構造的曖昧性に基づく読みづらさの検出

吉田あいり　河原大輔
早稲田大学理工学術院
yoshida-a.waihk@ruri.waseda.jp, dkw@waseda.jp

概要

日本語文の読みづらさを定量的に評価することを目的とし, 構造的曖昧性を活用した手法を提案する. 漸進的要素を組み入れ, BERT による構文解析の結果を用いて評価する. ベースラインには言語モデルによる尤度 (サプライザル) を用いて, 読みづらさの正解データはクラウドソーシングにより文ペアに対して複数人の回答を収集したものを使用する. 2 種類の方法で作成したデータにて評価を行い, 提案手法はベースラインに対して有意な結果を得た.

1　はじめに

テキストを読む機会は多く, 文が読みやすいことは消費される時間の短縮以外に読者の理解度の向上にもつながる. そのため, 文の情報を損なわない読みやすい文が必要とされる. 本研究では執筆支援システムにおいて想定される読点の挿入や語順整理といった文の読みやすさの改良の前段階として, 自動的に読みづらい文を検出することを目的とする.

人間における読みづらさは主に読み時間で評価され, 計算機においては言語モデルを用いて計算される負の尤度であるサプライザルで評価される [1]. これは予想と異なる内容が現れた際に読み負荷が高まるというサプライザル理論に基づいている [2, 3]. ここに日本語特有の要素を考慮することで, 本研究では日本語に適した読みづらさの検出を目指す.

日本語の逐次的な読みやすさにおける先行研究にて, 先行文脈に係り元が多い文節ほど読みやすいことや読み間違いが読み負荷を高めることがわかっている. 例えば, 文 (1) [4] は「男性を」の部分で再解釈に要する読み時間が増加する.

(1)警官が犯人を捕まえた男性を...

本研究では構造的曖昧性に着目し, 漸進的構文解析結果と係り先の保留のプロセスを利用して読みづらさの検出を行う.

2　関連研究

2.1　読みやすさの分析

人間の読みづらさは読み時間を用いて評価されており, 読み時間を主軸とした分析がされる. 日本語の読み時間データは BCCWJ-EyeTrack が整備され, 先行文脈に係り元文脈が多い要素ほど読み時間が短くなるということが分析されている [5]. また, 日本語の読解時間と統語・意味カテゴリの対比分析においても関連する先行文脈が読み時間の減少に関与すること [6] や, 文の意味的曖昧性の高さが構造的曖昧性の解消/保留に影響することがわかっている [4, 7]. 他にも個別の言語現象が読み時間に与える影響に対して分析が行われている [8, 9].

これらは読みやすい文における知見ではあるが, 統一的な傾向は定かではない. そこでサプライザル理論に基づいて, 日本語読み時間に関する傾向がサプライザルが大きいところで読みにくくなるという傾向に統一的に解釈できるかという仮説検証がなされた [10]. 読み時間の様々な傾向がサプライザルでも再現され, 情報量の観点から解釈できることが示されている.

2.2　ガーデンパス効果

読み時間が増加する要素の 1 つとしてガーデンパス (GP) 効果がある. これは, 文の理解の途中で一時的に誤った解釈をし, 続きを読んだ際に誤解に気づいて解釈をし直す事になるが, この再解釈により発生する処理負荷や読み直しにかかるコストにおける効果である. 文理解の初期段階で曖昧性を解消しようとするために, 結果として誤解釈が発生する. この効果がもたらす構造的曖昧文の性質は言語の構造により異なることが分析されている [11]. GP 文は最後まで解釈が曖昧な場合もある. 例えば,「かわいい少女の猫」の「かわいい」が修飾するのはこの部分のみでは決定できない.

2.3 単語親密度

本研究では構造要素による読みづらさに着目するために, 単語難易度による影響を減らす工夫が必要である. WLSP-familiarity[1)]は「分類語彙表」増補改訂版データベースに親密度情報を付与したもの [12] であり, 見知らぬ単語による読みづらさを軽減するためにこれを用いてフィルタリングを行う.

