『日本語話し言葉コーパス』における 自由要約・重要文抽出データについて Version 1.0

野畑 周　内元 清貴　高梨 克也　井佐原 均
(情報通信研究機構)

目次

1　本冊子の内容

本マニュアルは、『日本語話し言葉コーパス』における要約データの仕様を解説したものである。ここで述べる要約データは、自由要約・重要文抽出の 2 種類である。自由要約データは、転記テキストから直接書き言葉の文章の形式に講演を要約したものである。一方、重要文抽出データは、自由要約データとは独立に、節単位が与えられた講演の転記テキストから、各節のうち重要だと思われる部分を抽出したデータである (節単位についての詳細は「『日本語話し言葉コーパス』における節単位認定 (clause.pdf)」を参照)。これらのデータは、自動要約のための訓練データや正解データとして用いることを意図している。

要約データ作成の対象とした講演は、「コア」中の独話 177 講演 (対話・再朗読以外) とコア以外のテストセット 22 講演を含む 199 講演で、そのうち学会講演は 85 講演、模擬講演は 114 講演である。各講演に対して、それぞれ 10%と 50%の二種類の要約率についてデータを作成した。要約データの作成にあたっては、どちらの要約率に基く要約においても、話題の一部を示す要約 (indicative summary) ではなく、講演内容の要点を可能な限り網羅した要約 (informative summary) とすることを意図した。複数の作業者により各要約率ごとに 3 個の要約データが作成され、各講演につき自由要約 6 個、重要文抽出 6 個のデータが利用可能である[1]。重要文抽出データは、『日本語話し言葉コーパス』の統合データにおいて SUW 属性の一部として提供される (詳細は「『日本語話し言葉コーパス』XML 文書について (xml.pdf)」を参照)。一方、自由要約データは、XML 形式の統合データとは独立にテキストファイル形式で格納されている。

2　自由要約データ

自由要約データとは、講演の転記テキストを作業者が読み、内容をできるだけ保ちつつ書き言葉の文章の形で要約を行ったものである。作業の入力となる転記テキストでは、与えられているタグ情報を利用して識

[1] 2 節で言及するように、一部のデータについては講演者自身による要約も含まれる。

別できる言い誤りや言い淀みは作業前に削除してある (詳細は「転記テキストの仕様 (transcription.pdf)」を参照)。要約率は 10%と 50%の 2 種類とし、作業時には要約率に相当する文字数を示して、おおよそ目標とする要約率となるように指示した。ただし、とくに 10%の要約においては必要な内容を含めるのに字数の制限が厳しい場合が多く、その場合には制限字数の 1 割程度の超過を認めるものとした。要約データには単一の正解と定められるものはないため、複数の要約結果を得るために、総勢 9 人の作業者に要約データを作成してもらい、10%、50%共にそれぞれ 3 個ずつ要約データが提供できるようにしている。具体的な要約文の形式は、次のように定めた。

1. 要約文の文体は、常体 (である体) とする。
2. 要約文の形式は、日本語として自然な文のみで構成される文章の形式とし、基本的に箇条書き・表などの形式はとらないものとする[2]。
3. 要約文中の各文は、読みやすいように適切な長さで区切るものとする。
4. 要約文中の各文の末尾は句点で示し、句点の後には改行を入れるものとする。
5. 要約文中の各文を構成する文字は、全角のひらがな、カタカナ、アルファベット、漢字、引用符、読点、句点のみとし、半角文字は用いないものとする。
6. 段落の開始位置を示す場合には、文の先頭に全角空白 1 個を入れるものとする。

要約率 50%のデータ作成では編集操作を限定し、重要な部分の抽出と各部分内での表現の変更のみで要約作成を行っている。要約率 10%の場合は、基本的には要約率 50%の場合と同様の操作を中心にデータ作成を行うが、それでは必要な内容を十分に含めることができない場合には、自由な表現の書き換えや部分の入れ換えを許している。なお、自由要約作成後に、人名など一部の表現については転記テキストにおける扱いに沿って「×」記号によって伏字に変換することとしたため、情報が落ちている文があることを御了承されたい (詳細は「転記テキストの仕様 (transcription.pdf)」(R) タグの項目を参照)。

