K-028

複数要約筆記文連携による要約筆記品質向上の試み

Merge Multiple Sentences Function of Quality Improvement Support System of Summary Transcript

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。一般に日本語の発話速度は 200～400 文字/分であり、要約筆記者による入力量は PC の場合で 100～200 文字/分、手書きの場合で 40～80 文字/分となっている。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][3][4]。現在、PC 要約筆記は IPtalk[5]などを利用し、IP ネットワーク経由で 2～4 人連携で行なわれている。ひとつの発話文の前半と後半などに分けて分担入力し、要約筆記文をスクリーンに表出する際の制御は手動で行なっている。本論文では、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。

2．品質評価に利用した要約筆記データ

要約筆記研修プログラムで使用した発話テキスト(T 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。2 人～4 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 2～4 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。

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テキストと要約筆記テキスト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][4]。複数テキスト連携モジュールは、複数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。

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 2 つの文「西郷さんは戊辰戦争で江戸へ」と「官軍総大将として江戸にやってきた」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

表１．要約筆記文と複数人連携による要約評価の向上

					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		
1人	P1				533	136.7	54.8%	0.6477
		P2			492	126.2	50.6%	0.6005
			P3		370	94.9	38.1%	0.4902
				P4	497	127.4	51.1%	0.6180
2人連携	P1	P2			678		69.8%	0.7320
	P1		P3		669		68.8%	0.7130
	P1			P4	744		76.5%	0.7259
		P2	P3		629		64.7%	0.6744
		P2		P4	708		72.8%	0.7113
			P3	P4	649		66.8%	0.6752
3人連携	P1	P2	P3		778		80.0%	0.7726
	P1	P2		P4	814		83.7%	0.7814
	P1		P3	P4	821		84.5%	0.7612
		P2	P3	P4	771		79.3%	0.7530
4人連携	P1	P2	P3	P4	883		90.8%	0.7991

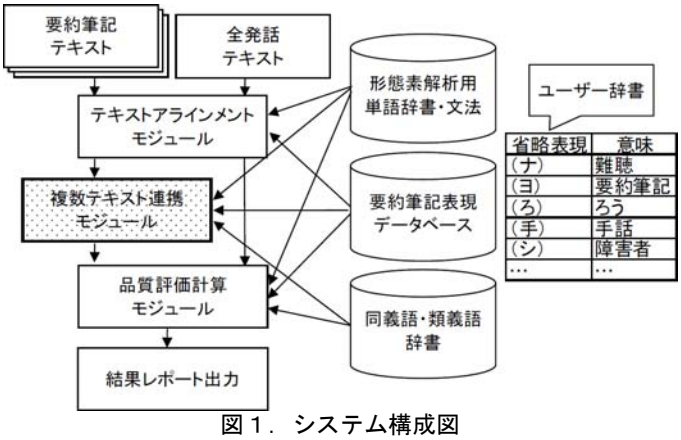

図１．システム構成図

† 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

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
・「戊辰戦争で」⇔「官軍総大将として」（相違部分とみなし、両方を出力）
・「江戸へ」⇔「江戸に」（セルの評価値から同じものとみなし、どちらか一方、例えば前後の文節とのつながりやベテラン要約筆記者のものをとるなど）
・文末⇔「やってきた」

が該当する。マージの結果、

「西郷さんは」
｛「戊辰戦争で」　「官軍総大将として」｝
「江戸へ」または「江戸に」
「やってきた」

の順となり、マージ結果の文として、「西郷さんは戊辰戦争で官軍総大将として江戸へやってきた」が得られる。文節境界判定は構文解析ツール CaboCha を利用し、形態素間のつながり保たれるようにした。

4. 実験結果

筆記者 P1～P4 について、単独の場合、2～4 人連携によりマージされた要約筆記文について、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。複数人連携を行なうことで品質が向上していることがわかる。各要約筆記者は他の要約筆記者が入力した文を気にする必要はないので労力軽減につながる。しかし、マージを行なうことで文字数基準の要約率が高くなり、要約筆記利用者が読まされる量が多くなる。その原因は、マージにより冗長度が増えること、その結果としてやや不自然な文体になることにある。以下にその例を示す。

