日本語 WiC データセットの構築と読みづらさ検出への応用

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

本論文では, 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, 語義曖昧性に基づく手法を提案する. 2 文に含まれる同じ単語の語義が一致するかを判定する WiC (Word in Context) データセット [1] の日本語版である JWiC を構築し, このデータセットを使用した語義曖昧性判定に基づく読みづらさの検出を行う. 評価はクラウドソーシングを活用し, 約 7 割の精度で読みづらさの検出ができることを確認した.

1　はじめに

文章を読む機会は多く, その文章が読みやすいことは消費時間と理解力に良い影響となる. そのため, 執筆者の意図が誤って伝わる要素の排除が望まれる. 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善の前段階として, 読みづらい文を自動的に検出することを目的とし, 改善自体は執筆者に委ねる.

読みづらさを扱うにあたり曖昧性に着目する. 曖昧性は構造的なものと意味的なものがある. 構造的曖昧性とは構造解釈に複数の選択肢があることであり, 我々はこれに着目した読みづらさの検出手法 [2] を提案した. 意味的曖昧性の一つに語義曖昧性がある. 例えば「動きを踏まえた分析」という文が与えられた時に「動き」が傾向を表すのか実際の活動を表すのかの判別が付かないというようなものである. 本研究では語義曖昧性に着目し, まず, 2 文に含まれる同じ単語の語義が一致するかを判定する WiC (Word in Context) データセット [1] の日本語版である JWiC を構築する. JWiC は日本語フレームネットから構築するが, それだけだとデータ数が少ないため, データ拡張を行う. 最後に, 拡張した JWiC を使用した語義曖昧性判定に基づき読みづらさの検出を行う.

2　関連研究

2.1　WiC

言語理解のためのベンチマークとして SuperGLUE [3] が構築されており, 8 つのタスクのうちの 1 つである語義曖昧性解消タスクが WiC (Word in Context) [1] である. 2 文に含まれる多義語である同一単語が同じ語義で使用されているかを判断する 2 値分類のタスクである. 多言語 WiC データセットとして XL-WiC [4] が整備されているが, 完全に整備されている言語はドイツ語, フランス語, イタリア語に止まり, その他の言語は dev と test データのみである. 日本語もこれに含まれているが, 1,000 組程度と数が少なく, 言語資源として活用するのは困難である. 従って, 日本語データセットの構築が求められているのが現状である.

2.2　読みづらさの分析

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・分析されており [5], 読解時間と統語・意味カテゴリの対比分析 [6] もなされている. また, 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 [7, 8]. 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 [9, 10]. また, サプライザル理論に基づいて統一的に解釈できるかという仮説検証もなされている [11]. 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.

これらは既存の読みづらさの研究であり, 我々は構造的曖昧性を利用して読みづらさを検出する手法を提案した [2]. 構造的曖昧性とは構造解釈に複数の選択肢があることである.

3 日本語版 WiC の構築

本研究では, 日本語フレームネット (JFN) [12] を用いて日本語版 WiC (JWiC) の構築を行う. オリジナルの WiC は多様性とバランスに注目しており, JWiC においてもこれを重視し収録語彙の汎用性を高めることを目指す. 以下では, JFN の特徴の確認からはじめ, 構築の流れ, モデルによる評価までを述べる.

3.1 日本語フレームネット (JFN) の概要

JFN は日本語における語彙・複合言語資源であり, 言語形式とその意味の関係を背景知識 (フレーム) との関係で捉えている. 対象語句を含む例文にフレームがアノテーションされている. 例えば「持つ」に関する以下の例文には次のフレームが付与されている.

1. …報告を持ってきた…：Bringing
2. …在庫の持ち方を変えて…：Storing
3. …特許権を持っています。：Possession

収録語彙は 5,000 語程度であり, 名詞が半分以上を占め, 名詞と動詞でほとんどを占めている. また, 単語ごとに定義されているフレーム数にも偏りが大きく, 最大で 11 フレームを持つ. フレームが付与されている例文の長さは 50 字程度が最も多く, 大抵の例文は 200 字以内の長さをとっている. また,「雇用再生集中支援事業」のような固有名詞が含まれていることから, 対象が汎用語句に留まらないことが問題であり, 活用にあたってフィルタリングが必要となる.

