

日本語小論文データについて v1.02

竹内孔一
岡山大学学術研究院
takeuc-k@okayama-u.ac.jp

1 はじめに

本文書では日本語小論文データを使ってシステムの精度を評価するために必要となるデータの内容とディレクトリ構成について説明し，採点済み結果を利用する際の最終評点の算出法や，現段階で付与されている評点の一致度について記述する．

2 小論文データの内容と構成

小論文データは模擬試験を実施して，9個のテーマに対して1から3件の課題(プロンプト)を出し，解答者が作成した小論文を電子化して，2名で採点したデータである．解答者は大学生と少数の大学院生(脚注6参照)である．小論文課題の形式として講義を解答者が聞いた後，指示された課題に従い小論文を記述する講義形式の他に，参考文書(文献の一部やグラフなど)を読んだ後に課題に従うもの，参考文書なしで課題に対して小論文を作成するものの3種類の形式を設定した．表1にテーマと課題形式，入力形式，最大文字数，答案数を示す．

表1 小論文の各テーマと形式および答案数

テーマ	課題形式	内容	入力	講義タグ	課題	最大文字数	答案数 (白紙数)
グローバル	講義	経済	筆記	global	1	300	328
					2	250	328 (1)
					3	300	328 (1)
自然科学	講義	教育	筆記	science	1	100	327
					2	400	327 (2)
					3	800	327
東アジア	講義	経済	筆記/電子	easia	1	300	290
					2	250	290 (2)
					3	300	290 (2)
批判的思考	講義	教育	筆記/電子	criticize	1	100	290
					2	400	290
					3	800	290
人口減少	参考文書	社会	筆記	gdp1	1	150	88
					2	100	88
					3	800	88
文化相対主義	参考文書	社会	筆記	koumin	1	200	110
					2	300	110
					3	600	110 (1)
創造性	参考文書なし	教育	電子	creativity	1	600	105
民主主義	参考文書	社会	電子	populism	1	400	105
対照実験	参考文書	科学	電子	sciexp	1	150	105
					2-1	150	105
					2-2	300	105
合計(白紙含)							4824
合計(白紙無)							4815

入力形式には筆記と電子の2種類あり，筆記は紙に解答を記述する形式であり，電子はパソコン

やタブレット入力したものを指す．よって紙に書いた場合は人手により答案を電子化している．その際、存在しない漢字を記入している場合は黒丸「●」に置きかえている．最大文字数は課題内で指定している答案で記述できる最大の文字数である¹⁾．「筆記/電子」は両方の答案が存在する．筆記の答案には後の節で示すデータのディレクトリ内で pdf の画像の答案が存在するため、入力形式の異なるディレクトリ内のファイルで個別に確認できる (12 月に収集した答案の一部が電子答案)．

また、表 1 の答案数では各課題で白紙の答案数を括弧内に示している．よって白紙答案を排除した合計は 4815 件である．白紙答案の解答者の番号 (受験番号) を表 2 に示す．

表 2 白紙答案の受験番号一覧

テーマ	講義タグ	課題番号	受験番号
グローバル	global	2	20160809020
グローバル	global	3	20161226062
自然科学	science	2	20160809090 20161226062
東アジア	easia	2	20170809016 20171226060
東アジア	easia	3	20170809093 20171226060
文化相対主義	koumin	3	20171227030

この受験番号は下記の解答者のファイル名、及び採点データでの個別答案を指す場合の値として利用している．

3 小論文データのファイル構成

本節では、まずファイル構成を説明した後、各ファイルの内容と形式について説明する．小論文データのディレクトリ (フォルダ) 構成の概略を図 1 に示す．図 1 に示すように、「基本データ」と「補足データ」が存在する．「補足データ」は採点の際に採点者が各答案に対して赤ペンで答案の気になる点などをメモしたもので、各答案に対して採点者の数だけ画像 pdf ファイルが収録されている．補足のために付与したデータである．

「基本データ」以下が小論文自動採点に利用するデータである．test2016、test2017、test2018 はそれぞれ模擬試験を実施した年度を表しており、各年度で講義タグごとに小論文データが整理されている．1 つの講義タグ内には「試験問題」「答案」「採点結果」の 3 つのデータが記録されている．基本的に pdf 以外は Word または Excel ファイルでデータが構成されているので、これらを読み出すソフトウェアが必要になる．図 1 では global についてのみ下位ディレクトリを展開している．まず、global を例に 1 つの講義データ内の「試験問題」「答案」「採点結果」について説明する．

