BCCWJ-SUMM：『現代日本語書き言葉均衡コーパス』を元文書とした要約文書コーパス

浅原 正幸 (国立国語研究所) *
杉 真緒 (国立国語研究所・津田塾大学)
柳野 祥子 (国立国語研究所・津田塾大学)

BCCWJ-SUMM: A Summarization Corpus of the 'Balanced Corpus of Contemporary Written Japanese'

Masayuki Asahara (NINJAL)
Mao Sugi (NINJAL, Tsuda College)
Shoko Yanagino (NINJAL, Tsuda College)

要旨

『現代日本語書き言葉均衡コーパス』を元にした要約文書コーパスの設計について報告する。要約文書作成においては、クラウドソーシングを用いて 1 文書に対して 100 件規模で要約文書を収集する方法と、実験室において 1 人の被験者に複数回要約文書作成を依頼する方法の 2 通りを試行する。さらに作成した要約データに対する人手による主観評価情報を付与する。本稿では現在の進捗を報告するとともに今後の課題について示す。

1. はじめに

人間の文書理解過程は多様である。背景知識が異なる書き手と読み手との間には認知に乖離があり、何を伝えたいのかと何を読み取りたいのかとが必ずしも一致するとは限らない。また複数人の読み手が 1 つのテキストに対して何を重要視するかについても必ずしも一致するとは限らない。さらに 1 人の読み手の認知についても時間や回数の経過とともに変わってくるだろう。

本稿では『現代日本語書き言葉均衡コーパス』(以下 BCCWJ; Maekawa et al. (2014)) を元文書とした要約文書コーパスの設計について報告する。要約文書コーパスの分析を通して文書理解過程の多様性をとらえることを第一義的な目的とする。コーパスのその他の用途として、成人母語話者の作文能力の評価データや単一文書自動要約のためのベンチマークデータを想定している。収集した要約文書コーパスには要約文の優劣を評価し、人手による主観評価情報を付与する。5 種類の評価指針を立て、作業者 2 人により 5 段階の主観評価を行う。

以下 2 節では要約文の収集方法について述べる。3 節では収集した要約文に対する主観評価情報の付与について議論する。4 節ではまとめと今後の予定について述べる。

* masayu-a@ninjal.ac.jp

2. 要約文の収集

要約文の元文書として BCCWJ の新聞 (PN) サンプル (アノテーション優先順位 A) を用いる。BCCWJ の PN 可変長サンプルは複数記事からなるものもあり、これらについては記事単位に分割して元文書データを 19 文書作成した。

クラウドソーシングにより安価で大量にデータを得る手法 (タイプ入力:BCCWJ-SUMM_C) と実験室にて被験者に 3 回繰り返し要約作成課題を依頼してデータを得る手法 (筆述:BCCWJ-SUMM_L) の 2 種類の方法を用いた。表 1 に収集した要約文の概要について示す。

表 1　収集した要約文の概要

言語資源名	収集場所	生成過程	繰り返し	取得人数	摘要
BCCWJ-SUMM_C	クラウドソーシング	タイプ入力	なし	100-200	19 文書の要約
BCCWJ-SUMM_L	実験室	筆述	3 回	のべ 47	8 文書の要約

以下各言語資源について解説する。

2.1 BCCWJ-SUMM_C

BCCWJ-SUMM_C は BCCWJ の新聞記事の要約を Yahoo! クラウドソーシング (15 歳以上の男女) により被験者実験を行い作成したものである。

40 文字毎に改行した元文書を画像として提供し、実験協力者に 50-100 文字に要約せよという指示で収集した。実験協力者は元文書をコピーして作業することができないために、画像を見ながらタイプ入力を行う必要がある。実験協力者の環境は PC 環境に限定した。元文書毎に約 100〜200 人の実験協力者が要約に従事した。実験実施時期は 2014 年 9 月である。

得られたデータ 19 文書の統計を表 2 に示す。収集要約数はクラウドソーシングで得られたファイルの総数である。得られたデータには、文字数制限を守っていないもの・実験の趣旨を理解していないもの・既に実験を行った実験協力者から同一回答を提供されたと考えられるものなどが含まれており、これらを排除したものを有効要約とした。

2.2 BCCWJ-SUMM_L

BCCWJ-SUMM_L は BCCWJ の新聞記事の要約を実験室環境で筆述により作成したものである。BCCWJ-SUMM_C で用いた元文書を印刷紙面で提供し、実験協力者に 50-100 文字に要約せよという指示で収集した。1 つの元文書に対して、3 回まで繰り返して要約文作成を行った。繰り返しに際しては、特別に「前と同じ要約文を作成してください」などといった指示は行わず、質問された場合にも「自由に要約文を作成してください」と教示した。被験者実験は強制ではなく被験者が拒否した時点で実験を終了するため、3 回繰り返していない事例も含めた。実験協力者は原稿用紙上で筆述 (鉛筆と消しゴム利用) で要約を行い、そのデータを電子化した。

