修士論文

Master's Thesis

論文題目

Thesis Title

主題化における人間と言語モデルの対照

Contrast of Humans and Langage Models about Topicalization

提出者

東北大学 大学院 情報科学研究科 システム情報科学専攻

Department of System Information Sciences

Graduate School of Information Sciences, Tohoku University

学籍番号 (ID No.) C1IM2034

氏名 (Name) 藤原　吏生 (Riki Fujihara)

指導教員	乾　健太郎　教授
学位論文指導教員	
審査委員 (〇印は主査)	〇乾　健太郎　教授 1 塩入　諭　教授　　2 鈴木　潤　教授

提出者略歴	
ふじはら　りき 氏名　藤原　　吏生	平成 10 年 7 月 5 日生
本籍　山梨県	国籍
履歴事項	
[学歴]	
平成 29 年 4 月 1 日	東北大学　工学部　電気情報物理工学科入学
令和 3 年 3 月 25 日	同　卒業
令和 3 年 4 月 1 日	東北大学　情報科学研究科　システム情報科学専攻 博士課程前期 2 年の課程　入学
令和 5 年 3 月 25 日	同　修了

Contrast of Humans and Langage Models about Topicalization*

Riki Fujihara

Abstract

Humans use different wordings depending on the context to facilitate efficient communication. For example, instead of completely new information, information related to the preceding context is typically placed at the sentence-initial position. In this study, we analyze whether neural language models (LMs) can capture such discourse-level preferences in text generation. Specifically, we focus on a particular aspect of discourse, namely the topic-comment structure. To analyze the linguistic knowledge of LMs separately, we chose the Japanese language, a topic-prominent language, for designing probing tasks, and we created human topicalization judgment data by crowdsourcing. Our experimental results suggest that LMs have different generalizations from humans; LMs exhibited less context-dependent behaviors toward topicalization judgment. These results highlight the need for the additional inductive biases to guide LMs to achieve successful discourse-level generalization.

Keywords:

Natural Language Processing, Language Model, Topicalization

*Master's Thesis, System Information Sciences, Graduate School of Information Sciences, Tohoku University, C1IM2034, January 27, 2023.

主題化における人間と言語モデルの対照*

藤原 吏生

内容梗概

人間は，効率的なコミュニケーションを行うために，文脈に応じて言い方を使い分けている．例えば，完全な新情報ではなく，先行する文脈に関連する情報を文頭に配置することが一般的である．本研究では，テキスト生成においてニューラル言語モデルがこのような談話レベルの選好を捉えられるかどうかを分析する．具体的には，談話の一側面である，主題化に着目する．言語モデルの言語的な知識を個別に分析するため，主題優勢言語である日本語を用いてプロービングタスクの設計および人間の主題化における判断を収集した．実験の結果，言語モデルは人間とは異なる汎化能力を持っており，言語モデルは主題化の選択において人間より文脈に依存しない振る舞いをすることが示された．これらの結果は，言語モデルが談話レベルの汎化を成功させるためには，さらなる帰納的バイアスが必要であることを示唆している．

キーワード

自然言語処理，言語モデル，主題化

*東北大学 大学院情報科学研究科 システム情報科学専攻 修士論文, C1IM2034, 2023 年 1 月 27 日.

目 次

図 目 次

表 目 次

1　はじめに

自然言語処理の分野におけるニューラル言語モデルの近年の成功を踏まえ，言語モデルの言語的な知識，主には言語モデルの統語的な知識を検証する研究が行われてきた [1, 2, 3, 4, 5, 6]．談話も統語的な構造とともに文章の生成において必要不可欠な側面であるが，言語モデルにおける談話レベルの言語的な知識はあまり研究されておらず，行われているとしても文の並び替えをタスクとして分析するなど，粗いレベルでの分析が一般的である [7, 8]．

本研究では，言語モデルの談話レベルの言語的知識を精緻に理解するために，談話の一つの側面である主題構造 (thematic structure とも呼ぶ) における言語モデルの汎化性能を調査する．主題構造は，文章の生成において必要不可欠な要素である [9, 10, 11]．

例えば，話し手は以下の文を文脈に応じてで使う．

(1) a. In Japan, my father bought the vase last year.
　　b. The vase was the one my father bought in Japan last year.