また、講演を行った講演者の方々にも依頼し、自身の講演の要約データを同様の作業仕様に基いて作業をお願いした。結果として、18 講演について (内 16 講演は 10%、50%の二種類とも) 講演者自身による要約結果を自由要約データに含めることができた。

自由要約データは、XML 形式の統合データとは独立に、別ディレクトリにテキストファイル形式で格納されている。ファイル名は、各講演の ID に、ピリオドをはさんで要約率、作業者の区別を示す文字を付し、さらに自由要約データであることを示す拡張子 fsum を付加したものとなる。アルファベットと要約率・作業者の対応を表 1 に示す。自由要約ファイルは、要約率ごとにディレクトリを分けて格納されている。例えば、講演 A01F0055 に対する自由要約データは、以下のように格納されている:

```
A01F0055/
     SUMMARY/
          10PER/
               A01F0055.1A.fsum, A01F0055.1B.fsum,
               A01F0055.1C.fsum, A01F0055.1S.fsum
          50PER/
               A01F0055.5A.fsum, A01F0055.5B.fsum,
               A01F0055.5C.fsum, A01F0055.1S.fsum
```

[2]ただし、講演の形式によっては例文を列挙する場合など、一部箇条書きの形式を許している場合もある。

表 1: 要約データのファイル名

要約率を示す文字	要約率	作業者の区別を示す文字	作業者
1	10%	A-I	作業者 (9 名)
5	50%	S	講演者

表 2: 統合データにおける重要文判定の属性名

属性名	説明	値 (0=非採択,1=採択)[3]
SE_Subject1_50p	重要文抽出作業者 1_50p 結果	0/1
SE_Subject1_10p	重要文抽出作業者 1_10p 結果	0/1
SE_Subject2_50p	重要文抽出作業者 2_50p 結果	0/1
SE_Subject2_10p	重要文抽出作業者 2_10p 結果	0/1
SE_Subject3_50p	重要文抽出作業者 3_50p 結果	0/1
SE_Subject3_10p	重要文抽出作業者 3_10p 結果	0/1

3 重要文抽出データ

重要文抽出データは、講演の転記テキストについて、節単位で重要な部分を認定したものである。自動要約の主要な手法の一つである重要文抽出は、文章中の各文についていくつかの評価尺度を用いて重要度を求め、その結果に基づいて重要と思われる文を抽出するものである。重要文抽出データは、講演の自動要約システムの研究をすすめる上で有用なものであると考えられる。

書き言葉の場合とは異なり、話し言葉の場合には単位となる「文」の定義自体が必ずしも自明でない。『日本語話し言葉コーパス』においては、単語の品詞などを用いた一定の基準に従って機械的に節単位を認定した後、人手でその結果を修正して最終的な節単位を与えている (「『日本語話し言葉コーパス』における節単位認定 (clause.pdf)」)。今回作成した重要文抽出データでは、そのようにして与えられた各節単位を「文」として、与えられた転記テキストを対象に各「文」について重要かどうかを判定したデータである。以降では、節単位のことを「文」と呼ぶことにする。

要約率は自由要約データと同様に 10%と 50%に設定したが、重要文抽出データでは要約率を文字数ではなく文数に変換して用いている。作業者は、まず内容を最大限保ちつつ全体の 50%になるように重要な文を選択し、更にその結果から全体の 10%になるように文を選択している。従って、要約率 10%のデータに含まれる文は必ず要約率 50%の文にも含まれている。自由要約データと同様に、一講演について複数の作業者が重要文を抽出し、10%、50%共に 3 個ずつ重要文抽出データを作成した。