現在のところデータは 8 文書のべ 61 人分に限定した。得られたデータの概要は表 3 のとお

表 2　BCCWJ-SUMM_C データ概要

FileID	有効要約数	収集要約数
A_01	106	198
A_02	112	195
B_02	98	149
B_03	74	100
C_01	63	100
C_02	63	99
C_03	53	100
D_01	55	100
D_02	55	100
D_03	48	99
D_05	55	99
E_01	58	99
E_02	46	98
E_03	54	100
E_04	60	99
E_05	48	100
E_06	56	98
F_01	57	100
F_02	58	100

表 3　BCCWJ-SUMM_L データ概要

FileID	有効要約数	被験者数
A_01	19	7
A_02	18	6
B_02	21	7
B_03	27	9
C_01	27	9
C_02	21	7
C_03	18	6
Q	30	10

り。本実験の実験参加者からは要約作業前に要約元文書の読み時間 (視線走査法もしくは自己ペース読文法) のデータも取得した。さらに被験者の特性 (最終学歴・語彙数・言語形成地・記憶力) などのデータについても収集した。実験実施時期は 2014 年 8 月～2015 年 1 月であるが、今後このデータは引き続き拡充していく予定である。

3.　人手による要約の主観評価

収集した要約文に対して、主に読みやすさに関して人手による要約の主観評価を付与する。

人手による要約の主観評価として DUC-2005 [(1)]で用いられた以下の 5 種類の評価指針を用いる：

- 文法性 (Grammaticality)： 誤字・文法的でない文が含まれていないか
- 非冗長性 (Non-redundancy)：全く同じ情報が繰り返されていないか
- 指示詞の明解さ (Referential clarity) : 先行詞のない指示詞（代名詞）が含まれていないか
- 焦点 (Focus)：要約全体と無関係な情報が含まれていないか
- 構造と一貫性 (Structure and Coherence)：接続詞を補ったり削除したりする必要のある箇所はないか

この 5 種類の評価指針について A (very good) - E (very poor) の評価を行う。現在主観評価付与作業は 2 人の作業者により行っている。基準の統制後、作業者を増やすことも検討する。DUC は対象言語が英語であるために、指針については DUC-2005 の quality question をそのま

[(1)] http://www-nlpir.nist.gov/projects/duc/duc2005/

ま用いず、作業者間で調整しながら基準を策定中である。現在までに得られている作業者メモから主観評価における細かい指針と論点について示す：

- 全体：
  特に問題がないものを A とし、作文として問題が軽度のものを B とする。C 以下は問題の程度に応じて付与する。
  C は欠陥が認められるがぎりぎり意味が通じる程度のものとし、程度や件数に応じて D 以下を付与する。
- 文法性 (Grammaticality)：
  問題のないものは A とする。誤字については「蓮舫」→「蓮坊」[(2)]のような単純なタイプミス、変換ミスは B とする。
  「法学部への進学し、」のような文法的な誤りが 1 件ある場合は C とし、1 件増えるごとに評価を 1 段階ずつ下げる。誤字の評価に加えて文法的でないものがあった場合、評価を 2 段階下げる。
  文法的なものについては、問題がないものには A、意味は通じるもの（読点の使い方や文のわかりやすさに改善点があるもの）には B を付与する。意味は通じるがわかりにくいもの（主語や目的語が省略されていてかつ意味が不明確なもの、コロケーションが不適切なもの）には C、日本語として不自然なもの（「たり」の使い方、助詞「の」の連続など）には D、明らかに文法的でないものには E を付与する。
  元文書にある誤用「レッテルを張る」についても漢字の誤用として評価を下げる判断を行った。
- 非冗長性 (Non-redundancy)：
  問題のないものは A とする。固有名詞や人を表す名詞（先生など）が重複しているような場合には B を付与し、普通名詞などの重複は C を付与する（喋る → しゃべりなど、品詞が変わっているものも含む）。表現の意味的な重複は D とする（才能 → 能力など）。冗長性が複数認められた場合は E とする。
  その他、言い換えられているが同じものを指す場合 C とする。
  現在のところ単語レベルの冗長性のみを検討しているが、句レベル・文レベルの基準についても事例が出現次第、随時検討する。
- 指示詞の明解さ (Referential clarity)：
  問題のないものは A とする。指すものが曖昧な場合、要約文を読むだけで曖昧性が解消できるものには B を付与し、推測はできるが書き手の指示するものが分かりにくいものには C を付与する。全く指示詞などの情報が示されていない、また明解でないものが複数ある場合、程度や件数に応じて D か E を付与する。
- 焦点 (Focus)：
  問題のないものは A を付与する。
  表現の仕方により、元文書の内容と違う読み方がされる可能性があるものは B か C を