これらの文は，主題構造という点で異なる (主題には下線が引かれている)．例えば，”I broke a vase yesterday.”という文の続きとしては，文 (1b) の方が文 (1a) よりも適している．本研究では，言語モデルがこのような文脈に依存した，談話レベルの語用論的な選択を人間のように行えるかどうかを検証する．

言語の認知科学では，言語モデルのような計算モデルの (非) 人間的な言語汎化能力は古くから注目されている [12, 13]．さらに，工学的な観点からも，言語モデルが主題化に対して正しい選好を持つかどうかということは，言語モデルを文章の自動品質評価に用いる際の妥当性を保証する上で重要な点である．

主題構造に関する言語モデルの行動の認知的な妥当性を検証するために，主題構造が明示的に示される主題優勢言語を用いた．主題優勢言語を用いることで，人間と言語モデルの振る舞いを分析するための文のペアを容易に作成することが可能となる．まず，クラウドソーシングを利用して，人間の主題化に対する選好をアノテーションし，分析を行った (3 節 ， 4 節)．

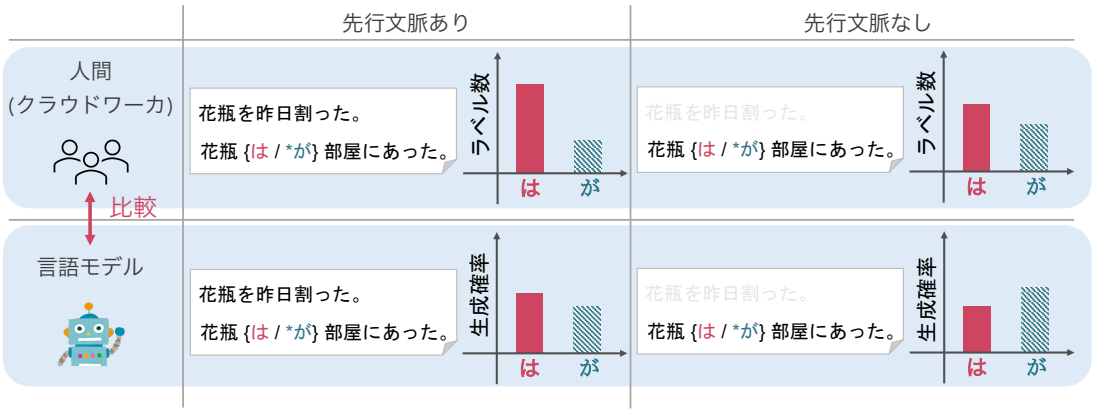

図 1: 人間と言語モデルの主題化に関する選択における振る舞いの比較

実験では，先行文脈に依存した主題化に関する選好を，言語モデルと人間とで比較した (5 節; 図 1)．実験の結果，言語モデルは人間のような文脈依存的な振る舞いをしないことが示された (6 節)．さらに，言語モデルは文脈に依存せず，特定の名詞 (項) を主題化するべきか否かということにだけ基づいて主題化に関する選択を行っていることが示された．この結果から，言語モデルは人間とは異なる振る舞いをしており，人間のような振る舞いを行うためには，学習においてさらなる帰納バイアスが必要であることが明らかになった．

2 関連研究

2.1 文の主題

文の主題は，メッセージの中で関心があるもの，つまり話し手が何を言おうとしているのかを表す．言語学などの知見では，文の主題は一般的に文頭の要素 [10, 14] に該当することが知られている[1]．例えば，例 (1a)(1 節) の主題は *Japan* であり，例 (1b) の主題は *vase* である．

文中の特定の要素をその文の主題として提示することを「主題化」と呼ぶ．主題化は，典型的には例 (1) に示すような語順によって示され，言語によって主題

[1]厳密に言えば，主題にはいくつかの定義があるが，この研究では， [10] で導入した主題テーマに着目している．

を提示する (主題化する) 方法 (助詞やイントネーションなど) が異なることがある (Section 3.1).