重要文抽出の情報は『日本語話し言葉コーパス』の統合データにおいて SUW 属性の一部として提供される (詳細は「『日本語話し言葉コーパス』XML 文書について (xml.pdf)」を参照)。統合データ中の重要文に関する属性を表 2 に示す。重要文抽出データについては、作業者のうち二人は 199 講演全体について重要文の判定を行ったが、3 個目のデータについては、時間の制約もあり 4 人の作業者が分担して判定を行った。このため、Subject3 の判定は実際には、複数の作業者の判定結果を統合したものとなっている。

3.1 重要文抽出データの比較

重要文抽出データについて、Kappa 値を用いてデータ間の比較を行った結果を紹介する。

[3]実際には、省サイズ化のため属性値が 0 の場合は属性の記述自体を省略している。

表 3: 3 種類の重要文抽出データに対する Kappa 値

データ	S1 ↔ S2	S2 ↔ S3	S3 ↔ S1	平均
10%(全体)	0.378	0.356	0.374	0.369
50%(全体)	0.477	0.459	0.499	0.478
10%(学会)	0.342	0.342	0.345	0.343
50%(学会)	0.473	0.433	0.486	0.464
10%(模擬)	0.404	0.366	0.395	0.388
50%(模擬)	0.481	0.479	0.508	0.489

表 3 は、三種類の文抽出データ（以下 S1, S2, S3 と区別する）のうち二種類を選んだ全組合せについて、重要文抽出データの Kappa 値を求めた結果である。

Kappa（κ）値とは、二つのデータの間で偶然に一致する割合を除いた一致度を示す指標であり、以下の式で定義される [Car96]：

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

$P(A)$ は実際に両データ間で一致した割合を、$P(E)$ は偶然に両データ間で一致する割合を表す。文数 S_n の講演に対して、一方で抽出された重要文の数を S_a、他方で抽出された重要文の数を S_b とし、双方のデータで共通して重要文として抽出された文の数を S_c とすれば、$P(A)$、$P(E)$ は以下のように定義される：

$$P(A) = \frac{S_c}{S_n} + \frac{S_n - S_a - S_b + S_c}{S_n}$$
$$P(E) = \frac{S_a}{S_n}\frac{S_b}{S_n} + (1 - \frac{S_a}{S_n})(1 - \frac{S_b}{S_n})$$

表 4 は、[CIK+97] に示された Kappa 値を解釈するための分類表である。Kappa 値が高いほど、両データの一致度が高く、データの内容に信頼性があると考えられる。[Kri80] における考察では、Kappa 値が 0.7 未満の場合には比較した両者の関連を示すことは困難な場合が多いと述べられている。表 3 の結果を表 4 に基いて解釈すれば、データ間の一致度は “Fair” から “Moderate” にあたり、作業者の結果を統一した重要文抽出データの作成は困難であることが分かる。

また、学会講演 85 講演と模擬講演 114 講演を分けて Kappa 値を計算してみると、どちらの要約率においても学会講演の方が模擬講演よりも値が低い。模擬講演では講演者が自分の話したい題目について自由に話をしているのに比べて、学会講演の場合は一般に講演の最初や最後に内容をまとめる部分があることが多く、重要文の位置が作業者間で一致する度合は模擬講演よりも高いと予想されたが、今回作成した重要文抽出データに関しては、必ずしもそうでないことが分かった。

表 4: Kappa 値の分類表

Kappa 値	値の解釈
< 0	Poor
.00 – .20	Slight
.21 – .40	Fair
.41 – .60	Moderate
.61 – .80	Substantial
.81 – 1.0	Near perfect

参考文献

[Car96] Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, Vol. 22, No. 2, pp. 249–254, 1996.

[CIK+97] Jean Carletta, Amy Isard Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, Vol. 23, No. 1, pp. 13–31, 1997.

[Kri80] Klaus Krippendorff. *Content Analysis: An introduction to its methodology.* Sage Publications, 1980.