(2) かな漢字変換ツールによっては変換が困難であるため。

付与する。要約におけるある部分要素（事例）にのみかかわる場合はBを付与し、要約全体の意味にかかわる場合はCを付与する。
要約作成者が元文書の内容理解に失敗している可能性があるものはCもしくはDを付与する。厳密には内容と合っていないものにはCを付与し、主体や語彙の意味などを取り違えているものはDを付与する。
元文書の要点とずれているものや、要約に不必要な情報が入っているものにはDを付与する。
内容と関係のない情報（原文に記述されていないことや書き手の意見）が入っているものにはEを付与する。

- 構造と一貫性 (Structure and Coherence)：
問題のないものにはAを付与する。
表記に一貫性のないものが高々1件の場合はBを付与し、複数あればCを付与する。具体的には漢字（ひらくかどうか）や呼称、記号の使用などを対象とする。
文章を通して、主語の交代が頻繁である場合はCを付与する。
接続詞の使用や、複文・重文の構成に改善点がある場合はDを付与する。具体的には接続詞の誤用、欠落など。またひとつの文を複数に切ったほうがよいものも対象とする。
文体に一貫性がないものにはD以下を付与する。具体的には語尾が一貫していないものなどを対象とする。

なお、細かい指針については今後修正される可能性がある。

表4　A_01サンプルに対する評価指標付与

	A	B	C	D	E	相関係数
文法性	9,5	7,3	3,8	3,7	1,0	0.72
非冗長性	21,9	2,5	0,4	0,5	0,0	0.07
指示詞	22,7	1,8	0,3	0,5	0,0	0.67
焦点	19,8	3,1	1,6	0,8	0,0	0.09
構造と一貫性	14,8	3,0	4,5	2,8	0,2	0.73

表4にBCCWJ-SUMM_CのA_01サンプルに対する評価指標付与結果を示す。元文書は付録A節に示す。表中カンマで区切られた2つの数字が、それぞれ2人の作業者が付与したA-Eの件数を表す。相関係数は2人の作業者の相関係数を表す。

「文法性」「指示詞」「構造と一貫性」の3つについては強い相関がみられたが、「非冗長性」と「焦点」の2つについては相関がみられなかった。表5に「文法」の、表6に「非冗長性」の、表7に「指示詞」の、表8に「焦点」の、表9に「構造と一貫性」の作業者間分割表を示す。「文法性」について対角線近くに分布しており作業者間で統制できていることがわかる。「非冗長性」・「指示詞」・「焦点」については基本的に厳しい作業者と厳しくない作業者との間に差が出ていると考える。「構造と一貫性」については評価が割れていることがうかがえる。作業者間の統制については今後検討していきたい。

表 5　文法性の作業者間分割表

	A	B	C	D	E	計
A	5	-	-	-	-	5
B	2	1	-	-	-	3
C	2	4	1	1	-	8
D	-	2	2	2	1	7
計	9	7	3	3	1	23

表 6　非冗長性の作業者間分割表

	A	B	計
A	8	1	9
B	5	-	5
C	4	-	4
D	4	1	5
計	21	2	23

表 7　指示詞の作業者間分割表

	A	B	計
A	7	-	7
B	8	-	8
C	3	-	3
D	4	1	5
計	22	2	23

表 8　焦点の作業者間分割表

	A	B	C	計
A	7	1	-	8
B	1	-	-	1
C	5	1	-	6
D	6	1	1	8
計	19	3	1	23

表 9　構造と一貫性の作業者間分割表

	A	B	C	D	計
A	4	3	1	-	8
B	-	-	-	-	-
C	3	-	2	-	5
D	7	-	-	1	8
E	-	-	1	1	2
計	14	3	4	2	23

最後に A_01 の評価事例について示す。

以下は評価が比較的高い例である：

A_01(No.18):

文法性 (A,A)・非冗長性 (A,B)・指示詞 (A,B)・焦点 (A,A)・構造と一貫性 (A,A)

> 蓮舫さんは幼いころから活発で、自分の意見をはっきり言うことができる人だった。池田弘子先生はそれを持ち前の長所だと考えて適切なアドバイスをし、蓮舫さんがキャスターになるきっかけを与えてくれた。

要約としてまとまっており、読みやすさも優れている。

以下は評価が文法性・構造と一貫性が比較的低く、指示詞・焦点の評価が一致していない例である：

A_01(No.23):

文法性 (D,C)・非冗長性 (B,A)・指示詞 (A,D)・焦点 (A,C)・構造と一貫性 (C,E)