3.1 試験問題について

「試験問題」には図 2 に示すように、「参考文書」(または「参考資料」)「設問」「ループリック」「評価意図」があり、課題によっては「解答例」が存在する．「参考文書」は解答者が小論文を書く際に参考にした内容である．講義形式の課題では解答者は 30 分の講義を受けており、参考文書はその講義の内容を書き言葉で整理した内容である．講義形式ではない「参考文書」形式の場合は解答者は参考文書を読んで小論文を作成している．またテーマによっては参考文書 (または資料) は出版されている文献や図表である場合がある．その時は文献や図表の引用部分を指示している．さらに、設問の中に参考文書が記述されている場合がある (対象実験 (sciexp) の場合)．この場合は参考文書は設問内に記述している．

1) 字数制限には最大文字数を指定する課題もあれば「70-150 字」のように範囲で指定している課題もある．各課題の設問を参照．



図1 小論文データのディレクトリ構成の概略 図2 「試験問題」「採点結果」「答案」内の内容 (一部詳細化)

「設問」は解答者に対してどのような小論文を作成するか指示する内容 (プロンプト) であり、最大文字数とともに指示内容が Excel 形式で記述されている。

「ループリック」は人手で採点する際の基準を記述している。評価軸として「理解力」「論理性」「妥当性」「文書力」の4軸について1から5点で評価する際の基準を記述している (文献 [1] も参照)。図3に global におけるループリックを示す。4行目から7行目の丸数字の1から丸数字の4までがそれぞれ「理解力」「論理性」「妥当性」「文書力」の評価項目を示しており、C列からG列にかけて5点から1点のスコアとなる小論文の基準について記述している。「評価意図」は試験により解答者のどのような能力を評価したいかについて作問者の考えを記述したものである。

global_ループリック						
ホーム 挿入 描画 ページレイアウト 数式 データ 校閲 表示 操作アシスト						
D5 fx 講義で出された数字等をあげ、論拠をもって手堅い議論を展開している。						
	A	B	C	D	E	F
1	グローバルゼーションの光と影					
2						
3	設問	評価の項目	5・優れている	4・良い	3・普通	2・悪い
4	問1	① 設問に対する理解力・・・出題者が何をテーマに小論文を作成させようとしているかを理解しているか。	格差の拡大または縮小がグローバルゼーション評価の一つの基準であることに気づき、その視点から議論を展開している。	世界的な格差の縮小と国内における格差の拡大の両側面を具体かつ論理的に記述している。	グローバルゼーションの「光」と「影」の両側面に触れている。	グローバルゼーションの「光」か「影」の一方しか記述していない。
5		② 文章内容の論理性・・・文章を論理的な構成で書いているか。	経済社会の仕組みを理解し、ある現象がなぜ起きるのかを論理的に説明できている。数字による説明が必要	講義で出された数字等をあげ、論拠をもって手堅い議論を展開している。	論理の飛躍、根拠のない断定、独りよがりな論理の展開が見られず、客観的に議論を展開している。	論理の飛躍、根拠のない断定、独りよがりな論理の展開が見られ、全般的に客観的に議論を展開していない。
6		③ 見解力、知識力、専門性・・・テーマに関係する見識や専門的な知識を知っているか。	「世界的な格差の縮小と国内における格差の拡大」メカニズムの論証、「ジニ係数」の説明をしている。	世界的な格差の縮小と国内における格差の拡大の現象を論理的かつ具体的に説明している。	世界的な格差の縮小と国内における格差の拡大の理由を記述している。	格差の縮小および拡大について記述しているものの、その理由について述べていない。または、一方しか述
		④ 言語活用の正確性・・・仮名遣い、漢字や言葉の	仮名遣い、漢字の間違いがなく、指定字数をほぼ埋め	仮名遣い、漢字の間違いが1カ所ある。指定字数をほ	仮名遣い、漢字の間違いが2つある。分かりやすい	指定字数の85%に達していない。または、マスを埋め

