K-037

要約筆記品質向上システムにおける音声認識ツール利用の検討

Examination of the use of a speech recognition tool in the summary writing quality improvement system

高尾 哲康†
Tetsuyasu Takao

1．はじめに

聴覚障害者や高齢者への情報保障手段である要約筆記には「PC 要約筆記」と「手書き要約筆記」があり、いずれも要約筆記者が講演や番組などを聞き取り、リアルタイムで要約を行ない、キーボードや手書きで入力する。要約筆記者は「速く」、「正確に」、「読みやすく」の 3 原則をもとに、技術の向上を目指してさまざまな研修プログラムで訓練を重ねる。個々の研修プログラムでは要約筆記の品質の尺度として、要約筆記利用者からのフィードバックや意見・要望を受けることが多い[1]。これらのフィードバックは個々の事例として受けることが多く、定量的な品質評価を受けることはほとんどなく、長期間の研修を経ても要約筆記の品質向上の実感が得られにくくなっていた。これまで筆者らは講演者の発話内容のテキストと要約筆記者が入力したテキストをもとに定量的な評価が行ない、要約筆記者支援としてよりよい要約筆記表現を抽出する機能をもつシステムを試作した[2]。さらに、現在の PC 要約筆記が IPtalk[4]などを利用し、IP ネットワーク経由でひとつの発話文の前半と後半などに分けて分担入力し、2～4 人連携で行なわれている実情に合わせ、複数人による要約筆記文を自動連携することにより、要約筆記者の労力軽減とともに要約筆記文の品質向上をめざす試みを行なった[3]。今回、実利用が十分可能となってきた音声認識システムを活用し、要約筆記と組み合わせて高品質な要約テキストを出力する実験を行なった。

2．要約筆記データおよび音声認識システム

要約筆記研修プログラムで使用した発話テキスト(T1 で表わす)と PC 要約筆記者 4 名がリアルタイム要約筆記したテキスト(P1～P4 で表わす)を利用した。データの詳細を表 1 に示す。発話テキストには観光ガイド(約 4 分)を利用した。要約筆記者 2 人連携の部分は P1～P4 の各要約筆記テキストをもとに、それぞれ 2 名の自動連携を行なった結果のデータである。要約率は文字数基準の数値(要約筆記の総文字数/発話の総文字数)であり、要約評価は本システムにて品質評価した数値である。

音声認識システムは入手が容易な次の 6 システムを利用した。ドラゴンスピーチ 11J、AmiVoice SP2、Julius-4.3.1、Siri(Mac OSX)、Voice Rep Pro(Google)、Windows 音声認識である。いずれもマイク入力以外にライン入力または音声ファイルからの認識が可能である。ライン入力の場合は講演者のワイヤレスマイクからの電波を広帯域レシーバで受信、屋外での観光ガイド用イヤホンクリボ (CleVo)や指向性マイク、IC レコーダなどから品質のよい音声を取り込んで認識できる。IPtalk[4]のテキスト入力域やエディタ画面などに認識結果を直接入力可能である。要約筆記と同じ音声データにてトレーニング前の要約評価を表 1 に R1～R6 で示す。要約筆記との連携には、トレーニング前の認識率が高く、かつユーザプロファイルごとのトレーニング作業(音響学習、主に固有名詞登録)も容易なシステムの 1 つ(表 1 の R1、トレーニング後を R1T と表わす)を対象とした。

3．要約筆記品質評価システム

本システムはテキストアライメントモジュールと品質評価計算モジュール、複数テキスト連携モジュールから構成される(図 1)。テキストアラインメントモジュールは発話テキストと要約筆記や音声認識テキストを入力とし、統計情報と言語情報をもとに、動的計画法を利用して対応する文や段落を関連づけるモジュールである(m 文対 n 文の対応付け)。アラインメント単位ごとに発話文と要約筆記文や音声認識結果文のペアが作成される(XML 形式)。これにより品質評価計算対象範囲を狭くすることで後段の品質評価計算モジュールなどにおける評価計算精度を高めることができる。品質評価計算モジュールは、表記のゆれ(漢字の読みのひらがな・カタカナ表記など)や要約筆記特有の