[野畑 02] 野畑周, 関根聡, 内元清貴, 井佐原均. 話し言葉コーパスにおける文の切り分けと重要文抽出. 第 2 回 話し言葉の科学と工学ワークショップ, pp. 93–100, 2 月 2002.

[野畑 04] 野畑周, 高梨克也, 内元清貴, 井佐原均. 『日本語話し言葉コーパス』における要約データの作成. 第 3 回 話し言葉の科学と工学ワークショップ, 2 月 2004.

付録

要約データ作成の対象である講演ファイル名一覧を以下の表に示す。全 199 講演のうち、「コア」に含まれるものは 177 講演である。それ以外の 22 講演 (コア以外のテストセット) は表中では斜体で示している。

講演ファイル名 (199 講演)				
A01F0055	A03F0153	S00F0177	S02F0129	S04F1495
A01F0067	A03M0004	S00F0197	S02F0180	S05F0463
A01F0122	A03M0005	S00F0209	S02F0189	S05F1041
A01F0132	A03M0010	S00F0210	S02F0852	S05F1517
A01F0143	A03M0018	S00M0025	S02M0011	S05F1600
A01F0145	A03M0045	S00M0053	S02M0043	S05M0412
A01M0007	A03M0059	S00M0065	S02M0068	S05M0613
A01M0015	A03M0061	S00M0071	S02M0076	S05M1236
A01M0020	A03M0138	S00M0075	S02M0092	S05M1505
A01M0021	A04M0026	S00M0112	S02M0103	S05M1666
A01M0025	A04M0047	S00M0115	S02M0161	S06F0167
A01M0030	A05F0039	S00M0117	S02M0191	S06F1034
A01M0048	A05F0043	S00M0153	S02M0198	S06F1566
A01M0056	A05F0154	S00M0199	S02M0245	S06M0373
A01M0065	A05F0502	S00M0213	S02M1698	S06M0894
A01M0070	A05M0002	S00M0218	S03F0062	S06M0895
A01M0074	A05M0031	S00M0221	S03F0072	S07M0833
A01M0083	A05M0040	S01F0006	S03F0108	
A01M0096	A05M0068	S01F0038	S03F0119	*A01F0001*
A01M0097	A06F0028	S01F0050	S03F0133	*A01F0034*
A01M0099	A06F0049	S01F0074	S03F0184	*A01F0063*
A01M0103	A06F0073	S01F0151	S03F0214	*A01M0141*
A01M0110	A06F0075	S01F0157	S03F0224	*A02M0012*
A01M0115	A06F0120	S01F0166	S03F0232	*A03M0016*
A01M0131	A06F0128	S01F0183	S03F0314	*A03M0106*
A01M0133	A06M0092	S01F1522	S03F0383	*A03M0112*
A01M0137	A07F0844	S01M0005	S03F1477	*A03M0156*
A01M0140	A11M0369	S01M0051	S03F1577	*A04M0051*
A01M0142	A11M0469	S01M0091	S03M0003	*A04M0121*
A01M0147	A11M0846	S01M0101	S03M0046	*A04M0123*
A01M0157	S00F0014	S01M0182	S03M0089	*A05M0011*
A02F0038	S00F0031	S01M0205	S03M0098	*A06F0135*
A02F0082	S00F0041	S01M0225	S03M0106	*A06M0064*
A02F0116	S00F0066	S01M0227	S03M0141	*S00F0019*
A02M0076	S00F0082	S01M0706	S03M0194	*S00F0148*
A02M0098	S00F0083	S02F0012	S03M0201	*S00F0152*
A02M0107	S00F0088	S02F0094	S03M0317	*S00M0008*
A03F0072	S00F0131	S02F0100	S03M0996	*S00M0070*
A03F0108	S00F0134	S02F0113	S03M1133	*S00M0079*
A03F0109	S00F0173	S02F0121	S04F0013	*S01F0105*