

日本語語順分析に言語モデルを用いることの妥当性について

栗林樹生^{1,2}, 伊藤拓海^{1,2}, 鈴木潤^{1,3}, 乾健太郎^{1,3}

1 東北大学 2 Langsmith株式会社 3 理研AIP

日本語語順分析に言語モデルを用いることの妥当性について

- 日本語の語順は比較的自由

- 理沙が 健に バトンを 渡した =  = Risa gives Ken the baton
- 健に 理沙が バトンを 渡した =  ≠ Ken gives Risa the baton
- バトンを 渡した 理沙が 健に =  ≠ *The baton gives Risa Ken

日本語語順分析に言語モデルを用いることの妥当性について

- 日本語の語順は比較的自由

- 理沙が 健に バトンを 渡した =  = Risa gives Ken the baton
- 健に 理沙が バトンを 渡した =  ≠ Ken gives Risa the baton
- バトンを 渡した 理沙が 健に =  ≠ *The baton gives Risa Ken

- 語順が完全に自由というわけではない (基本語順規則の存在)

Mazukaら2002, Tamaokaら2005

- 品質に 影響を 与えた
- ?影響を 品質に 与えた

日本語語順分析に言語モデルを用いることの妥当性について

● 日本語の語順は比較的自由

- 理沙が 健に バトンを 渡した =  = Risa gives Ken the baton
- 健に 理沙が バトンを 渡した =  ≠ Ken gives Risa the baton
- バトンを 渡した 理沙が 健に =  ≠ *The baton gives Risa Ken

● 語順が完全に自由というわけではない (基本語順規則の存在)

Mazukaら2002, Tamaokaら2005

- 品質に 影響を 与えた
- ?影響を 品質に 与えた

日本語学習者も混乱



● 基本語順規則に諸説あり

- 二重目的語構文はすべて「にを語順」であると考えられる (Hoji1985, Takano1998 など)
 - すべて「にを語順」であるとは考えにくい (Sasanoら2016など)

日本語語順分析に言語モデルを用いることの妥当性について

検証のプロセス

1. 仮説の提唱

例: 二格名詞が生物であれば, 「□を<生物>に△した」ではなく
「<生物>に□を△した」という語順になる

2. 検証

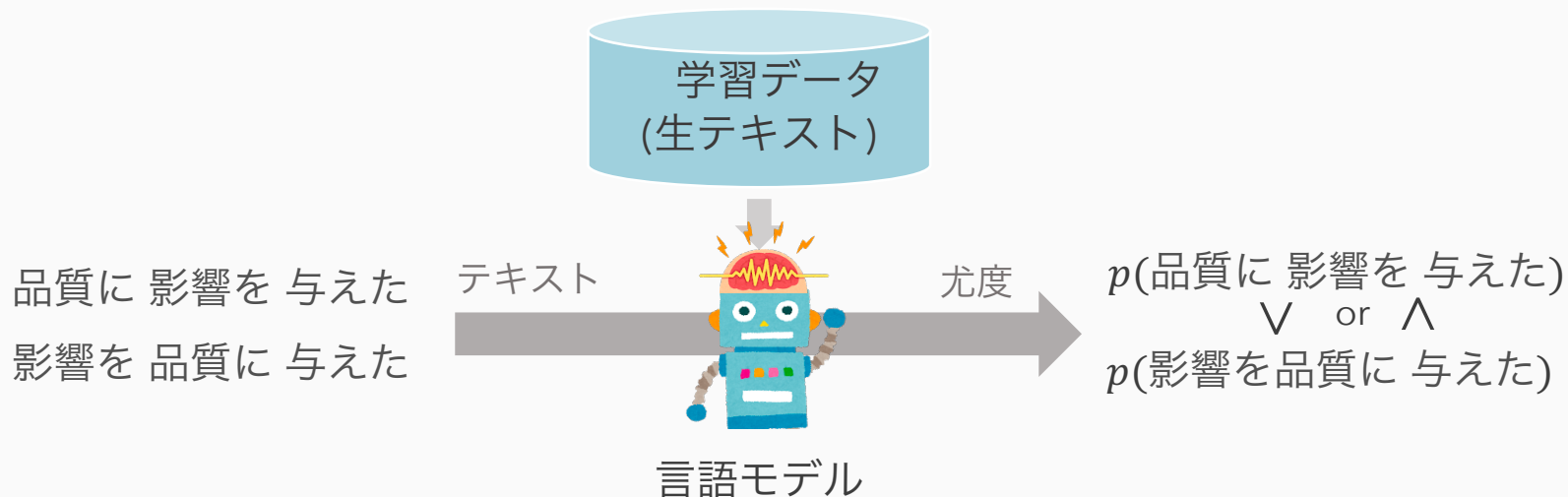
仮説に従う語順の方が, 人間の処理負荷が低い (心理実験・脳科学など)
たくさん使われている (コーパス言語学)