3.2 JWiC の構築手法

JWiC の構築は, JFN 例文のフィルタリング, 例文のペアリング, 作成したペアのクラウドソーシングによる検証という 3 段階で行う.

JWiC を特定単語に特化したものではなく, 汎用性を高めるために対象語句を制限する. 適切な汎用性を保つために, 対象語句の構成単語数とフレーム数に着目する. 本研究では 4 単語以上で構成されるものと, 6 フレーム以上定義されている語句を除くこととした. 次に, 対象語句の制限に続いてバランスの調整を行う. 例文は, 文として成立していることと簡潔であることが求められるため, 例文長は 15 字以上 100 字未満とする.

WiC の形式にするために, 対象語句ごとに例文のペアを作成する. ここで, 1 対象語句に対して例文ペア数は 50 ペアまでと制限することで偏りを少なくする. ペアである 2 文が持つフレームが一致するか否かで WiC 用のラベル「同じ」「違う」を付与する. 例を表 1 に示す.

表 1: JWiC の例

対象	ラベル	例文ペア
聞く	違う	社内で聞いてみても誰もわからない。 奥さんがまだお元気と聞いておりますが」
見る	同じ	「ただ、見てるだけでないのか！！ 窓辺にたち、海のほうをみました。

表 2: クラウドソーシングの結果における正答率の平均と四分位数

文ペア	平均	最小	25%	50%	75%	最大
3,230	0.590	0	0.350	0.650	0.850	1.00

作成したデータに関して, 付与されたラベルの妥当性を確認するためにクラウドソーシングによる検証を行う. ここで, 人手であらかじめ正解ラベルを付与したチェック設問の正答率を確認することでクラウドワーカーの品質を保証する. チェック設問は最も正答率が低いもので 70%を上回るものを使用し, 平均 8 割程度の正答率のものを選択した. クラウドワーカーには短文によって例を提示し, チェック設問があることを説明している. 文ペアを提示し, 対象語句の語義が「同じ」か「違う」かを選択してもらうが, 適当な回答をできる限り取り除くために「わからない・判断できない」という選択肢も用意した. クラウドソーシングには Yahoo!クラウドソーシングを使用し, 1 ペアに対して 20 人から回答を回収した. 解答の結果は表 2 の通りで, クラウドワーカーの正答率のばらつきを確認するために四分位数を示す.

この結果から極端に精度の悪い文ペアを除いて JWiC の完成とする. JFN フレーム付けは細かく, 見分けのつかないものも含まれるため正答率がさほど高くならなかったと考えられる. 3,230 ペアに対して精度が 0.35 以上のものを採用し 2,495 ペアを得た. 含まれる対象語句は 259 個である.

3.3 モデルによる評価実験

JWiC を用いて, 事前学習モデル BERT [13] をファインチューニングし, 評価実験を行う. JWiC は, 2 文における同一語句が同じ語義で使用されているか否かを判定する 2 値分類タスクである. そのため, 入力文中のどの語句が対象であるかを明示する必要がある. そこで, BERT の埋め込みに対象語句を示す

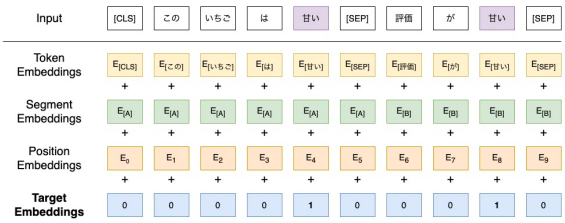


図 1: Target Embeddings の導入

Target Embeddings を導入する. 図 1 に概要図を示す.

JWiC のデータは [train:validation:test] = [0.75:0.1:0.15] の比率で分割し, NICT BERT (BPE あり)[1] をファインチューニングした. 表 4 において, JWiC の行が対象語句を明示しないモデル, JWiC_{Target} の行が Target Embeddings を使用したモデルを表し, test セット列の JWiC が分類精度である. Target Embeddings を用いることでモデルが対象語句を認識し, 1.1%の精度向上が見られた.

4 JWiC の拡張

前節で構築した JWiC は 259 個の対象語句からなり, 読みづらさ検出に応用するには対象語句が少ない. そのため JWiC のデータ拡張に取り組む.

4.1 単語親密度を用いたデータ拡張

データ拡張は, 追加する対象語句の選定, 例文の収集・ペアリング, 作成したペアのアノテーションという 3 段階で行う.

