自動リーダビリティ推定の研究動向と展望

NLP2023 深層学習時代の計算言語学

東京学芸大学

江原遥

i@yoehara.com

readability.jp　yoehara.com

一言自己紹介:
英語の語彙学習支援・
読解支援の研究をしてきました

概要

リサーチクエスチョン:

人手評価による読みやすさ(リーダビリティ)は、大規模言語モデルから出力されるパープレキシティのような言語の流暢性の指標とどの程度相関するのか？

- 大規模母語話者コーパスの言語モデル(BERT, bert-base-uncased)だけから、結構、人手評価のリーダビリティと相関する指標RSRSを作れる [Supervised and Unsupervised Neural Approaches to Text Readability, Martinc他, Computational Linguistics 2021]
- ツッコミ(相関は順位相関で見たほうがいい & 単語テスト結果のデータから、平均的な英語学習者がテキスト中の全単語を知っている確率を使った方が、より強く相関する指標が作れる) [Ehara,Eval4NLP workshop21]

展望:最近、江原が新しく気が付いたこと:

実は有名なリーダビリティ指標Flesch-Kincaid Grade Level (1975)も、ある言語モデルのパープレキシティ(の線形和)とみなせる

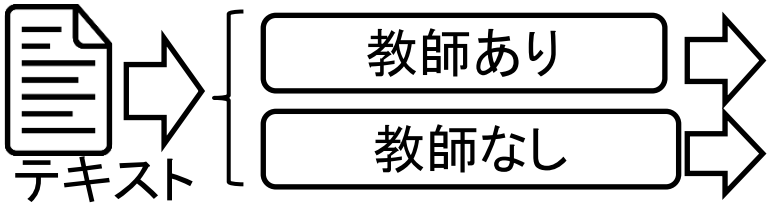
自動リーダビリティ推定
Automatic Readability Assessment (ARA)

所与のテキストのリーダビリティ（可読性）を自動的に推定する事.
[Martinc et al., CL2021]

入力: テキスト, 出力 : 可読性ラベル・スコア

可読性コーパス（人が読んでテキストに可読性ラベルをアノテート）
[Vajjala&Lucic,18]等

- 教師あり (可読性コーパスの一部を訓練データに使う)
- 教師なし (可読性コーパスを訓練データに使わない（それ以外のデータで教師あり学習したりした結果を使ってもよい））



出力はラベル
性能評価は精度

出力は連続
どのように評価するか？

有名なFlesch-Kincaid Grade Level(1975)も、
この分類では「教師なし」
可読性ラベルを使っているわけではないため。

教師なし自動リーダビリティ推定

- 自動リーダビリティ推定
  - テキストの難しさの自動推定
  - 人が難しさを評価したコーパス Weebit [Vajjala and Meurers他12], Newsela[Xu他15], OneStopEnglish [Vajjala&Lucic,18] 初級，中級，上級
  - [Martinc et al., Computational Linguistics 2021]
    - bert-base-uncased＋パープレキシティのような独自の難しさの尺度RSRSで出来るよ
  - 相関係数より順序相関係数 & 英語学習者がテキスト中の全単語を知ってる確率[Ehara,Eval4NLP2021]
  - Trends, Limitations and Open Challenges in Automatic Readability Assessment Research, S. Vajjala, LREC22
    - *Martinc et al. (2021) and Ehara (2021) proposed unsupervised approaches to measuring text readability in the recent past.*
  - A Neural Pairwise Ranking Model for Readability Assessment, J. Lee & S. Vajjala, ACL Findings 22他、最近の論文も全て「教師あり」

Ranked Sentence Readability Score (RSRS) [Martinc他,2021]

NLL: Negative Log Likelihood
（言語モデルの対数尤度）

$$\mathrm{NLL} = -\sum_{i=1}^{n} \log P(w_i | w_{1:i-1}, w_{i+1:n})$$

$$\mathrm{PPL} = e^{(\frac{\mathrm{NLL}}{N})}$$ パープレキシティ

It rains ____ and dogs.