文中の非主題の部分をコメントと呼ぶ．主題とコメントは，言語学では ”topic-comment structure” や ”theme-rheme structure” として研究されることが多い [9, 10, 11]．これらは文の文法構造や論理構造とは区別され，例えば，例 (1a) の主題は *Japan* であり，文法上の主語は *my father* となる．なお，文は必ずしも主題を持つとは限らない (例：“There is a pen.” という文は主題を有さない).

2.2 主題化と談話

テキスト生成において，文の主題はメッセージの中での関心事を示し，語順を制御する重要な役割を果たす (例 (1))．一般に，ある文脈の中で特定の概念が顕著であるほど，その概念は主題化されやすいと言われている [10, 15]．この性質は談話における **センタリング理論** [16, 17, 9] と密接な関係があることが知られている．センタリング理論では，ある文の主題は文法的な要素に従って高い Centering-forward (Cf) レベルを持ち，この Cf が高い要素が主題になることが知られている．また，ある文で Cf が高かった要素は，次の発話でも高い Cf 順位を持つことが多いことが期待される (すなわち，主題が連続した発話を越えて続くことが期待される).

本研究では，言語モデルが意味を伝える際に文中のどの要素を主題に (主題化) すべきかを判断する際に，人間に近い選好を持っているかどうかを明らかにする．主題化は談話レベルの文脈に依存した言語現象であることから，主題化に関する言語モデルの振る舞いを調べることで，言語モデルが文レベル，さらには談話レベルでの文章の流れを捉えられているかということについての理解を深めることができると考える．

2.3 言語的知識の分析

自然言語処理の研究では，ブラックボックスであるニューラルモデルの内部構造および (または) ニューラルモデルの動作を分析し，それらが実際に言語を理解

しているかどうかを調査している．このような分析 (プロービング) の対象は，例えば，構文的な知識 [3, 13, 18] から常識的な知識 [19, 20] に至るまで幅広いものとなっている．また，共参照 [21, 22] や談話構造 [23, 24, 25] など談話レベルの現象についても分析を行ったものが存在する．本研究は談話レベルの言語現象である主題化に注目したものであり，我々の知る限り主題化を題材とした研究が存在しないという点で，本研究はニューラル言語モデルについて新たな知見が提供可能であると考える．

言語モデルの言語な知識を調べるために，既存研究ではしばしば，特定の言語的側面においてのみ異なるテキストペア (minimally-different-text-pair) を使用し，言語モデルによって計算される確率 (パープレキシティ) をもとにその知識分析する [3, 13, 26]．この実験のデザインには利点があり，追加の分類器を設計することなく，直接モデルの振る舞いを分析することができる [27, 28]．そこで本研究では，主題構造の観点でのみ異なるテキストペアを用いて言語モデルの分析を行う．

3　データセット作成

3.1　主題優勢言語の使用

英語を用いて，主題構造の観点でのみ異なるテキストペアを作成することは非常に困難である．なぜなら，英語において文中の主題化する要素が変わると，それに伴って文の構造や構文の複雑さも大きく変わってしまうためである (例 1)．そのため，英語で異なる主題構造を持つテキストペアを用いて言語モデルの選好を分析すると，言語モデルが実際にどの言語的な情報をもとに主題化の判断を行ったのかという結論づけを行うのが困難になる可能性がある．一方，主題優勢言語の一つである日本語では，主題構造の観点でのみ異なるテキストペアを作ることが可能である．例えば，主題優勢言語である日本語の次の 2 つの文は，主題構造の観点でのみ異なる (文の主題が異なる)．

(2) a. 花瓶は　部屋に　あった.
Vase-TOP room-DAT exist-PAST.
The vase was in the room.

b. 花瓶が　部屋に　あった.

Vase-NOM room-DAT exist-PAST.
There was a vase in the room.

