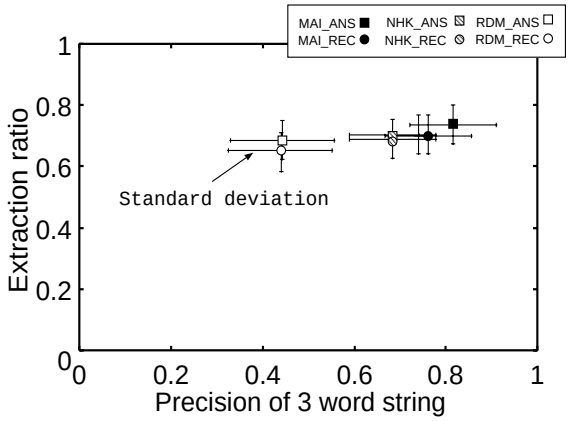

図 1．3 単語連鎖の適合率に対する重要単語の抽出率

図 2 に，単語連鎖数と単語連鎖の適合率の関係を示す．単語連鎖数が大きくなるに従い，単語の連接に対する制約が強まるため，誤りのある自動要約文の適合率は減少する．無作為に抽出した単語列では，単語連鎖数が 3 単語以上で適合率が 50%以下となり，単語の連接に不整合が多いことが分かる．一方，自動生成された要約文では単語連鎖の適性を保持している．

音声認識結果から自動生成した要約文で，言語モデルの違いによる適合率の差異は無い．また，書き起こしテキストから自動生成された要約文では，NHK のニュース原稿に対し毎日新聞を用いた要約文が適合率が高い．このことから，新聞記事に基づく言語モデルが要約文生成に有効であることがわかる．これらの結果は，被験者による文意の評価結果に一致する [3]．

図 1 の重要単語の抽出率において，無作為に抽出した単語列と自動要約文で有意差が無かったのは，1 単語の単語連鎖の適合率が 90%以上ということから，被験者数 10 人では正解要約文が多様で単語間の重要度に有意な差が現れなかったためである．

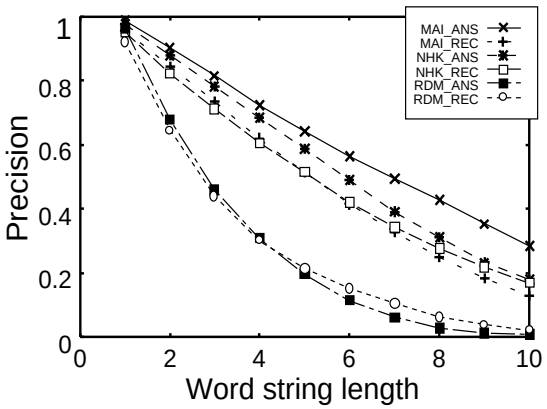

図 2．単語連鎖数と単語連鎖の適合率の関係

図 3 に，単語連鎖適合率による評価結果と被験者による文意の評価結果の相関を示す．ただし，図中の直線は回帰直線である．各自動要約文に対する 3 段階の評価に対し，「同意」2 点,「包含」1 点,「異なる」0 点とスコアをつけ，各文で被験者間の平均スコアを求めた．単語連鎖は 10 単語である．相関係数は 0.68 で，被験者による文意の評価と単語連鎖適合率に相関が認められる．

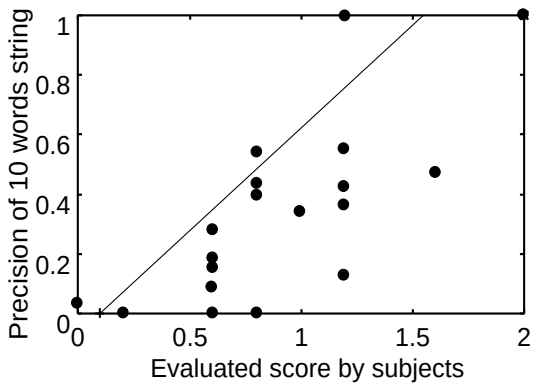

図 3．被験者による文意の評価と単語連鎖適合率の関係

5. まとめ

話題語と言語尤度に基づく要約スコアを最大化する規準で，動的計画法を用いる自動音声要約により生成した要約文の評価尺度を検討した．これまでの被験者による要約文の評価結果に対して，被験者の作成した正解要約文に基づく評価尺度を対応付けることができ，自動生成された要約文に対し定量的な評価ができることが確認された．今後さらに正解要約文を収集することにより，一層精度の良い評価を行いたい．また，正解要約文を分析することにより，自動要約処理の改良を行っていきたい．

謝辞

放送ニュースのデータベースを提供して下さった NHK 放送技研と新聞記事コーパスを提供して下さった毎日新聞社に感謝致します．

参考文献

[1] 堀 智織 他: 音学秋季講論,3-1-11, 1999.

[2] C. Hori et al.: Proc. 1999 Japan-China Symposium on Advanced Information Technology, pp. 75-82, 1999.

[3] 堀 智織 他: 信学技報, 99-SLP-29, pp.29–18(1999-12).