> 蓮舫さんは思い出の先生についてこう語っている。おしゃべりだと言われていただけの自分を仕事に生かしてみたらと目を開かせてくれた。違う角度から相手の身になってくださる方だった。

以下に評価が低い理由についてのアノテータコメントを示す。

文法性：「自分を生かす」「目を開かせる」

指示詞：「仕事とは何か」、「何と違う角度からか」、「相手とはだれか」
焦点：「仕事に生かす（活かす）ことをアドバイスしたわけではない」
構造と一貫性：「『くれた』『くださる』一貫性がない」

文法性については 2 人の作業者ともに 2 文目の不自然さを指摘している。構造と一貫性については待遇表現についての指摘がある。焦点については 1 人の作業者が元文書において言及されていない点を含むことを問題視している。

以下は評価が文法性・焦点が低く、構造と一貫性の評価が一致していない例である：

A_01(No.31):
文法性 (C,D)・非冗長性 (A,A)・指示詞 (A,A)・焦点 (C,D)・構造と一貫性 (A,D)

蓮舫さんは、通っていた青山学院高等部では、ピアスをしたりしていたので、注意をする先生もいたが、二、三年時に担任だった池田弘子先生だけは、頭ごなしではなく、子どもの目線に立って聞く耳を持たせてくれた。

以下に評価が低い理由についてのアノテータコメントを示す。

文法性：「したりしていたので」「1 つの文の中で主語の違う節が多すぎる」
焦点：「先生と蓮舫さんのつながりが表わされていない」
構造と一貫性：「文を切るべき」

構造と一貫性については 1 人の作業者により 1 文中の節の多さが指摘されている。

4. おわりに

本稿では『現代日本語書き言葉均衡コーパス』を元文書とした要約文書コーパスの設計について議論した。要約元文書として BCCWJ のコアデータの PN サンプルを用い、クラウドソーシングと実験室においての被験者実験により、複数人・複数回の要約作文を収集した。収集した要約作文に対して人手による主観評価を進めている。少量ではあるが、現在までに作成した主観評価結果について検討した。

引き続きデータを拡充するとともに人手による指標付与の相関の向上に努めたい。さらに複数人間・複数回間の評価の揺れを被験者属性を含めて分析することで、最終目標である文書理解過程の多様性の定量評価を行いたい。

謝辞

本研究の一部は科研費基盤 (B)「言語コーパスに対する読文時間付与とその利用」、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

付録 A.　要約元文書 A_01 サンプル

以下に要約元文書 A_01 サンプル (PN1c_00001) を示す：

ＡＬＢＵＭ　　私の先生
キャスター　蓮舫さん
「おしゃべり」才能後押し
東京都生まれ。
９５―９７年、中国・北京大に留学し、帰国後に双子を出産。
子育てのかたわらテレビ、ラジオなどで活躍中。
３３歳。
　幼稚園から大学まで通った青山学院では、とにかく活発で、目立つ生徒だったという。
高等部では自由な校風もあって、流行に乗ってかばんを薄くつぶしたり、ピアスをしたり。
呼び出して注意する先生もいたが、二、三年時に担任だった池田弘子先生（７５）は違った。
「そんな薄いかばんじゃ遊び道具も入らないよ」「体育や部活では、危ないからピアスをはずしたほうがいい」。
やんわり語りかける。
「頭ごなしでなく、子どもの目線に立って、聞く耳を持たせてくれるんですよね」
　保健の担当でスクールカウンセラーでもあった先生の授業は、型破りだった。
障害や難病に苦しむ人の話をよく取り上げ、生徒同士で討論させた。
「世の中には様々な人がいるということが、よくわかった。
ホスピスという言葉を初めて聞いたのもこの授業でした」
　台湾人の父を持ち、「家で自己主張するよう教えられていた」蓮舫さんは、いつも率先して自分の意見を言った。
「どこかみんなとは違っていたのかもしれない」。
ほかの先生たちには、「おしゃべり」のレッテルを張られていた。
　それなのに、池田先生は言ってくれたのだ。
「しゃべるのが得意なんだから、能力を生かしてみたら」と、初めて「おしゃべり」を評価してくれた。
ブラウン管の中で話すなんて、思ってもみないころだった。
　大学に進学する時も、「あなたは論理的に考えるのが得意」と、法学部に行くよう促したのは池田先生。
大学在学中にデビューし、キャスターとして活躍するその後の進路を思うにつけ、「本当によく見ていてくれた」と感謝する。
　池田先生も、蓮舫さんにアドバイスしたことを覚えていた。
「生意気という人もいたけれど、私は、彼女のようにモノをはっきり言えることがこれからは大切だと思っていました」。
ひときわ元気だった教え子に、「持ち前の才能を生かしていってほしい」とエールを送る。

参考文献

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.