日本語では，例 2a のように，助詞の「は」が，文の主題を表す「主題マーカー」として使われることが知られている [29, 30, 31, 32] [2]．例 (2) の 花瓶は (*Vase*-TOP) と 花瓶が (*Vase*-NOM) は，ある要素 (花瓶; *vase*) が主題としてマークされているかどうかという点で異なっている．

日本語は膠着語 (agglutinative language) であり，要素の機能情報 (文法的格など) は後置詞 (助詞) によって提示されることに注意されたい．ある特定の要素に「は」が付くと，本来使われる助詞 (例えば「が」) が省略されることがある．つまり，例 (2a) の 花瓶は (*Vase*-TOP) は，文法上の主語 (NOM) と文の主題 (TOP) の両方の役割を果たしているが,「は」しか伴わない．

一方，例 (2b) の 花瓶が (*Vase*-NOM) は主題ではなく単なる文法的主語 (NOM) である．花瓶についての話をする文脈では，文法主語 (*kabin*; *vase*) が主題化されている例 (2b) が選好されるはずである．我々は，言語モデルがこのような文脈に依存した主題マーカーの使用に関する選好を持つかどうかを分析した．このような主題化に関する選択（「は」と「が」の選択）は厳密なルールによって決定されるものではない．むしろ，例 (2a) と (2b) のどちらも通常はどちらも文として許容されるが，文脈によってどちらかの文がより自然になることが生じるというものである．このような，人間の主題化に関する選好度合いをアノテーションしたデータセットを作成した．

なお，日本語のセンタリング理論では，文法的な主題 (TOP) は主語 (NOM) よりも Cf の順位が高いとされており，この文脈に依存した「TOP か NOM のどちらが伴うのが自然か」という判断は，センタリング理論から見ると Cf を推定するタスクとみなすことができる可能性もある．

[2]厳密には,「は」と「が」の対立は主題化だけでなく，日本語学で古くから議論されてきたものである．また，全ての 「は」 が主題マーカーとして機能すると主張する意図はなく，主題マーカーとして使用されていることが考えられるデータポイントを抽出して実験に用いた (3 節).

3.2　アノテーションタスクの作成

本研究では，述語の項のうち，特に主格，与格，対格の主題化を対象とする．以下に，ある項が主題化された文とされていない文のペアを例示する．また，主題化されている項とそれに伴う助詞に下線を引いた．

主格の主題化の例

(3) a. 花瓶が部屋にあった．

b. 花瓶は部屋にあった．

与格の主題化の例

(4) a. 飛行機で沖縄に向かった．

b. 沖縄には飛行機で向かった．

対格の主題化の例

(5) a. 沖縄で会議を開催した．

b. 会議は沖縄で開催した．

実験では，特定の文脈のもと，続く文として項を主題化した/しない文のどちらが自然かを人間と言語モデルに問い，選好を比較する．なお，主格についてはより精緻な分析を行い (6.1)，与格・対格については追加の分析として，語順レベルでの変化も伴う現象として言語モデルの分析を行う (6.2 節)．

3.2.1　データの抽出

我々は主題優勢言語の一つである日本語に着目した．具体的には，日本語の文におけるある項の主題化の選好度合いを人間と言語モデルとで分析した．本研究では，日本語の言語現象の分析によく用いられる NAIST テキストコーパス

文章		人間の選好		言語モデルの選好	
文脈 c	文 s と主題化を問う項 a	$r_{\text{human}}^{\text{C+S+A}}$	$r_{\text{human}}^{\text{S+A}}$	$r_{\text{LM}}^{\text{C+S+A}}$	$r_{\text{LM}}^{\text{S+A}}$
私たちは日々，大量のエネルギーを消費しています.	経済成長-{TOP/NOM} エネルギーの使用と密接な関わりがある.	0.83	1.00	0.66	0.64
ヨットレースはスタートダッシュが勝負を決めると言われる.	強い風-{TOP/NOM} *15* ノットの海上に吹いた.	0.00	0.00	0.31	0.25

表 1: インスタンスの例 $(c, s, a)_d$ および人間と言語モデル (Trans-l) の選好

(NTC; [33]) を使用した．NTC から以下の条件を全て満たす項を含む文およびその文脈を収集した[3].