3. 結論

仮説は正しい

日本語語順分析に言語モデルを用いることの妥当性について

- 言語モデルは、学習データの観測に基づき、テキストの尤もらしさを求めることが可能



日本語語順分析に言語モデルを用いることの妥当性について

1. 仮説の生成

2. 検証

仮説に従う語順の方が、人間の処理負荷が低い (心理実験など)
たくさん使われている (コーパス言語学)
大規模テキストで学習した言語モデルの尤度が高い

3. 結論

仮説は正しい

日本語語順分析に言語モデルを用いることの妥当性について

本研究の問い:

ある仮説に従った語順の言語モデル尤度が高いことを,
その仮説を支持する証拠として用いて良いか

仮説に従う語順の方が, 人間の処理負荷が低い (心理実験など)
たくさん使われている (コーパス言語学)
大規模テキストで学習した言語モデルの尤度が高い

3. 結論

仮説は正しい

なぜ言語モデルを語順分析に用いたいか

既存の検証方法と比較しながら

人間の観察に基づく検証
Shigenaga2014など



高コスト

- 偏りのない被験者を集める
- 心理実験などの実験内容・設定の統制

なぜ言語モデルを語順分析に用いたいか

既存の検証方法と比較しながら

人間の観察に基づく検証

Shigenaga2014など



高コスト

- 偏りのない被験者を集める
- 心理実験などの実験内容・設定の統制

数え上げに基づく検証

Sasanoら2016

仮説に従う語順



比較的低コスト

- 被験者の収集や心理実験を行う必要がない
- 低コスト -> 大規模・網羅的な検証が可能