2.4 漸進的係り受け解析

人間の言語理解過程には漸進性があり, 音声言語アプリケーションの基盤技術として漸進的係り受け解析技術の研究がなされている [13]. 言語処理過程のデータを構築・分析することにより人間の漸進的係り受け解析能力や入力予測能力の解析がなされている [14, 15]. 逐次処理を行うために, 係り先を未入力部分と見做し, 未入力の中でも同一文節であるのかまで考慮してアノテーションが行われている. 人間の言語理解には漸進性がある [16] が, 特に音声では入力と同時に処理することが求められ, 漸進的言語処理システムが開発されている. 本研究ではテキストに漸進性を組み込むことで, GP 効果等の特徴を考慮する.

3 構文解析

読みづらさを検出するために, 構造的に曖昧な文を示すラベル付きデータなしで活用できる機械学習を用いた構文解析を行う.

3.1 BERT に基づく構文解析

BERT を活用し, 日本語構文解析の精度を向上させる手法 (BERTKNP) が提案されている [17]. 構文解析を head [18] の選択と見なすことにより BERT に 1 層追加する形で実装を行っている. 事前学習されたモデルを活用することで, 大量の生コーパスを利用し構文解析の精度向上を達成している.

3.2 漸進的構文解析手法

BERT による構文解析に漸進的な視点を入れることで, 人間が読む際の要素を追加する. 単語ごとに係り受けのアノテーションがなされた文 (Original) に対して, 単語が 1 つずつ入力される過程 (Slided) を考えていく. ある単語が入力されるまでの部分における, 係り受け構造を学習データとして作成する. 入力

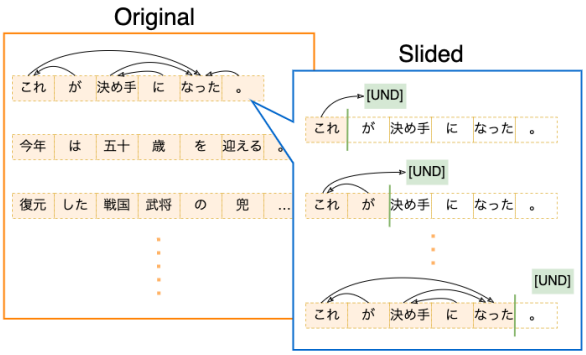


図 1: 漸進的構文構造データ作成

される単語数を 1 ずつずらしていくことで漸進的なデータを実現する. 図 1 のように Slided の入力は文の途中までであり, 未入力部分に係る単語も存在する. その場合は係り先を特定単語ではなく, 未決定 ([UND]) と指定する [14]. 単語ごとにずらして作成した文は Original に対しての分量が非常に多くなるため, Slided について過学習を起こす可能性がある. そこで Slided には n%の制限を設ける. どの程度漸進的データを追加するべきかとシャッフルの必要性は予備実験にて確認する.

3.3 解析結果

漸進的構文解析において適切な漸進的データ量を確認するために, 実験の結果を表 1 に示す. 精度は単語単位での精度と文単位の精度の平均の 2 通りで算出した. 加えて [UND] の再現率も測定した. 日本語 BERT は NICT BERT 日本語 Pre-trained モデル[2)]を使用した. fine-tuning には京都大学テキストコーパス (新聞) と京都大学ウェブ文書リードコーパス (ウェブ) の 2 種類を混合したコーパス約 5 万文を使用し, 元論文にならい 3 epoch 回した. また, 評価には京大コーパス約 4,000 文を使用した.

京大コーパスにおいては加工しない BERTKNP が一番精度が高い. 漸進的要素を加えた Slide All (Orig+Slided 100%) においては元のデータに 5%混合したデータを用いて学習したモデルが一番精度が良く, 再現率は 50%混合した場合が良い値となった. また, 予測精度と再現率はデータをシャッフルすることによる悪影響を受けることがわかった. 漸進的データの追加は精度の向上に直結しない上, 計算資源と時間を消費する. 本研究では, 精度を重視して 5%のみ追加したデータで fine-tuning したモデルを用いてこれ以降の実験を進める.