図3 globalにおけるループリック

3.2 答案について

次に解答者の小論文を保持している「答案」ディレクトリについて説明する。「答案」ディレクトリ下では小論文を電子化した Excel 形式のデータと元々の手書き入力による pdf データが模擬試験実施日毎に整理されて保存されている²⁾。例えば 0808_global_excel 下では anstxt_20160808_global_20160808002.xlsx のようなファイル名で講義タグと日付と受験番号を利用して1解答者に対して1ファイルの答案として整理されている。この答案ファイル内では1つのテーマに対する複数の課題が「問1」、「問2」..としてまとめて記載されている。よって前節で白紙答案について記述したが、答案のxlsxファイルは1つのテーマに対して少なくとも1つの課題を解いているとxlsxが存在するため合計のファイル数は表1における合計(白紙含)となる。

具体的には図4に示すように1答案ファイルは2行目B列に日付(20160808)、C列に講義タグ(global)、D列に受験番号(20160808002)が半角英数字で記述されている。各課題に対する小論文はA列に対して「問」と数字で記載されている同じ行のB列のセル内にテキストで記述されている。解答者の元々の小論文はマス目に対して記述しているが改行や文頭に空白を入れる答案がある。その場合にはExcelの小論文テキストでは同様に全角空白や改行を入れている。よってプログラムで読み出す際にはセル内に改行が入っていることに注意する。各解答者の小論文テキストは1つ1つ別のファイルになっているため小論文テキストを収集する場合はこれら全てのファイルを対象の課題に対して取り込む必要がある。

3.3 採点結果について

次に人手による評価結果である「採点結果」ディレクトリについて説明する。採点は複数の採点者によって採点されているため、採点作業毎にExcel形式でファイルに保存されている。globalでは採点作業者はI、S、KWの3人である。それぞれの作業で8月のデータに対

2) 例えば0808は8月8日を示している。

	A	B	C	D	E
1					
2		20160808	global	20160808002	
3					
4					
5					
6					
7					
8					
9					
10	問1	グローバリゼーションは世界全体の所得格差を、これは、グローバリゼーションに伴う自由主義質昇があった一方で、先進国での競争の活発化に。			
11					
12					

図4 答案ファイル内の記述例 (anstxt_20160808_global_20160808002.xlsx)

しては 08, 12 月のデータに対しては 12 という文字列がファイル名に付与されている³⁾。よって score_global_i_08.xlsx は 作業者 I による 8 月分の課題に対する採点結果を記録したデータを示す。図 5 に score_global_i_08.xlsx の内容を記述する。まず 2 枚のシートが存在し、最初のシートが 0808, つまり 8 月の 8 日の答案に対する採点結果を示しており、後の 0809 は 9 日の答案に対する採点結果を示している。3 行目に各列の説明があり、4 行目から各答案に対する採点結果が記載されている。A 列は欠席者について赤字で欠席と記述しており、欠席の場合は一切の採点データは無

出	受験番号	問1理	問1文	問1英	問1理	問2理	問2文	問2英	問2理	問2文	問2英	問3理	問3文	問3英
4	欠席	20160808001												
5		20160808002	4	3	4	5	4	4	4	5	3			
6		20160808003	3	3	3	2	85%未満	2	光のみ(伊)	3	3	3	3	3
7		20160808004	2	3	2	5	4	4	4	4	5	3		
8		20160808005	3	3	3	4	3	3	3	3	4	3	3	3
9		20160808006	4	3	4	5	2	光のみ(伊)	3	3	2	85%未満	3	
10		20160808007	4	4	4	5	3	3	3	3	5	4		
11		20160808008	3	3	3	2	85%未満	2	3	2	1	60%未満	4	
12		20160808009	3	3	3	2	85%未満	5	5	4	2	85%未満	3	
13	欠席	20160808010												
14		20160808011	3	2	2	4	3	2	2	2	5	3		
15	欠席	20160808012												
16		20160808013	3	3	2	4	3	5	5	4	5	4		
17		20160808014	1	1	1	3	3	2	3	3	5	4		
18		20160808015	4	4	4	5	4	4	4	4	3	3		
19		20160808016	2	3	2	2	85%未満	4	4	4	2	85%未満	2	

図5 global に対する作業者 I の採点ファイル内の記述例 (score_global_i_08.xlsx)