	予測値	実測値
elephants	0.6	0
cats	0.3	1
snakes	0.1	0

パープレキシティ(PPL)を計算する時には、log 0.3しか使わない

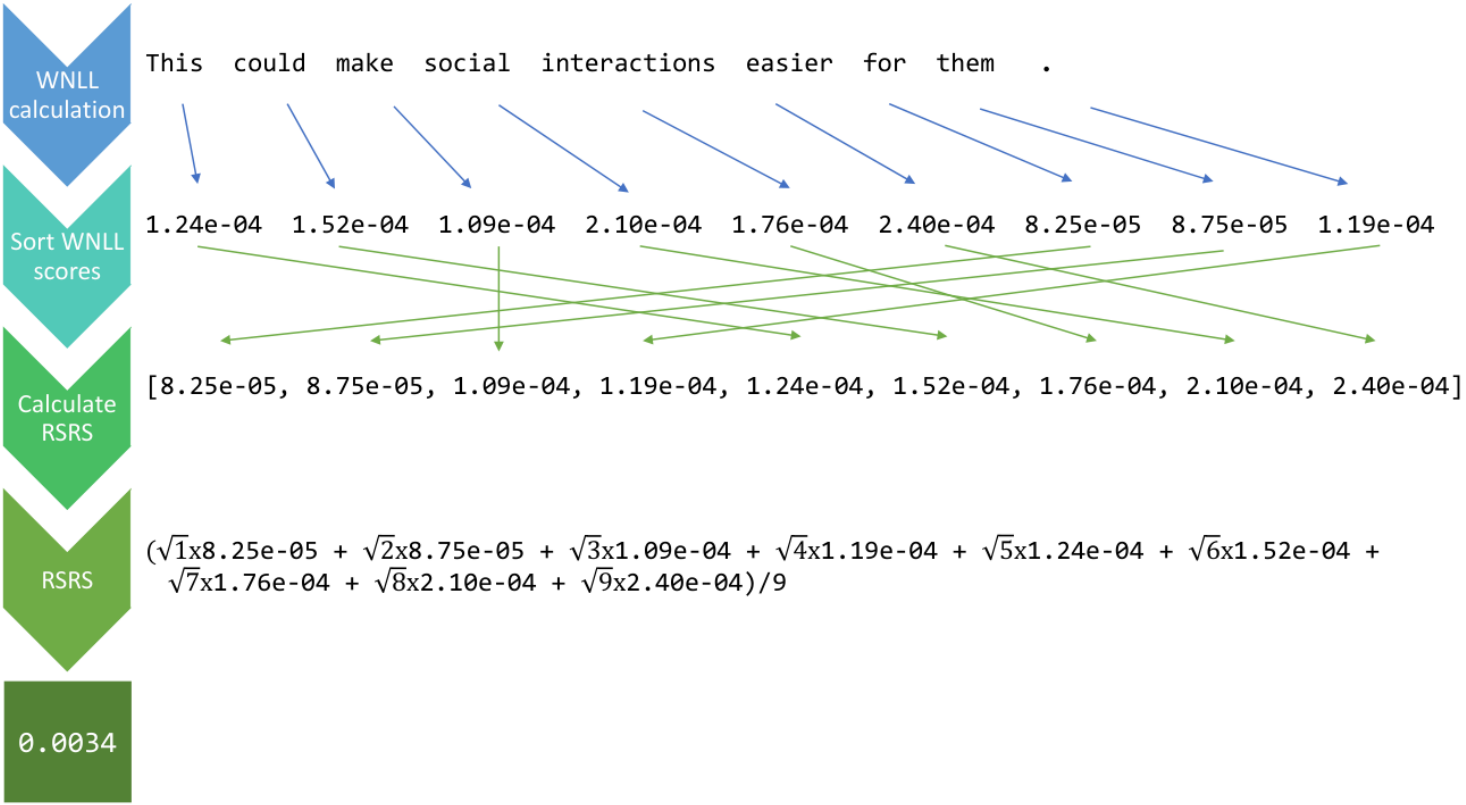
WNLL: 予測が間違っていた場合にlog(1.0-0.6)を加える

$$\mathrm{WNLL} = -(y_t \log y_p + (1 - y_t) \log (1 - y_p))$$

WNLLを文中の全ての単語について計算して、文中のWNLLスコアの高いものを重み付けよう。
√（WNLLスコアの低い（珍しい）ものから数えたときの順位

$$\mathrm{RSRS} = \frac{\sum_{i=1}^{S} \sqrt{i} \cdot \mathrm{WNLL}(i)}{S}$$

RSRS（[Martinc他,2021]より引用）



[Martinc他,2021]より引用

Measure/Data set	WeeBit	OneStopEnglish	Newsela	Slovenian SB
RLM perplexity-balanced	−0.082	0.405	0.512	0.303
RLM perplexity-simple	−0.115	0.420	0.470	/
RLM perplexity-normal	−0.127	0.283	0.341	/
TCN perplexity-balanced	0.034	0.476	0.537	0.173
TCN perplexity-simple	0.025	0.518	0.566	/
TCN perplexity-normal	−0.015	0.303	0.250	/
BERT perplexity	−0.123	−0.162	−0.673	−0.563
RLM RSRS-balanced	0.497	0.551	0.890	0.732
RLM RSRS-simple	0.506	0.569	0.893	/
RLM RSRS-normal	0.490	0.536	0.886	/
TCN RSRS-balanced	0.393	0.601	0.894	**0.789**
TCN RSRS-simple	0.385	**0.615**	0.894	/
TCN RSRS-normal	0.348	0.582	0.886	/
BERT RSRS	0.279	0.384	0.674	0.126
GFI	**0.544**	0.550	0.849	0.730
FRE	−0.433	−0.485	−0.775	−0.614
FKGL	**0.544**	0.533	0.865	0.697
ARI	0.488	0.520	0.875	0.658
DCRF	0.420	0.496	0.735	0.686
SMOG	0.456	0.498	0.813	0.770
ASL	0.508	0.498	**0.906**	0.683

評価指標

人手評価: OneStopEnglish [Vajjala and Lucic, 2018]

離散　例えば… 0, 1, 2 : 初級，中級，上級.