表 1．要約筆記・音声認識テキストと連携による要約評価

		文字数	入力速度(文字数/分)	要約率(%)	要約評価
発話者	T1	972	249.2		
音声認識	R1	904	231.8	93.0%	0.8069
	R1T	944	242.1	97.1%	0.9258
	R2	901	231.0	92.7%	0.8092
	R3	981	251.5	100.9%	0.5165
	R4	831	213.1	85.5%	0.7634
	R5	811	207.9	83.4%	0.6777
	R6	857	219.7	88.2%	0.6640
要約筆記者	P1	533	136.7	54.8%	0.6522
	P2	492	126.2	50.6%	0.6052
	P3	370	94.9	38.1%	0.4902
	P4	497	127.4	51.1%	0.6180
要約筆記者1人＋音声認識	P1+R1	882	226.2	90.7%	0.8724
	P2+R1	812	208.2	83.5%	0.8558
	P3+R1	877	224.9	90.2%	0.8445
	P4+R1	931	238.7	95.8%	0.8287
	P1+R1T	925	237.2	95.2%	0.9261
	P2+R1T	871	223.3	89.6%	0.9051
	P3+R1T	890	228.2	91.6%	0.9001
	P4+R1T	951	243.8	97.8%	0.8904
要約筆記者2人連携	P1+P2	678	173.8	69.8%	0.7320
	P1+P3	669	171.5	68.8%	0.7130
	P1+P4	744	190.8	76.5%	0.7259
	P2+P3	629	161.3	64.7%	0.6744
	P2+P4	708	181.5	72.8%	0.7113
	P3+P4	649	166.4	66.8%	0.6752

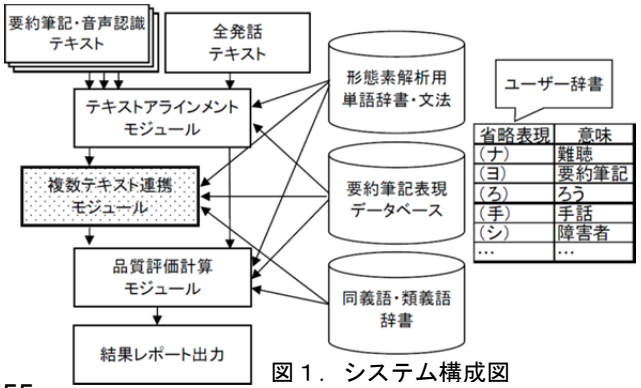

図 1．システム構成図

† 富山国際大学現代社会学部

Toyama University of International Studies, Faculty of Modern Society, Specialty of Management Information

省略表現などを吸収して正規化した形態素解析結果の形態素列に対し(形態素解析ツール MeCab を利用)、単語コスト、品詞コスト、単語間連接コスト、重複出現コスト(出現のたびに単調減少)を統計処理することにより、要約の品質評価(要約評価)の計算を行なう[2][3]。複数テキスト連携モジュールは、複数人の要約筆記文をマージしてよりよい要約筆記文にする機能をもち、品質評価計算モジュールのアルゴリズムを流用することで実現している。今回は音声認識結果文に要約筆記文をマージさせることで品質向上をめざした。

複数文のマージは 2 つの文の類似位置を評価値計算することで調べながら相互に異なる部分を抽出し、合成することで行なう。表 2 に音声認識結果の文「西郷さんは軍曹対象として江戸に来た」と要約筆記の文「官軍総大将として江戸へきた」をマージする例を示す。各文の形態素列において、コストは形態素解析用単語辞書に格納されている形態素コストを初期値としている。各形態素を 0〜1 の重み付き編集単位要素とみなして編集距離を求める。編集距離とは列 A と列 B について、A を編集操作(削除、挿入、置換)して B にするときの必要最低限の操作数のことである。評価値は編集操作コストを 2 つの文の形態素コスト値の総数で割り、数値の範囲を 0〜1 に正規化した数値にした。0 に近ければ 2 つの文の相違が多く、1 に近ければ相違が少なくなる。各セル値 E_{ij} の計算は表 2 の式にて全セルについて計算を行ない、表の最右下のセル値を 1 から引いた値が 2 つの文の評価値となり、この値が 1 に近いほど類似度が高いことになる。

2 つの文のマージの際の書き換え候補の抽出は次のように行なう。表 2 の評価値を算出するマトリクスにおいて、最右下のセルから最左上のセルまで評価値が最も小さくなる方向(上方、左方、左上方のいずれか)に順次たどることで 2 つの文の各形態素の対応セルが求まる。次に、2 つの文の対応関係のうち相互にマッチしないもの(前後のセル間で評価値の差が大きい場合)を抽出する。表 2 の例では、列方向と行方向の文を対応させて、

・「西郷さんは」⇔文頭
・「軍曹対象として」⇔「官軍総大将として」（相違部分とみなし、要約筆記文のほうを出力）
・「江戸へ」⇔「江戸に」（同評価値）
・「来た」⇔「きた」（同上）

が該当する。マージの結果、「西郷さんは」、「官軍総大将として」、「江戸へ」または「江戸に」、「きた」の順となり、マージ結果の文として、「西郷さんは官軍総大将として江戸へきた」が得られる。

4. 実験結果

音声認識システム R1〜R6, R1T、要約筆記者 P1〜P4 について、単独の場合と連携した場合について、発話文と比較した品質評価結果の要約評価を表 1 右側および図 2 に示す。要約筆記者の連携人数を 3 名以上に増やしてもそれほど要約評価が高くならないことがわかっている[3]。音声認識システムと要約筆記者 1 名を組み合わせるほうが要約筆記者 2 名連携の場合よりも要約評価が高くなっている。特にトレーニングを行なった音声認識システムと連携する場合は要約評価も 0.9 を越えることもあり、要約筆記者のミスによる悪影響が出やすくなる。次にその例を示す。