「西郷隆盛は人望があつく人気者」
「西郷は人に愛されたので人気が高い」
→「西郷隆盛は人望があつく人に愛されたので人気が高い者」
「戦いは終結した」
「その日のうちに終わった」
→「戦いは終結しその日のうちに終わった」
「ここにこの墓が建てられました」
「その場所にお墓がたてられました」
→「ここにその場所にお墓がたてられました」
「テレビ放送が始まりました」
「TV 放送スタート」
→「テレビ放送がスタート始まりました」

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などに対していくぶん寛容性があるのが現実である。しかし、いきすぎると可読性(「読みやすさ」の原則)が損なわれるので適度なバランスが必要と考える。

5. まとめ

本実験から複数人の要約筆記文を組み合わせることによりよい要約筆記文となることがわかった。複数人で要約筆記する場合、各要約筆記者のベストな組合せを推測できる可能性もある。冗長性と不自然な文体の箇所の自動検出などの対策が課題である。今後は、さまざまな要約筆記データを収集し、要約評価精度の向上や失敗箇所についての分析を進めるとともに、コスト計算手法やパラメータの最適化などを行なっていく。

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、要約筆記品質評価システムの改良、FIT2011、3Q-5、(2011)
[3] 高尾哲康、要約筆記品質評価システムにおける要約候補文提示機能、FIT2012、2M-6、(2012)
[4] 高尾哲康、要約筆記品質向上支援システム、FIT2013、7M-7、(2013)
[5] IPtalk、http://www.geocities.jp/shigeaki_kurita/

表２．評価値計算と要約筆記テキスト連携

i / j		西郷	さん	は	戊辰	戦争	で	江戸	へ
	コスト	3000	192	10	3000	1893	10	3000	10
	コスト	0.0000	0.1307	0.1390	0.1395	0.2701	0.3526	0.3530	0.4836	0.4841
官軍	3649	0.1589	0.2896	0.2979	0.2984	0.4290	0.5115	0.5119	0.6426	0.6430
総大将	3749	0.3222	0.4529	0.4612	0.4617	0.5923	0.6748	0.6752	0.8058	0.8063
として	479	0.3431	0.4737	0.4821	0.4825	0.6132	0.6956	0.6960	0.8267	0.8271
江戸	3000	0.4737	0.6044	0.6127	0.6132	0.7438	0.8263	0.8267	0.6960	0.6965
に	10	0.4742	0.6048	0.6132	0.6136	0.7443	0.8267	0.8271	0.6965	0.6960
やってき	949	0.5155	0.6461	0.6545	0.6549	0.7856	0.8680	0.8685	0.7378	0.7374
た	10	0.5159	0.6466	0.6549	0.6554	0.7860	0.8685	0.8689	0.7383	0.7378

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

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

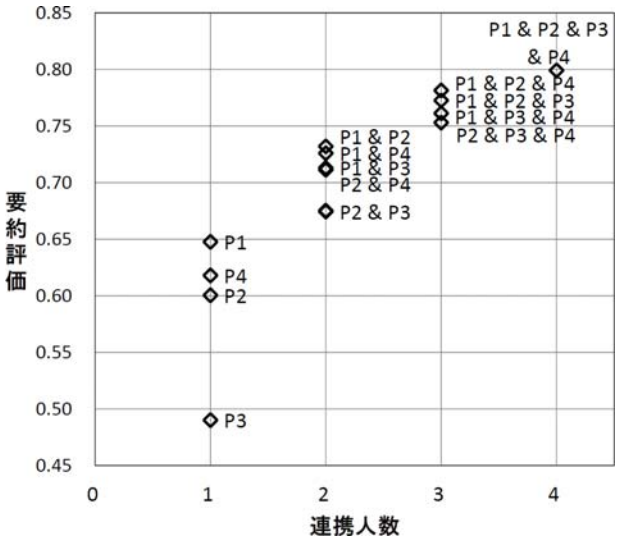

図２．複数人連携時の要約評価値の変化