なぜ言語モデルを語順分析に用いたいか

既存の検証方法と比較しながら

人間の観察に基づく検証

Shigenaga2014など

時に、仮説に従う語順と従わない語順を分類して数えることは容易でない

詳細は3.2節

- 心理実験などの実験内容・設定の統制

数え上げに基づく検証

Sasanoら2016

仮説に従う語順



比較的低コスト

- 被験者の収集や心理実験を行う必要がない
- 低コスト -> 大規模・網羅的な検証が可能

なぜ言語モデルを語順分析に用いたいか

既存の検証方法と比較しながら

時に、仮説に従う語順と従わない語順を分類して数えるよりも、仮説検証用データを作成し、言語モデルに入力する方が現実的

詳細は3.2節

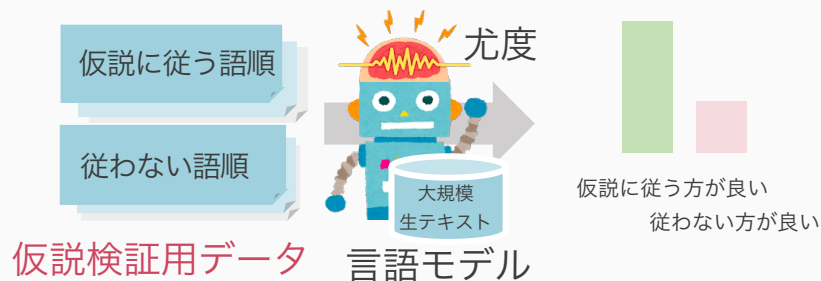
数え上げに基づく検証

Sasanoら2016 仮説に従う語順



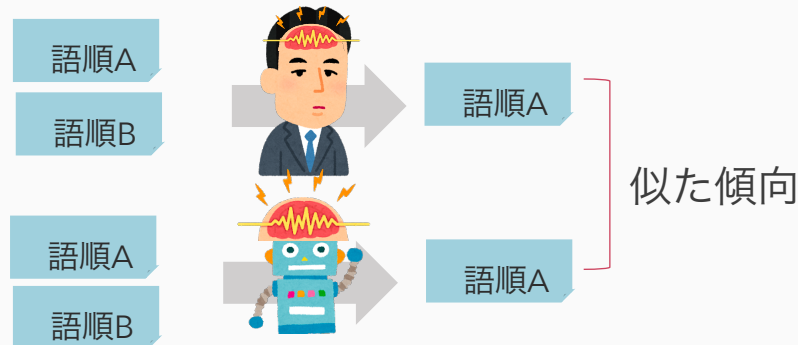
言語モデル尤度の比較に基づく検証

近年は言語モデルの実装コストも低い

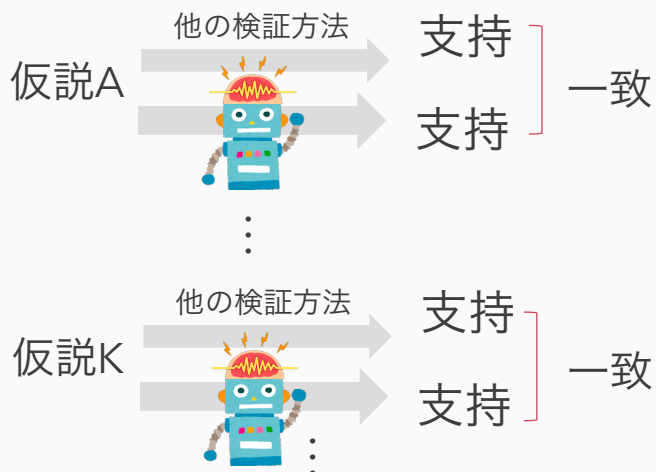


実験: 言語モデルを語順検証ツールとして使って良い?

1. 人間の語順選好との相関の高さ

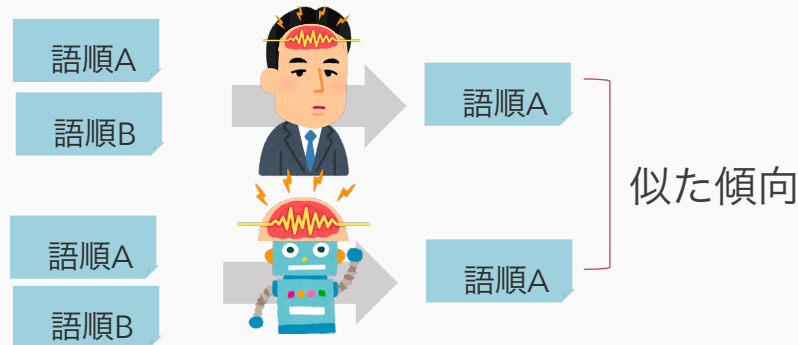


2. 様々な仮説において、既存研究と同様の結論が導かれる

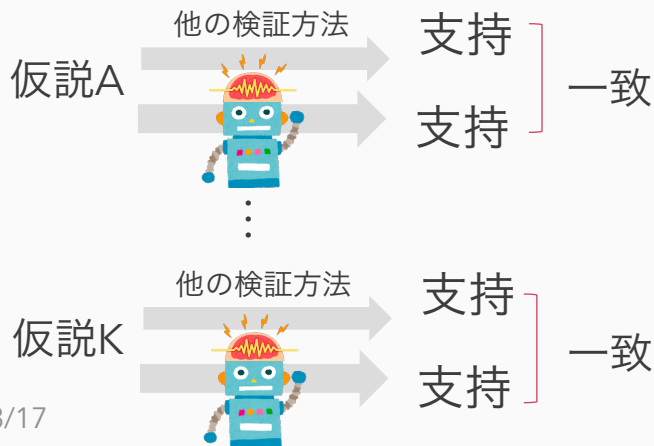


実験: 言語モデルを語順検証ツールとして使って良い?