R:「江戸城を無血開城し明治維新を…」「明治 39 年…」
P:「江戸城を無血会場し明治維新を…」「明治 30 年…」

キーボード入力に関わるミスが影響している。音声認識システムについても認識誤りや失敗による影響が長く引きずらない(立ち直りが早い)ようになっているので要約筆記者が入力した正しい文節に置き換えやすい。

5. まとめ

本実験から品質のよい音声が得られる環境であれば音声認識システムと要約筆記者の連携による品質向上の効果が高いことがわかった。リアルタイム PC 要約筆記の本来の目的は情報保障にあるので、発話に対するのと同様、少々の冗長性や文体の不自然さ、誤りなどに対して寛容性があるのが現実である。しかし、特に固有名詞や数字などについて正確性は高いほうがよいので、音声認識システムとの最適な役割分担方法などを検討していく。

参考文献

[1] 話しことばの要約、三宅初穂、全国要約筆記問題研究会 (2012)
[2] 高尾哲康、要約筆記品質向上支援システム、FIT2013、7M-7、(2013)
[3] 高尾哲康、複数要約筆記文連携による要約筆記品質向上の試み、FIT2014、K-028、(2014)
[4] IPtalk、http://www.geocities.jp/shigeaki_kurita/

表 2．評価値計算と複数テキスト連携

i / j		西郷	さん	は	軍曹	対象	として	江戸	に	来	た
	コスト	3000	192	0	3649	2108	479	3000	0	145	0
	コスト	0.0000	0.1256	0.1337	0.1337	0.2865	0.3747	0.3948	0.5204	0.5204	0.5265	0.5265
官軍	3649	0.1528	0.2784	0.2865	0.2865	0.4393	0.5275	0.5476	0.6732	0.6732	0.6793	0.6793
総大将	3749	0.3098	0.4354	0.4434	0.4434	0.5962	0.6845	0.7046	0.8302	0.8302	0.8363	0.8363
として	479	0.3298	0.4555	0.4635	0.4635	0.6163	0.7046	0.6845	0.8101	0.8101	0.8162	0.8162
江戸	3000	0.4555	0.5811	0.5891	0.5891	0.7419	0.8302	0.8101	0.6845	0.6845	0.6906	0.6906
へ	0	0.4555	0.5811	0.5891	0.5891	0.7419	0.8302	0.8101	0.6845	0.6845	0.6906	0.6906
き	431	0.4735	0.5991	0.6072	0.6072	0.7600	0.8482	0.8282	0.7026	0.7026	0.7086	0.7086
た	0	0.4735	0.5991	0.6072	0.6072	0.7600	0.8482	0.8282	0.7026	0.7026	0.7086	0.7086

$$E_{i,j} = \min(E_{i-1,j}+C_{i-1}/C,\ E_{i,j-1}+C_{j-1}/C,\ E_{i-1,j-1}+A)$$

$$A = \begin{cases} 0 & : \text{ i-1 と j-1 の位置の形態素がマッチ（表記基本形、品詞、同義語）した場合} \\ (C_{i-1}+C_{j-1})/C & : \text{ 上記以外（C:コスト値の総和）} \end{cases}$$

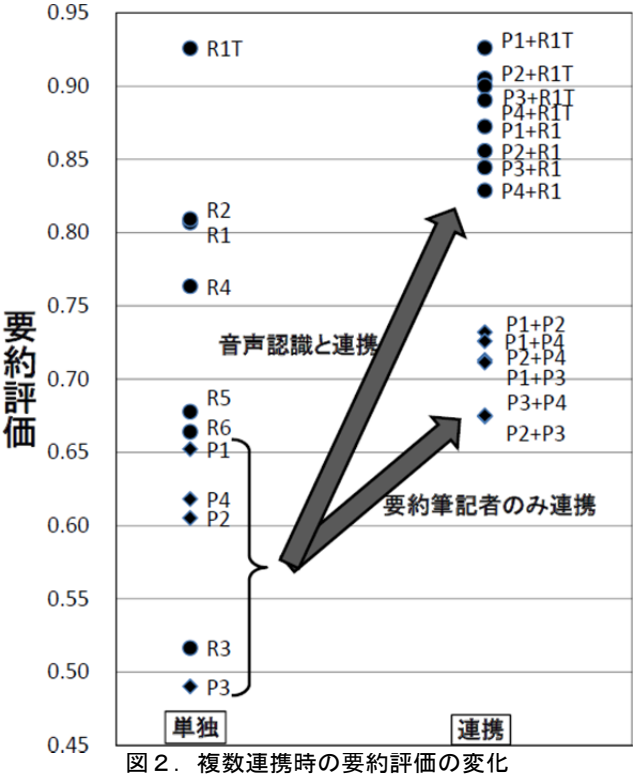

図 2．複数連携時の要約評価の変化