く、また小論文も存在しない。B 列が解答者の受験番号を示している⁴⁾。C 列からが採点が記入されている。評価基準として上述の通り、「理解力」「論理性」「妥当性」「文書力」について各問に対してスコアとコメントを付与している。具体的には図 5 では C 列は「問 1 理解力」のスコア (1 から 5 点) を示しており、D 列は「問 1 理解力」に対してスコア付けをした際に付与作業でコメントがある場合にコメントを記述している。よって 2 列毎に各課題に対するスコアが付与されている。白紙答案に対してはルーブリックに従って採点しているためスコアが付与されている (大抵 1 点である)。よって白紙答案を評価から排除する場合は、白紙答案の受験番号と課題を記録して、評価の際に排除するプログラムを作成する必要がある。

3) 小論文データの収集は大学の休暇期間を利用しているため 8 月か 12 月となっている。

4) 受験番号の数字は全角の場合と半角の場合があるので注意すること。

3.4 システムを評価するための最終評点について

小論文採点システムを評価するためには各小論文答案に対して1つの評点が必要になる．本データでは2人以上で採点をしているが，様々な制約から作業期間などの違いなどで，安定している評価作業者の2名を推薦する．つまり各課題では推薦の2名の採点結果の四捨五入した整数値を最終評点とする．安定している2名の評価作業者の一覧は下記の表3に示す．

表3 各課題と評価作業者および最終評点作成のための推薦作業者

テーマ	講義タグ	評価作業者				
		KW	I	S	KK	Y
グローバル	global	O	R	R		
自然科学	science		R	R		
東アジア	easia		R	R		
批判的思考	criticize		R	R		
人口減少	gdp1		R	R		
文化相対主義	koumin		R	R		
創造性	creativity				R	R
民主主義	populism				R	R
対照実験	sciexp		R	R		O

表3は作業者は全体で5名存在し，評価したテーマに対して表中でOまたはRとして示している．その中で優先して最終評点に利用するペアをRで示している⁵⁾．例えば global では3人の作業者が評価を行っているが，最終評点に利用する作業者の点数は作業者IとSである．

4 採点データについて

各小論文の答案には「理解力」「論理性」「妥当性」「文書力」の評価軸にしたがって1から5の評点が付与されている．この4軸の大まかな考え方として「理解力」では「解答者が課題を理解して設問に即した内容を理解しているか」であり，「論理性」では書き方に注目して，「文書の構成が論理的に記述されているか」であり，「妥当性」が小論文の深さに注目して，「専門的な知見からみて深く記述されているか」であり，「文書力」が文章の文法など，「文法的に正しく記述されているか」を評価している．

そこで各課題における各評価軸に対して，表3で推薦する2名についてどの程度のゆれで付与できているかを明らかにするため QWK の値を表4に示す．このとき表2に示した白紙答案を含めずに計算する．

表4では平均 QWK をみると概ね作業者間の揺れが大きい値となっていることがわかる．文書力については各課題に対して安定して付与できている．これは文字数や誤字脱字といった形式に関する評価であるため揺れが少ないことが考えられる．また理解力も総じて安定している．これは課題に対して書くべき内容が記述されているかどうかを判定するため，書くべき内容がループリックである程度指定できたことが要因として考えられる．ただしこれは課題に依存するところであり，自分の考えに基づく内容を記述させる課題(例えば gdp1 の課題3など)ではループリックでの指示が抽象的になるため作業者間でのゆれは大きくなっている．

評価軸において論理性と妥当性について平均的には安定して付与されているが課題によっては揺れが大きい．これは論理的に書けているかをループリックで指定する際に抽象的な表現になって表現が難しい点，および論理的かどうかの判断が採点者でずれるところがあると考えられる．また妥当性については内容により深く書かれているかという点でループリックでは具体的に指定できない部分があるため採点者のその分野の理解とどこまでの表現を良しとするかの判断の部分で他の採点者と一致させるところが難しい部分があると考えられる．