1. 人間の語順選好との相関の高さ



2. 様々な仮説において、既存研究と同様の結論が導かれる

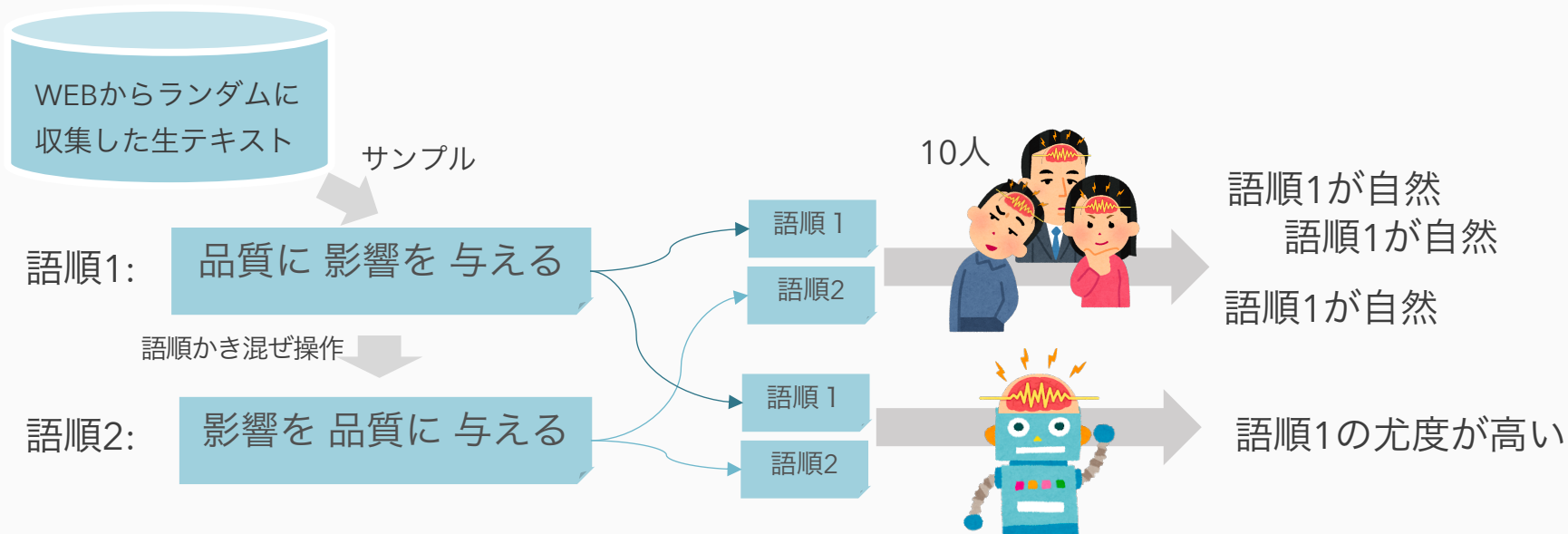


実験設定

- (次の単語を予測する) 言語モデル
- 学習データ: WEBからランダムに収集した1.6億文 (検証用データとの重複はない)
- 言語単位: 文字レベルとサブワードレベルのLM両方で実験
- Left-to-right LMによる尤度と, Right-to-left LMによる尤度の和を使用

実験1: 人間の選好との相関

- Yahoo!クラウドソーシングを利用 (計756人が参加)