- 文内で最も後方に出現する能動態の動詞の項である
- 項がとりたて助詞「は」を伴うか，各格に対応する格助詞 (が・に・を) を伴って表出している
- 与格/対格を対象とする場合は，対応する述語が対格/与格の項を持たない
- 文中で対象の項以外が主題化されていない
- 項が各文章の先頭から 2 文目以降に出現する

これらの基準を用いて，$\mathcal{D} = \{(c, s, a)_d\}_{d=1}^{|\mathcal{D}|}$ で表されるアノテーション用データセットを抽出した．ここで c は先行文脈 (同じ文書内の先行文脈)，s は文内文脈，a は前述の基準を満たす項である．この処理により，主格については 1,661 件のデータが得られ，そのうち 939 件はもともと項 a に対して TOP の助詞が伴っており，残りの 722 件は NOM を伴っていた．得られたデータセットの例を表 1 に示す.

[3]主格の場合のみ,「対象とする格以外が主題化されていない」という条件は用いず，また事例を十分な量収集できたため，作業者の先行文脈を読む負荷を考慮して文章の先頭 2 から 4 文目までを対象とした.

3.2.2　アノテーションタスク

各インスタンス $(c,\ s,\ a)_d$ について，先行文脈 c の有無について 2 つの設定でアノテーションを行った．ただし，主格については文内文脈 s の有無についても，すなわち計 4 つの設定でアノテーションを行った．文内文脈 s がある場合には，(6a) のように主格 (花瓶) に続く助詞のみを隠した状態にし，文内文脈 s がない場合には，(6b) のように主格の項 (花瓶) 以外を全て隠した状態にした．

(6) a. 花瓶-__ 部屋に　　あった.
Vase-__ room-DAT exist-PAST.

b. 花瓶-__
Vase-__

次に，空白の部分を補完する助詞として，TOP (は; 主題化する) と NOM (が; 主題化しない) のどちらが自然かをアノテーターに質問した．また，その際には「どちらでも良い」という選択肢も用意した．

各インスタンス $(c,\ s,\ a)_d$ に対して，4 人から 8 人のワーカが選好のアノテーションを行った．最後に，各インスタンス $(c,\ s,\ a)_d$ について，以下のように **主題化率** r_{human} を算出した．

$$r_{\text{human}} = \frac{n_{\texttt{TOP}}}{n_{\texttt{TOP}} + n_{\texttt{NOM}}} \ , \tag{1}$$

ここで，$n_{\texttt{TOP}}$ と $n_{\texttt{NOM}}$ はそれぞれ空欄 (花瓶-____) を TOP と NOM で補完した場合の票数である.「どちらでも良い」の場合，TOP と NOM はそれぞれ 0.5 票とみなした.

3.2.3　文脈の効果の検証

文脈に依存した主題化という現象を分析しやすくするため，アノテーターに対し，3 つの異なる条件でタスクを実施した．表 2 に各条件を示す．ここで，**文脈** c の欄の ✓ は先行文脈をアノテーターに見せることを，**文** s の欄の ✓ は文内文脈を見せるかどうかを示している．文内文脈を表示しない場合は，項 a のみをアノテーターに見せる．

	文脈 c	文 s	項 a	設問: TOP か NOM か
C+S+A	✓	✓	✓	花瓶を昨日割った． 花瓶-__ 部屋にあった．
S+A		✓	✓	花瓶-__ 部屋にあった．
A			✓	花瓶-__

表 2: 主題化の選択を問う際の設定．

設定	ラベル数	ラベルの分布		
		TOP	どちらでも良い	NOM
C+S+A	8,039	55.3%	1.49%	43.2%
S+A	8,094	52.9%	3.47%	43.6%
A	7,621	1.86%	96.7 %	1.42 %

表 3: データセット全体の統計量

その後，各インスタンス $(c,\, s,\, a)_d$ に対する主題化率 r_{human} を異なる設定の下で 3 種類定義した．(i) 設定 C+S+で得られた主題化率 $r_{\text{human}}^{\text{C+S+A}}$, (ii) 設定 S+A で得られた主題化率 $r_{\text{human}}^{\text{S+A}}$, (iii) 設定 A で得られた主題化率 $r_{\text{human}}^{\text{A}}$ とそれぞれした．スコアの例を表 1 に示す． 5 節 では，言語モデルの主題化に対する選好を観察し，人間の選好と比較した．