追加する対象語句の選定にあたり, 読みづらさ検出を執筆支援システムに応用することを考慮する. 筆記において使用される語句による拡大を行うために単語親密度を用いる. 「分類語彙表」増補改訂版データベースに親密度情報が付与された WLSP-familiarity[2] を使用する [14]. 対象単語の選別に単語親密度の「書く」の項目を使用し, ある程度筆記に使われる語句を採用し, かつ解釈に迷わないものは省くために 0.5 から 1.2 未満[3] のものを対象とし収集した. 特殊文字を含むものやカタカナを含むもの, また, 2 単語以上からなるものを削除した. 対象をさらに使用頻度が高いものとするために, 品詞を動詞と形容詞に絞り, 最終的に 700 語句を収集した.

得られた対象語句に対して, 対象語句を 1 つのみを含む例文を「用例.jp[4]」から収集する. 収集した例文からペアを作成する. 同じ対象語句に対する例文ペア数は 10 ペアまでとし, JWiC にすでに含まれる対象語句に関しては数を減らすことで追加データに関してもバランス調整を行う.

表 3: 拡張データのラベル分布

全ペア	同じ	違う	棄却	採用ペア
5,222 ペア	2,593	1,793	836	4,386 ペア

得られた例文ペアに対して, 対象語句における語義が一致するか否かのアノテーションをクラウドソーシングで行う. 3.2 節と同様に, クラウドワーカーに [同じ, 違う, わからない・判断できない] の 3 択から選択してもらう. チェック設問は 3.2 節と同様のものを使用した. 5,222 ペアに対して 10 人分の回答を収集した.

得られた回答集合を用いて, 回答比率を考慮した平均スコアに基づきラベル付けを行う. まず, [同じ, 違う, 不明] の各回答を [1, -1, 0] の数値と見做して, 10 人の回答の和をスコアとする. 閾値 m を使用し, $[-10, -m)$, $[-m, m]$, $(m, 10]$ の範囲において [違う, 棄却, 同じ] と分類を行う. m の値はラベルの偏りが大きすぎず, かつ判別のつかない文ペアが十分に除かれるように選択する. 本研究では閾値 $m = 2$ を採用した. 拡張したデータのラベル分布を表 3 に示す.

4.2 モデルの再学習と評価

データ拡張の前後における JWiC の精度変化を確認する. まず, 拡張したデータを学習データにのみ追加した場合の精度を表 4 の JWiC+ExTrain および $\text{JWiC+ExTrain}_{Target}$ の行に示す. データ拡張によって, 元の JWiC の対象語句についてはドメイン外の学習データが増えたが, Target Embeddings を用いたモデルは精度が落ちなかった.

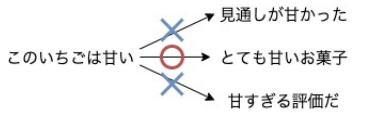
次に, 拡張したデータを JWiC の train, validation, test セットに追加し評価した. この新しい test セットによる精度を表 4 の JWiC+Ex 列に示す. 新しい train セットで学習したモデルの精度を JWiC+Ex と JWiC+Ex_{Target} の行に示す. 元の JWiC, JWiC_{Target} モデルと比べて精度が向上していることが分かる. また, 各モデルにおいて Target Embeddings はやはり効果があることを確認した.

次節で述べる語義曖昧性判定による読みづらさ検出と評価においては, 拡張データを含めて学習した

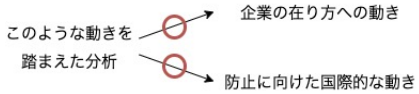
1) https://alaginrc.nict.go.jp/nict-bert/index.html
2) https://github.com/masayu-a/WLSP-familiarity
3) 単語親密度は標準正規分布に従う.
4) https://yourei.jp/

表 4: 各モデルの分類精度

モデル	test セット JWiC	JWiC+Ex
JWiC	0.886	0.700
JWiC$_{Target}$	0.897	0.724
JWiC+ExTrain	0.872	-
JWiC+ExTrain$_{Target}$	0.900	-
JWiC+Ex	-	0.757
JWiC+Ex$_{Target}$	-	0.781



(a) 曖昧でない場合



(b) 曖昧な場合

図 2: 語義曖昧性検出の例

BERT モデル (Target Embeddings あり) を使用する.