例 : [2,　　1,　　0,　　0,　　1]

教師なしのリーダビリティ評価尺度 : リーダビリティを表す実数値の列

例 : [45.3, 39.2, 10.7, 13.2, 24.4]

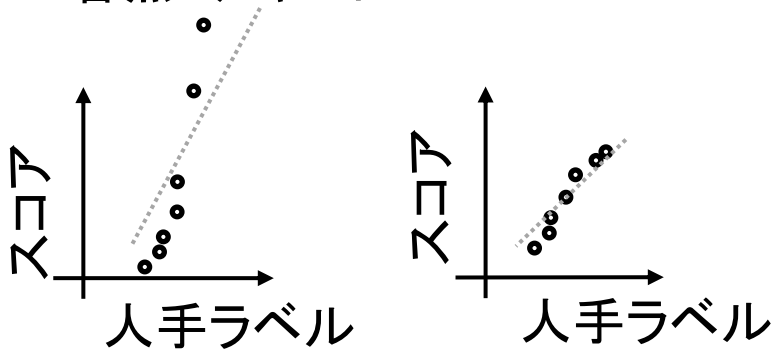
これを普通の相関係数（Pearson’s ρ)で測ると…

相関係数：Pearson's ρ

もし大小関係が完全に一致していたとしても
各点：テキスト

$$\rho_{y,s} = \frac{cov(y,s)}{\sigma_y \sigma_s}$$

cov(y,s): y（人手ラベル）と
s(スコア)の共分散.
σ_y: yの標準偏差
σ_s: sの標準偏差



相関係数は、要するに回帰直線との
ずれが重要なので、人手ラベルと
順序が完全に一致していても、
線形に相関している度合いに
良さが左右される

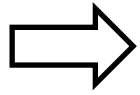
順序相関 – Spearman's ρ

人手ラベル:

値: [2, 1, 0, 0, 1]

順位: [5, 3, 1, 1, 3]

midrank: [5, 3.5, 1.5, 1.5, 3.5]

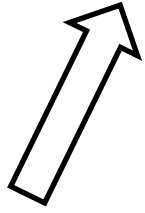


midrank法で同順を考慮した順位の間の相関係数 Pearson's ρ がSpearman's ρ

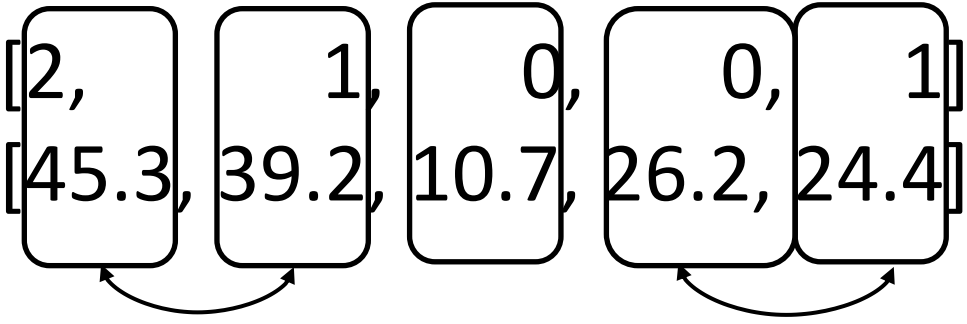
(教師なしの)予測値:

値: [45.3, 39.2, 10.7, 13.2, 24.4]

midrank: [5 , 4 , 1, 2, 3]



Kendall's τ : 順序があっているペアの比率 – あっていないペアの比率



2 > 1

45.3 > 39.2

○順序

0 < 1

26.2 > 24.4

×順序

$$\tau = \frac{n_c - n_d}{\text{Num. of all pairs}}$$

同順を考慮しない

Scipyのデフォルト

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

$$\tau_c = \frac{2(n_c - n_d)}{N^2 \frac{m-1}{m}}$$

離散値と連続値を比較するときはこちらの方がいい説

exp(45.3,39.2,10.7]) = [$e^{45.3}$,$e^{39.2}$,$e^{10.7}$]