3.2.4 アノテーションの使用方法

設定 C+S+A および設定 S+A で得られたアノテーション (比較的高い一致率) を主な実験に用いた (5 節)．設定 A のデータは，言語モデルがこの設定で人間と同様にこの判断をするのに困る様子を示すかどうかを検証するための追加解析 (6.1 節) に使用された．その結果，いくつかの興味深い傾向が観察された．文脈の情報を完全に制限した設定 A では，言語モデルは主題化の判断において不当に高い性能を示した．

3.3 クラウドソーシング

3.3.1 ワーカの選定

アノテーションを行うにあたり，日本語を母語とするワーカにアクセスするために，Yahoo!クラウドソーシング[4]を利用した．クラウドワーカーに対し，TOP または NOM を選択するタスクを解くように指示をした．クラウドワーカーを選定するために，まず，1 人のクラウドワーカーが 10 問の設問に答え， TOP または NOM を選択するトライアルタスクを作成した．その際に，TOP または NOM の一方に回答が偏ることが考えられる設問を用意し，10 問のうち 1 問をチェック設問とした．そして，そのチェック設問に正解したワーカを優秀なワーカとみなし，今後のタスクの参加者としてホワイトリストに登録した．この段階で，設定 C+S+A と 設定 S+A では 164 名，設定 A では 153 名の作業者が選出された．ここで，クラウドソーシングは 2 回に分けて実施し，1 回目は C+S+A，C+S の 2 つの設定，2 回目は設定 A のタスクをそれぞれ行った．各セッションでは，トライアルタスクの後，MACE [34] により作業者の信頼度を算出した．そして，信頼度の上位 80% のワーカーを選出した．

3.3.2 アノテーションの実施

3.2 節で収集した 1,661 インスタンスのデータのうち，各インスタンス $(c,\ s,\ a)_d$ に対して，8 人の優秀なワーカが主題化の判断に関する選好をアノテーションした．各ワーカは少なくとも 10 個のインスタンスを解いた．同じワーカーが異なる設定で同じインスタンスにアノテーションを行うことはないようにアノテーションを実施した．全アノテーション終了後，MACE [34] を用いて，信頼度の下位 30%のワーカーを除外する統計的後処理を行った．その結果 4 人以上の選好が付与されなかったデータポイントを削除した．

最終的に，1,355 個の主格の項に対して， TOP または NOM を選択する 23,745 個の選好を収集した．そのうち 758 個のインスタンスはもともと項 a に対して

[4] https://crowdsourcing.yahoo.co.jp/

TOP （助詞として「は」）を伴っており，597 個のインスタンスは NOM(助詞として「が」) を伴っていた．

3.4　統計量

収集した人間のラベル数とラベルの分布を表 3 に示す．先行文脈 c, 文内文脈 s, 項 a の長さは，文字数でそれぞれ 96.3±53.7, 50.1±24.4, 4.5±2.2 (± は標準偏差) であった[5].

アノテーションの一致度は，設定 C+S+A，設定 S+A において，Krippendorff の α でそれぞれ 0.689，0.699 となった[6] ．これらの α の値はデータの信頼度の最低基準 (0.667) を超えている [36].

一方，設定 A での一致度は 0.074 と，基準を大きく下回っていた．このような低いスコアの原因として，大多数のアノテーターが「どちらでも良い」と回答したことが考えられる．このため，クラス分布が歪み，α の値はクラスの不均衡の影響を受けたと考える [37]．また，多くの作業者が「どちらでも良い」と回答したことから，この設定での主題化判断は日本語話者にとって困難であることも考えられる．

4　データセットの分析

実験に先立ち，収集したデータセットの分析を行った．もし，すべてのアノテーションされた選好が先行文脈に依存して変化しないのであれば，我々のデータは言語モデルの談話レベルの振る舞いを分析するのには適していないことになる．この節では，収集したデータセットが文脈依存な性質を含んでいるのかということを調べる．