5) 文献[1]で述べている2名の作業者は表3のRの作業者である．

表 4 各課題の各評価軸に対する推薦作業員間 (2 名) のゆれ

テーマ	講義タグ	課題	評価した 答案数	理解力	論理性	妥当性	文書力
グローバル	global	1	328	0.636	0.213	0.517	0.785
		2	327	0.786	0.378	0.344	0.708
		3	327	0.399	0.316	0.366	0.791
自然科学	science	1	327	0.844	0.823	0.776	0.730
		2	325	0.669	0.575	0.521	0.861
		3	327	0.468	0.516	0.516	0.887
東アジア	easia	1	290	0.671	0.506	0.647	0.839
		2	288	0.696	0.701	0.642	0.913
		3	288	0.620	0.518	0.459	0.863
批判的思考	criticize	1	290	0.885	0.887	0.818	0.712
		2	290	0.660	0.659	0.644	0.925
		3	290	0.517	0.542	0.494	0.826
人口減少	gdp1	1	88	0.691	0.684	0.652	0.676
		2	88	0.491	0.524	0.408	0.532
		3	88	0.306	0.161	0.264	0.627
文化相対主義	koumin	1	110	0.465	0.527	0.434	0.952
		2	110	0.676	0.717	0.621	0.918
		3	109	0.523	0.462	0.422	0.978
創造性	creativity	1	105	0.769	0.668	0.322	0.654
民主主義	populism	1	105	0.450	0.488	0.497	0.658
対照実験	sciexp	1	105	0.442	0.535	0.230	0.717
		2-1	105	0.746	0.708	0.746	0.911
		2-2	105	0.791	0.792	0.790	0.409
平均			4815	0.617	0.561	0.527	0.777

採点システムなどを評価する際は上記の QWK を参考に利用できる部分を利用していただければと思う。また、独自に評点を付与する場合にも、これらの付与結果を参考にしていただければ幸いである。

5 その他の情報

5.1 小論文データ作成に関して

小論文の解答者は大学生および大学院生⁶⁾である。またオンラインなどではなく実際に教室に学生を集めて決められた時間内に解答を記述する。例えば 2016 年、2017 年に実施した global から criticize までは講義を 30 分聞いた後、1 時間かけて 3 問の課題に対する小論文を記述する。その他の課題も時間を十分与えているため、英語ライティングの試験の様に、時間切れで小論文が書けていないという例は答案には少ない。電子入力ではパソコン及び iPad を利用した入力を行った。その際、仮名漢字変換ソフトの利用が可能であるため、解答を記入する際に誤字が少なくなっていることが想定される。

5.2 掲載予定の文献 [2] との比較について

現段階で掲載予定であるが文献 [2] では本小論文データを利用して様々な採点手法を適用してシステムと人手による採点評価を行っている。掲載された場合で比較を行いたい場合に必要となる情報を記述する。文献 [2] では評価の際、模範答案を各課題で 1 件だけ削除している。よって比較のために全く同じテストデータを作成するには表 5 に示す模範答案をテストデータから削除することで可能になる。また、文献 [2] ではシステムの QWK を評価する際、システムの出力を 1 から 5 点に整数化してから評価している。よって評価する際にはシステムの出力を整数化した値で評価する

6) 2018 年の解答者に 10 人以下で大学院生の答案が存在する。

表 5 文献 [2] で評価から除外した答案の受験番号

テーマ	講義タグ	課題	除外した答案の 受験番号
グローバル	global	1	20160809011
		2	20161226022
		3	20160808036
自然科学	science	1	20160808015
		2	20161226049
		3	20160808018
東アジア	easia	1	20170809020
		2	20170809034
		3	20170809057
批判的思考	criticize	1	20170809020
		2	20171226006
		3	20171227025

と同じ条件となる。

5.3 小論文データの確認日

上記の内容は 2021 年 8 月 31 日現在のものである。

6 おわりに

日本語小論文データの内容について記述した。小論文のテーマと課題、さらに答案データや採点データの形式について説明し、採点作業者のゆれを QWK を用いて評価した。本日本語小論文データ利用の際には言語資源協会の該当サイトを引用いただくか、掲載されれば文献 [2] を引用していただければ幸いである。本データが小論文の採点支援システムの開発および自然言語処理の発展に役立つことを期待する。

参考文献

- [1] 竹内孔一, 田口雅弘, 稲田佳彦, 飯塚誠也, 阿保達彦, 上田均. 模擬試験による研究利用可能な日本語小論文データの構築と採点. 電子情報通信学会 言語理解とコミュニケーション研究会 (NLC) 信学技法, 2021 (掲載予定).
- [2] 竹内孔一, 大野雅幸, 泉仁宏太, 田口雅弘, 稲田佳彦, 飯塚誠也, 阿保達彦, 上田均. 研究利用可能な小論文データに基づく参照文書を利用した小論文採点手法の開発. 情報処理学会論文誌, 2021 (掲載予定).