結果 1/2

出力されたスコアにexpをかけて、わざと線形性を壊すことで、どの程度スコアが下がるか？

Supervision	Method	Spearman's ρ	Kendall's τ-b	Kendall's τ-c	Pearson's ρ
Unsupervised	**Flesch-Kincaid**	0.324	0.253	0.308	0.359
	exp(**Flesch-Kincaid**)	0.324	0.253	0.308	0.149
	ARI	0.317	0.248	0.302	0.351
	exp(**ARI**)	0.317	0.248	0.302	0.136
	Coleman-Liau	0.373	0.295	0.359	0.372
	exp(**Coleman-Liau**)	0.373	0.295	0.359	0.185
	FleschReadingEase	-0.387	-0.301	-0.366	-0.426
	exp(**FleschReadingEase**)	-0.387	-0.301	-0.366	-0.169
	GunningFogIndex	0.331	0.257	0.313	0.362
	exp(**GunningFogIndex**)	0.331	0.257	0.313	0.151
	LIX	0.348	0.273	0.332	0.383
	exp(**LIX**)	0.348	0.273	0.332	0.129
	SMOGIndex	0.456	0.360	0.438	0.479
	exp(**SMOGIndex**)	0.456	0.360	0.438	0.306
	RIX	0.437	0.340	0.414	0.462
	exp(**RIX**)	0.437	0.340	0.414	0.181
	DaleChallIndex	0.495	0.387	0.472	0.506
	exp(**DaleChallIndex**)	0.495	0.387	0.472	0.431

結果 2/2

	TCN RSRS-simple	-	-	-	0.615(*)
	BERTLMavg	-0.220	-0.173	-0.210	-0.040
	exp(**BERTLMavg**)	-0.220	-0.173	-0.210	-0.005
	BNC	-0.012	-0.009	-0.010	-0.006
	exp(**BNC**)	-0.012	-0.009	-0.010	-0.123
	COCA	-0.018	-0.016	-0.020	-0.039
	exp(**COCA**)	-0.018	-0.016	-0.020	-0.121
	Proposed	**0.730**	**0.592**	**0.709**	**0.715**
	exp(**Proposed**)	**0.730**	**0.592**	**0.709**	0.260
Supervised	**spvBERT_half**	0.751	0.729	0.725	0.747
	spvBERT	**0.866**	**0.856**	**0.854**	**0.864**

[Martinc他,CL21

- 母語話者コーパスを一切使わず、単語テスト結果データから計算
- 平均的な英語学習者がテキスト中の全単語を知っている確率

最近、江原が気が付いたこと:

実はFlesch-Kincaid Grade Level (1975)も言語モデルのパープレキシティとみなせる

パープレキシティの考え方: 1/p(正解単語|モデル)

モデルが選択肢をどれだけ狭められているか？

Flesch Kincaid Grade Level (FKGL, 1975) 有名なリーダビリティ指標

FKGL=0.39×文中の平均単語数 + 11.8 ×単語中の平均音節数 – 15.59

文中の平均単語数(平均文長)=単語数/文数=単語数/EOSの数
=1/(EOSの数/単語数)=1/p(EOS|unigramモデル)

unigramモデル、EOS: 2/6
その他: 4/6

p(EOS|unigram)=2/6
テキスト末尾の
1単語でテキストの
パープレキシティを
測ったもの

平均文長: 6+1/2+1, 点線部のパープレキシティ: 6/2

平均音節数も同様に、音節列中のEnd Of Syllableで考えられる

まとめ

- 教師なしリーダビリティ推定は順位相関を使おう！[Ehara,2021]
- Flesch Kincaid [1975]は言語モデルのパープレキシティだ！ [Ehara,2023研究中]

要するに、最初から皆、言語モデルの話だった。

https://researchmap.jp/yo_ehara

今の話が面白いと思った方は本発表を引用 &

ぜひ江原まで、共同研究などのご連絡をお気軽にいただければ幸いです。

この研究はJST ACT-X (JPMJAX2006)の支援を受けました.

参考文献

[Ehara,Eval4NLP2021] Evaluation of Unsupervised Automatic Readability Assessors Using Rank Correlations, Proc. of Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP, EMNLP2021 workshop)

[Kincaid, 1975] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

[Vajjala&Lučić, 2018] Sowmya Vajjala, Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In Proc. of BEA, pages 297–304

[Martinc他,2021] Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. Computational Linguistics, 47(1):141–179

[Vajjala, 2022] Sowmya Vajjala. Trends, limitations and open challenges in automatic readability assessment research. In Proceedings of the 13th Language Resources and Evaluation Conference (LREC), 2022.

[Vajjala&Meurers,2012] S. Vajjala and D. Meurers. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. BEA 2012.

[Xu,2015] Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. TACL, 3:283–297.