以下の条件を満たす1万文ペアを使用

- どちらの語順も意味的に崩壊していない(ワーカーによる判断)
- ワーカー10人中9人以上の選好が一致

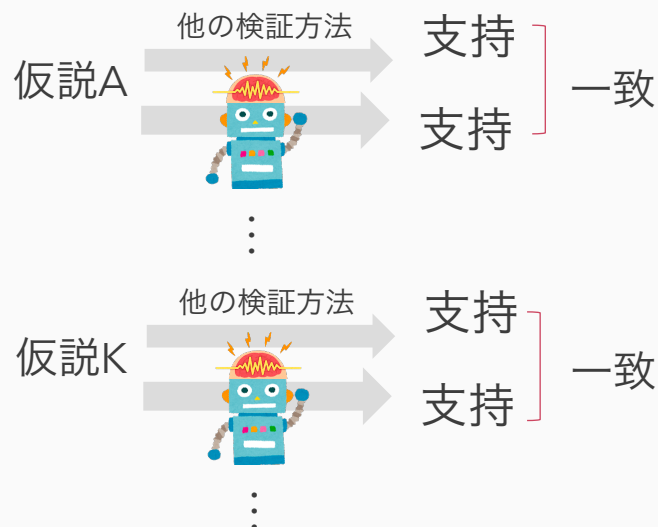
実験1: 人間の選好との相関

- 結果: 相関が非常に高い
 - 人間と文字レベル言語モデルの選好の相関: 0.89
 - 人間とサブワードレベル言語モデルの選好の相関: 0.90

基本語順分析に言語モデルを用いることの妥当性を支持

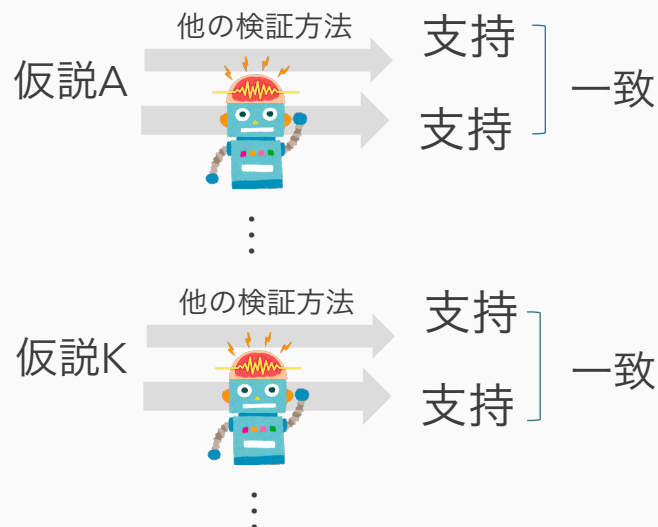
実験2: 検証道具としての振る舞いの一致

- 様々な仮説において、既存研究と同様の結論が導かれるか



実験2: 検証道具としての振る舞いの一致

- 様々な仮説において、既存研究と同様の結論が導かれるか



検証した10の仮説で全て結論が一致

基本語順分析に言語モデルを用いることの妥当性を**支持**

詳細は、栗林ら2019 (https://www.anlp.jp/proceedings/annual_meeting/2019/pdf_dir/P5-5.pdf)
の3節と栗林ら2020の5節を読む or 質問してください