1) https://github.com/masayu-a/WLSP-familiarity

2) https://alaginrc.nict.go.jp/nict-bert/index.html

表 1: 漸進的構文解析の精度

モデル	京大 (新聞)	Slide All	UND Recall
BERTKNP	**0.965 / 0.964**	0.760 / 0.852	0.000
Orig+Slided 5%	0.963 / 0.963	**0.958 / 0.956**	0.889
Orig+Slided 20%	0.961 / 0.961	**0.958** / 0.955	0.891
Orig+Slided 50%	0.959 / 0.960	**0.958** / 0.954	**0.892**
Orig+Slided 90%	0.957 / 0.958	0.957 / 0.953	0.890
Slide All	0.958 / 0.958	0.957 / 0.953	0.888
Slide All (shuf)	0.957 / 0.957	0.956 / 0.952	0.887

表 2: 各データの文ペア数

データ	n / m	A	B	同等	total
多数決	5	75	100	72	247
	6	42	54	43	139
	7	21	32	18	71
平均	0	131	160	-	291
	1	100	125	-	225
	2	67	91	-	158
	3	52	70	-	122

4 実験

4.1 正解データ作成と評価方法

読みづらさの正解データを作成するために, 京大コーパスに前処理として単語親密度によるフィルタリングを行い, 親密度が負のスコアを持つ形態素を含む文を除外した. また, 形態素数が 20 に満たない文も除去し, 形態素数が 20 から 30 のものと 30 以上のものに分割後, 文長によるソートを行ってから文字数順に 2 文のペアを作成した. これにより, 単語難易度による難読性を排除した上で, 文長だけでなく形態素数の近い分ペアを作成している. このペア文の比較による評価を行う. Yahoo!クラウドソーシングにより 1 ペアに対して 10 人の回答をフィルタリングされた 322 文について収集した. 2 文のうちどちらが構造的に読みづらいかという質問に対して, [文 A, 同等, 文 B] の 3 つの選択肢を用意した.

文法構造を考慮した読みづらさとして以下の 2 つの要素を挙げ, これに基づき評価させた.

1. 複数の解釈ができる
2. 理解に反復が必要となる

収集データから 2 種類の正解データを作成し, それぞれのデータにおいて評価を行う. 作成された正解データのペア数を表 2 に示す.

多数決 10 人中 n 人の回答が一致したペア文を正解データとする. 例えば, [文 A, 同等, 文 B] の回答数が [6,3,1] の際は, n=5 の場合は文 A が読みづらいデータとして採用され, n=7 の場合は使用しない. 評価は [文 A, 同等, 文 B] の全ての精度により行う.

平均 [文 A, 同等, 文 B] の回答を [-1,0,1] の数値と見なして各回答数の和をスコアとする. 閾値 m を使用し, $[-10, -m)$, $[-m, m]$, $(m, 10]$ をそれぞれの回答として分類する. m の値は難易度の判別がつかない文ペアが十分除かれたものを選択する. 正負でスコアリングしているため, [文 A, 同等, 文 B] の回答数が [2,7,1] と [5,1,4] の場合のスコアはどちらも-1 となる. 従って, 同等である回答が多くとも判別できないため, 文 A・B のどちらが読みづらいかのみを扱い, 同等なものは除外した精度により評価する.

多数決・平均手法共に, n/m を大きくするほど確実に読みづらい文ペアを取得できるが, 使用できるペア数が減少するため適当な n/m を抽出した.

4.2 読みづらさの検出方法

本研究では, 作成した漸進的データを用いて学習した構文解析モデルによる解析結果から係り先が未決定 ([UND]) である数を長さで正規化したものを読みづらさのスコアとする. 人間の言語理解の漸進性を取り入れた上で, 係り受け構造の保留により受ける読みづらさを考慮している. 本研究では 2 文のどちらが読みづらいかまたは同等であるかを正解として扱うが, スコアは数値で算出される. そこでスコアの差分を取り, その絶対値が閾値以下のものを同等である場合として評価する. 閾値は精度が最も高くなる値を選んだ.