5 読みづらさの検出

5.1 読みづらさの検出方法

読みづらさの検出は, 入力文中の対象語句が語義曖昧性を持つかを判定することによって行う. 語義曖昧性の判定は, 対象語句が持つ語義ごとに例文 1 つと入力文で文ペアを作り, WiC タスクを実施することによって行う. 複数の語義の例文と「同じ」と判定すれば複数の語義の可能性があり曖昧, 1 つのみと同じであれば曖昧でないとする. 図 2a の場合では,「このいちごは甘い」という入力文は「とても甘いお菓子」のみと一致しており, 入力文中の「甘い」において使用されている語義が明白である. 一方, 図 2b の場合では, 傾向を示す「動き」と実際の活動を示す「動き」のどちらとも一致するため, どの語義で使用されているかが不明となり, 曖昧と判定される.

5.2 評価用データセットの作成と評価

JWiC から評価用データセットを作成する. 対象語句が持つ全ての語義との比較により評価するため, 各語義から例文を 1 文ずつ選択する. [A, B] という 2 種類の語義を持つ場合は, [A, A, B] の語義が付与された例文セットを抽出し, 語義 A を持つ例文 1 文を評価対象文とする. これにより 184 セットを得た.

表 5: 閾値ごとのラベル分布と語義曖昧性の検出精度

閾値 n	6	7	8
曖昧	114	123	137
曖昧でない	70	61	47
精度	0.701	0.717	0.663

次に, 各セットにおいて評価対象文が語義曖昧性を持つか否かのアノテーションを行う. これは, JFN のフレーム情報を用いるのではなく, 一般的な人間による語義曖昧性の判断を得るために行う. 評価対象文に対して, どの例文が同じ語義であるかの複数選択式でクラウドソーシングにより 10 人分の回答を収集した. この回答をもとに 2 つの基準を設けて, 正解ラベルを作成した. 1 つ目は回答が分散していれば曖昧であり, 2 つ目は複数の語義の例文を選択していれば曖昧であるという基準である. まず, 最も回答数が多い選択結果 (セットで扱う) に対して n 人以上の回答が一致したものは回答が分散していないとみなし, n 人未満であれば分散しているため曖昧とみなす. n は 6 から 8 の値において試行した. 次に, n 人以上が一致した選択結果において, 複数の例文が選択された場合は曖昧であり, 単一であれば曖昧でないとする.

各閾値 n における曖昧か否かのラベル数と, 曖昧さ検出の精度を表 5 に示す. n=6, 7 においては 7 割程度の精度を達成しており, 語義曖昧性に基づく読みづらさの検出がある程度できた. n=8 において精度が低下したのは, 閾値が厳しいために, 分散しておらず, 本来曖昧でない回答についても曖昧であるというラベル付けとなったことが原因と考えられる.

6 終わりに

本研究では, JFN を使用した JWiC の構築及びデータ拡張を行い, これを応用して語義曖昧性に着目した読みづらさの検出を行った. 構造的曖昧性に続き語義曖昧性も絶対評価は難しく, 本研究では WiC タスクを語義の数だけ行う形で評価を行った.

収集した例文に対して語義の一致という 2 値分類のタスクに落とし込むことで, 語義ラベルを付与する必要性を排除した. これにより比較的容易にデータ拡張が可能となる. 評価においては語義ラベル自体は必要ないものの, 語義数分の例文収集が必要となる. 本研究では JWiC に含まれる対象語句を活用した評価を行ったが, 今後, 対象語句を拡張した評価を行うことが望まれる.

謝辞

日本語フレームネットを提供いただいた慶應義塾大学の小原京子教授に感謝する. 本研究は JSPS 科研費 JP21H04901 の助成を受けて実施した.

参考文献

[1] Mohammad Taher Pilehvar and Jose Camacho-Collados. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations, 2018.

[2] 吉田あいり, 河原大輔. 構造的曖昧性に基づく読みづらさの検出. 言語処理学会 第 28 回年次大会 発表論文集, pp. 425–429, 2022.

[3] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2019.

[4] Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. Xl-wic: A multilingual benchmark for evaluating semantic contextualization, 2020.

[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 (2). 日本認知心理学会発表論文集, Vol. 2011, pp. 74–74, 2011.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Kyoko Ohara. Relating frames and constructions in Japanese FrameNet. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 2474–2477, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1103.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] 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.