実験2:検証に用いた仮説

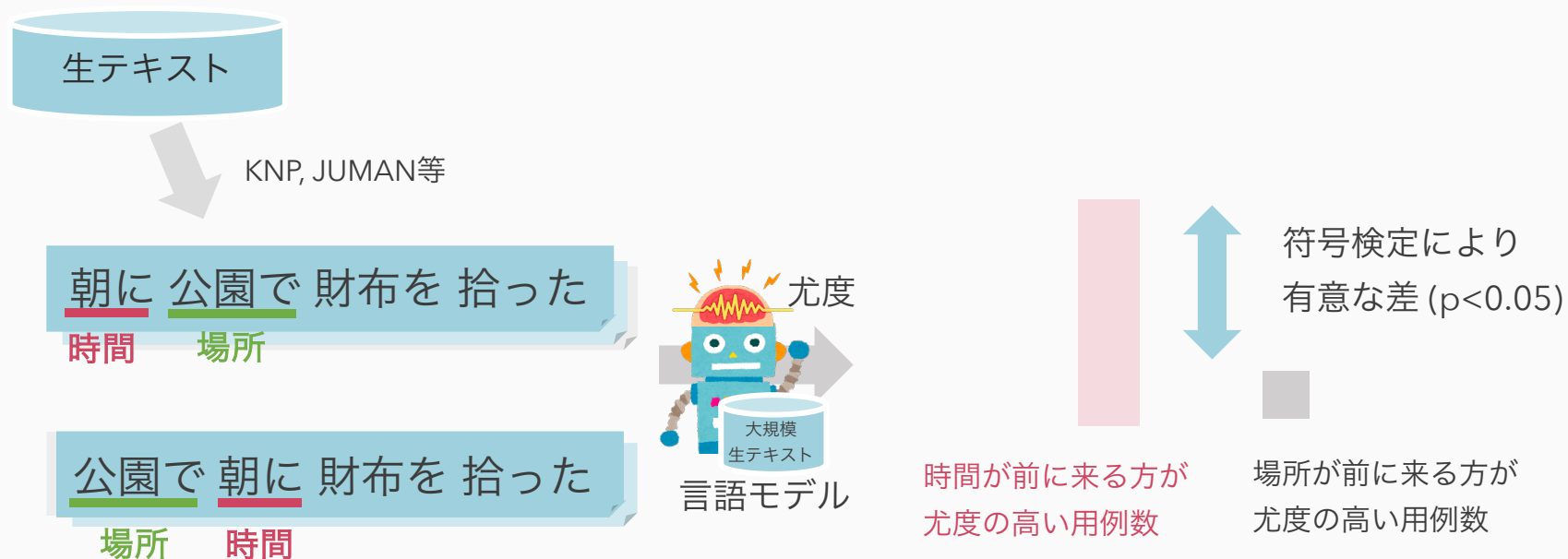
対象	仮説	既存の 検証結果	言語 モデル
二重目的語	動詞によらず基本語順は「にを」である	棄却	棄却
	基本語順は動詞のタイプによって異なる	棄却	棄却
	省略されにくい格は基本語順において動詞の近くに位置する	支持	支持
	基本語順は二格名詞の意味役割や有生性によって異なる	支持	支持
	対象の動詞と高頻度に共起するヲ格，二格名詞は動詞の近くに位置する	支持	支持
副詞	副詞のタイプにより基本生起位置は異なる	支持	支持
主語	主語は時間・場所を表す格よりも後ろにくる	支持	支持
場所	場所を表す格は時を表す格より後ろ，主語よりも前に位置する	支持	支持
時	時間を表す格は場所，主語よりも前に位置する	支持	支持
(一般)	モーラ数の多い句が少ない句よりも前に来る	支持	支持

栗林ら2019の
実験設定を
アップデート

詳細は，栗林ら2019 (https://www.anlp.jp/proceedings/annual_meeting/2019/pdf_dir/P5-5.pdf)
の3節と栗林ら2020の5節を読む or 質問してください

一致

検証の例: 時間を表す格は、場所を表す格よりも前にくる



分析: 言語モデルの長所を生かした検証

- とりたてに着目し、格助詞の明示的に出現しない格が存在する文の語順を分析

とりたて助詞「は」による 彼に 猫を 渡した
とりたて 猫は 彼に 渡した

- とりたて助詞ごとの語順に対する影響の分析
- 「基本語順で前方にある格ほど主題になりやすい」という仮説
(佐伯1960, 1996)や、二重目的語の主題になりやすさを検証し、経験的な証拠を提示した