ベースラインにはサプライザルとして $-\log p(x|$ 先行文脈 $)$ を使用する. 実装では –log_softmax を使用した. サプライザルは尤度の反転であり, これが大きいほど読みづらいと言える. これを計算するために言語モデル GPT-2[3] を利用した. こちらもスコアの差分から同等である場合の評価を行った.

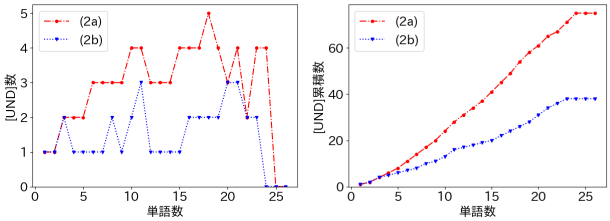
4.3 結果

表 3 に各データにおける精度の結果を示す. 正解は漸進的構文解析が全て正しく付与されたデータを用いて算出した精度である. 提案手法は多数決と平均どちらのデータに対しても読みづらい文の検出をすることができた. 確実に難易度に差があると判断された文ペア (多数決 n=7, 平均 m=3) に対しては特に有意な差が現れた.

3) https://huggingface.co/colorfulscoop/gpt2-small-ja

表 3: 読みづらさ検出の精度

データ	n / m	ベースライン	提案手法	正解
多数決	5	0.360	**0.368**	0.397
	6	**0.396**	**0.396**	0.453
	7	0.394	**0.451**	0.521
平均	0	0.478	**0.485**	0.519
	1	0.476	**0.493**	0.542
	2	0.456	**0.483**	0.557
	3	0.410	**0.500**	0.549



(a) 各単語長でのカウント　　**(b)** 各単語長での累積

図 2: [UND] 数の比較

5 考察

5.1 難しい文と平易な文

構造的に難しいと判別された文の特徴として主語が省略されているものや, 並列構造等が見られた. 文長が短いために一定数主語が無い文が存在しており, これらを排除することでより構造に焦点を当てた実験を行うことができると考えられる. また, 長い名詞句やカタカナは中身を読まずに塊であると判別できるため簡単と評価される傾向があった.

ここで, 図 2 にて難易度判定により有意な差を得た文における係り先未決定の動きを見る. 以下の文 (2a) は 10 人中 9 人が読みづらい文として選択し, 1 人が同等であると回答した.

(2) a.しかし、旧民社党は大半の議員が新進党に参加し、さきがけとの連携も流動的で連携相手は不確定だ。
b.初期の警察署が置かれたセントラル地区のハリウッド通りには、インド料理店やインド系企業が多い。

(2b) は入力に適宜係り先が決定されているが, (2a) は保留されていることがわかり, 文の構造的難易度により未決定数に差が生じることが確認できる.

5.2 読みづらさの正解データ作成

多数決方式の n=5 においては [文 A, 同等, 文 B] の回答数が [1,4,5] の様な回答が存在し, 小さい n に対しては難易度の定まらない文が混在していると考えられる. 平均方式はこの欠点はカバーできるものの, 同等の文に対しての評価ができなくなる.

多数決方式にて n の値により収集される文ペアを見ていく. n=5 の例として (3), n=7 の例として (4) を示す. (3a) と (4b) が読みづらい文である.

(3) a.九七年まで村山政権が続くのは安定かもしれないが、確信を持って政治がやれるのか。
b.この時は大分舞鶴がロスタイムに入って追いつき、抽選で長崎北陽台が決勝に進んだ。

(4) a.同県警では、県警山岳救助隊など約七十人で捜索したが、午後五時、捜索を打ち切った。
b.本に囲まれ、埋もれ、重みで床が抜けそうだ、と心配するような生活にあこがれている。

(3) は文構造が似通っており難易度にさほど差が感じられないが, (4b) には並列構造が見られ, 係り先の理解に時間を要する. n を大きくすることで確かに難易度差がある文ペアを取得している.

5.3 校正における活用

構造的曖昧性の絶対評価は難しく, 本研究では 2 文の比較による評価を行った. 難易度のスコアはサプライサルと提案手法のどちらも各文ごとに付与している. そのため読みづらいと判別された文のスコアの平均・分散における解析が, その文単体での難易度の判別に繋がる. これを用いて難しい文の指摘を行うことで筆者による修正や自動書き換え等の校正への活用が見込まれる.

