テキストの読みやすさについて

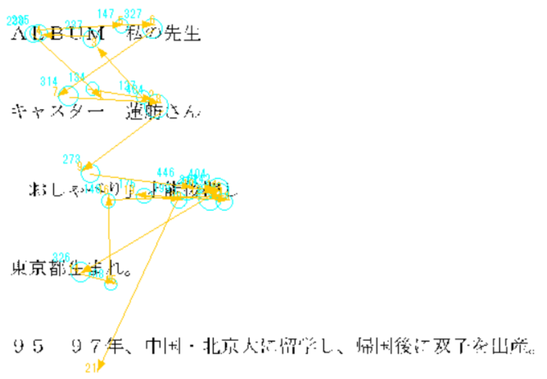
C1

浅原正幸　コーパス開発センター

『現代日本語書き言葉均衡コーパス』に対する読み時間の付与データ BCCWJ-EyeTrack
これに対して、様々なアノテーションと重ね合わせることにより、
テキストの読みやすさに関わる要因を統計的に分析する

BCCWJ-EyeTrack: [Asahara 2016: COLING-2016]

『現代日本語書き言葉均衡コーパス』の新聞記事
に対する日本語母語話者24人分の読み時間データ
- 視線走査法によるデータ
  視線走査順をテキスト正規順にしたデータを公開



- 文節境界の空白
文節境界に半角空白を入れたほうが読み時間が短くなる（レジビリティ）

- 実験の進捗
実験が進むと読み時間が短くなる（慣れてくる）

単語埋め込みによるモデル化 [浅原 2018: 言語学会]

サプライザル理論 [Hale 2001]
頻度は文処理のパフォーマンスに影響を与える

$$\mathrm{Effort}(w_i) \propto \log \frac{1}{P(w_i | w_{1\ldots i-1,}, \mathrm{Context})}$$

日本語の場合、読み時間を評価する文節単位と
頻度を計数する単語単位に齟齬がある

<u>単語埋め込みの加法構成性の利用</u>
　単語単位のベクトルの総和を文節のベクトルとして
利用
・当該文節のノルムのベクトル
→ノルムが大きいほど読み時間が長くなる
ノルムが大きいほど様々な単語と結びつきやすい
・左隣接文節ベクトルと当該文節ベクトルのコサイン
類似度
→類似度が大きいほど読み時間が短くなる
隣接確率に相当

アノテーションとの重ね合わせによる分析

○係り受けアノテーション BCCWJ-DepPara [Asahara and Matsumoto 2016: ALR12] との重ね合わせ
　係り受けの数が多くなるほど読み時間が短くなる
　係り元の文節が係り先の文節の予測に寄与する

補足語修飾節【関係節 ウチの関係】:
「被修飾名詞が修飾節内述部と格関係にあるもの」
　SELF で読み時間が短くなる

　(1) 幼稚園から大学まで<u>通った</u>青山学院では、
　【MSa200:名詞修飾節:補足語修飾節:非限定的】
　(読売新聞2001 年[BCCWJ: 00001_A_PN1c_00001_A_1])

内容節【関係節 ソトの関係】:
「被修飾名詞が発言・思考・事柄に関する意味を持ち、
　被修飾名詞と修飾節が同格にあるもの」

　SELF で読み時間が短くならない
　SPT で読み時間が長くなる（二度見が多い）

　(2) 支払利息や減価償却費の計上額が<u>少ない</u>傾向がある。
　【MSb:名詞修飾節:内容節】
　(北海道新聞2002 年[BCCWJ: 00005 A PN2e 00001 A 2])

○被験者属性　[浅原+ 2017: 言語処理学会年次大会]
　記憶力がある群は読む速度が速いが、全読み時間は変わらない
　語彙力がある群は読み時間が長い

○節境界との重ね合わせ[Asahara 2018: PACLIC-32, 浅原 2017: 言語学会]
　節末で読み時間が短くなる
　関係節ウチの関係は関係節ソトの関係より読み時間が短くなる

○分類語彙表番号との重ね合わせ [Asahara & Kato 2017: IJCNLP-2017, 浅原・加藤 2017:認知科学会]
　統語分類：用の類＜相の類＜体の類
　意味分類：「関係」＜「主体」≒「活動」≒「生産物」≒「自然物」

○情報構造との重ね合わせ [Asahara 2017: PACLIC-31, 浅原 2017: 言語処理学会]
　共有性：　旧情報＜ブリッジング＜新情報
　定性：　不定名詞句＜定名詞句
　有生性：　有生名詞句＜無生名詞句

○述語項構造との重ね合わせ [Asahara 2018: READ-2018]
　主語が明示されていないが、読み手が主語の場合（外界二人称照応主語）を持つ述語の
二度見が少なくなる（SPT が短くなる）