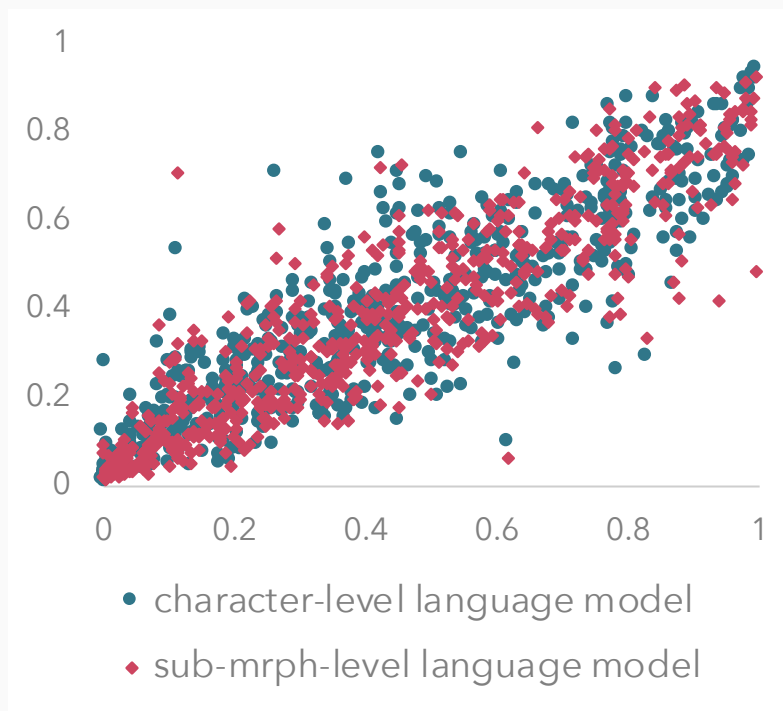
詳細は6節

まとめ

- 語順分析に言語モデルを用いることの利点・妥当性について議論
- 人間の語順選好との高い相関, 既存の検証方法による検証結果との一致から, 言語モデルを用いることの妥当性を支持
- 取り立てと主題について, 言語モデルによる示唆を提供
- 本論文では, 一旦妥当性が支持されたと結論づけるが, さらなる検証が必要

Appendix

実験2: 二重目的語構文



実験2: 副詞の位置

友達が 道具を 乱暴に 扱った
主語 目的語 副詞

友達が 乱暴に 道具を 扱った

乱暴に 友達が 道具を 扱った



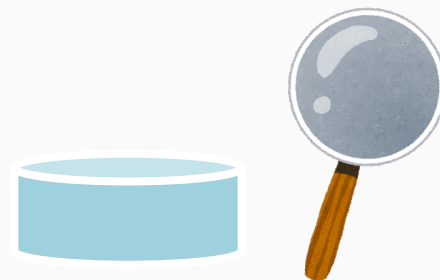
副詞の種類と傾向

- 陳述 (あいにくなど): 主語の前
- 時 (昨日など): 主語の前 (主語と目的語の間)
- 様態 (ゆっくりなど): 主語と目的語の間, 目的語の後ろ
- 結果 (こなごなになど): 目的語の後ろ (主語と目的語の間)

実験2: 句のモーラ数と語順

昨日 彼に レシートがたくさん入った財布を 渡した

レシートがたくさん入った財布を 昨日 彼に 渡した



言語モデルが基本語順から
逸脱した語順を支持したデータ

最もモーラ数の大きい句が前に動いていることが、
後ろに動いていることよりも有意に多かった

分析: とりたて助詞ごとの語順に対する影響

● とりたて助詞別の語順に対する影響

とりたて助詞「は」による
とりたて

格助詞が提示されないため、
解析器のミスを引きかねない

彼に 猫を 渡した
猫は 彼に 渡した

「も」によるとりたて

彼に 猫を 渡した
猫も 彼に 渡した

「こそ」によるとりたて

彼に 猫を 渡した
猫こそ 彼に 渡した

「だけ」によるとりたて

彼に 猫を 渡した
猫だけ 彼に 渡した



猫は 彼に 渡した

とりたてによって文頭に移動する現象は、
主題(対比)のとりたて助詞「は」特有

基本語順で文頭側にある格要素ほど、
取り立てられた際に前に動きやすい傾向

分析: 格の主題になりやすさ

- 「基本語順で前方にある格ほど主題になりやすい」

火曜日には 学校で 私が 友達に 本を 渡した
学校では 火曜日に 私が 友達に 本を 渡した
私は 学校で 火曜日に 友達に 本を 渡した
友達には 学校で 火曜日に 私が 本を 渡した
本は 学校で 火曜日に 私が 友達に 渡した



私は 学校で 火曜日に 友達に 本を 渡した

Kを格の種類とすると, $\{(a,b) \in K \times K \text{ ただし, } a \neq b \text{ かつ 基本語順において } a \text{ が } b \text{ より前にある}\}$ について

$$\frac{N_{a|b}}{N_{a|b} + N_{b|a}} > \frac{N_{b|a}}{N_{a|b} + N_{b|a}}$$

$N_{a|b}$ は, 言語モデルが指示した語順で, bではなくaが取り立てられている用例の数

t検定 (対応あり) より有意な傾向