6 終わりに

日本語構造に着目した読みづらさの検出を行なった. 漸進的構文解析による, 未入力文脈への係り構造を利用することで, サプライザルに比べて構造的に読みづらい文の検出を行うことができた. 今後は構文解析結果の top-K 比較による構文確率の比較や語義曖昧性への拡張などによる読みづらさの検出を検討する. また, 特定の曖昧性に的を絞った正解データの用意は困難であり, 収集方法も改善の余地がある.

参考文献

[1] Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. **bioRxiv**, 2020.

[2] John Hale. A probabilistic earley parser as a psycholinguistic model. In **Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies**, NAACL '01, p. 1–8, USA, 2001. Association for Computational Linguistics.

[3] Roger Levy. Expectation-based syntactic comprehension. **Cognition**, Vol. 106, No. 3, pp. 1126–1177, 2008.

[4] 井上雅勝. 文の意味的曖昧性が構造的曖昧性の解消と保留に及ぼす影響 (2). 日本認知心理学会発表論文集, Vol. 2011, pp. 74–74, 2011.

[5] 浅原正幸, 小野創, 宮本 エジソン正. Bccwj-eyetrack : 『現代日本語書き言葉均衡コーパス』に対する読み時間付与とその分析. 言語研究, No. 156, pp. 67–96, 2019.

[6] 浅原正幸, 加藤祥. 読み時間と統語・意味分類. 認知科学, Vol. 26, No. 2, pp. 219–230, 2019.

[7] 井上雅勝. 文の意味的曖昧性が構造的曖昧性の解消と保留に及ぼす影響. 日本認知心理学会発表論文集, Vol. 2010, No. 0, pp. 81–81, 2010.

[8] Masayuki Asahara. Between reading time and clause boundaries in japanese— wrap-up effect in a head-final language—日本語の読み時間と節境界情報—主辞後置言語における wrap-up effect の検証—. **Journal of Natural Language Processing**, Vol. 26, pp. 301–327, 06 2019.

[9] Masayuki Asahara. Between reading time and the information status of noun phrases 名詞句の情報の状態と読み時間について. **Journal of Natural Language Processing**, Vol. 25, pp. 527–554, 12 2018.

[10] 栗林樹生, 大関洋平, 伊藤拓海, 吉田遼, 浅原正幸, 乾健太郎. 日本語の読みやすさに対する情報量に基づいた統一的な解釈. 言語処理学会 第 27 回年次大会発表論文集, pp. 723–728, 2021.

[11] 井上雅勝. ガーデンパス現象に基づく日本語文理解過程の実証的研究. 大阪大学博士論文, 2000.

[12] 浅原正幸. Bayesian linear mixed model による 単語親密度推定と位相情報付与. 自然言語処理, Vol. 27, No. 1, pp. 133–150, 2020.

[13] Tomohiro Ohno and Shigeki Matsubara. Dependency structure for incremental parsing of Japanese and its application. In **Proceedings of the 13th International Conference on Parsing Technologies (IWPT 2013)**, pp. 91–97, Nara, Japan, November 2013. Assocation for Computational Linguistics.

[14] 大野誠寛, 松原茂樹. 漸進的係り受け解析の出力構造 -人間の文解析過程のアノテーション-. 言語処理学会 第 22 回年次大会 発表論文集, pp. 457–460, 2016.

[15] 後藤亮, 大野誠寛, 松原茂樹. 人間の漸進的言語処理能力の分析. 情報処理学会 第 82 回全国大会, pp. 457–458, 2020.

[16] Gerry Altmann and Mark Steedman. Interaction with context during human sentence processing. **Cognition**, Vol. 30, No. 3, pp. 191–238, 1988.

[17] 柴田知秀, 河原大輔, 黒橋禎夫. Bert による日本語構文解析の精度向上. 言語処理学会 第 25 回年次大会発表論文集, pp. 205–208, 2019.

[18] Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. Dependency parsing as head selection. In **EACL 2017**, pp. 665–676, 2017.