[5]日本語には明確な単語境界線がない．単語数の近似値として，先行文脈 c，文内文脈 s，項 a はそれぞれ 56.4±31.7, 29.5±14.4, 2.8±1.1 の形態素を持つ．形態素解析には JUMAN [35] を使用した．

[6]本研究では weighted Krippendorff α を用い，距離尺度は TOP $\prec$ 「どちらでも良い」 $\prec$ NOM とした．

言及回数	インスタンス数	$r_{\text{human}}^{\text{C+S+A}}$
0	1,082	0.48 ± 0.43
1	197	0.87 ± 0.25
2+	76	0.90 ± 0.22

表 4: 先行文脈での言及回数に対する人間の主題化の選好の強さ

4.1 主題化の選択における文脈の効果

まず,「日本語では既に言及された旧情報は主題マーカーによってより主題化されやすい」という言語学の分野での理論について，主題化の判断に関する選好と文脈での言及回数の関係を分析する [38]．文脈で何度も言及される要素は主題化率が高い傾向があることが分かった (表 4)．このことは，作成されたデータが談話レベルの言語学的な自然な傾向を反映していることを裏付けている．なお，言及回数は NTC の共参照アノテーションを用いて，名詞と同じエンティティが先行文脈に何回出現したかをカウントしている．

次に，各インスタンス $(c,\, s,\, a)_d$ に対して，文脈に依存した主題化の判断に関する選好の変化を以下のように定量化した．

$$\Delta_{\text{human}} = r_{\text{human}}^{\text{C+S+A}} - r_{\text{human}}^{\text{S+A}} \quad . \tag{2}$$

ここで，Δ は先行文脈の有無による TOP を選好する際の確信度の変化に相当する．直感的には，Δ が大きいほど，先行文脈を見ることで主題化 (TOP) を選好する票数が増えたということを意味する．

図 2 には，先行文脈の有無による主題化の選好度合いの変化の分布 (Δ_{human} の分布) を示す．図 2 より，先行文脈の有無によって人間の主題化の判断に関する選好が変化することがわかった (約 53%のデータポイントが非ゼロの Δ_{human} を持つ).

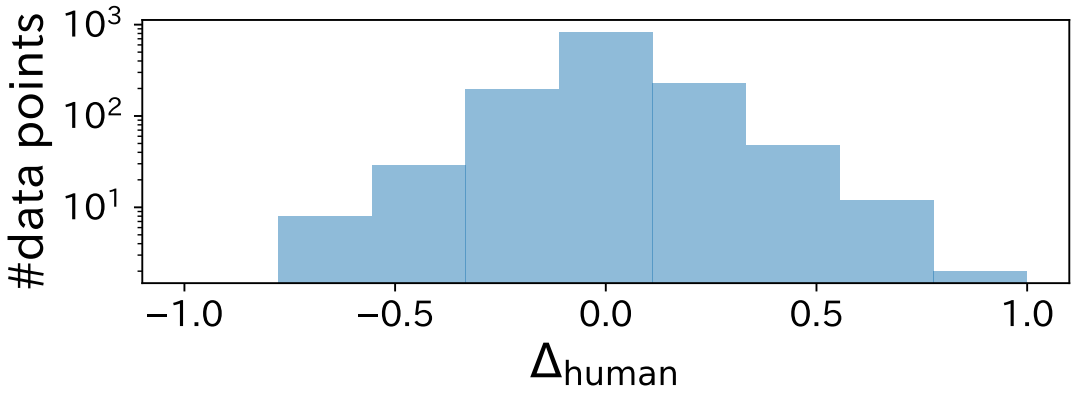

図 2: 先行文脈の有無による主題化率の変化 (Δ_{human}) の分布

4.2 チャレンジングセットの作成

3.3 節で収集したデータの中には，文脈の影響が見られない，すなわち文脈を見せた場合と見せない場合とで作業者の選好がほとんど変化しないインスタンスが確認された．また,「は」と「が」の使い分けは，談話レベルの主題構造以外の観点の作用が無視できないという調査もある [32, 39]．そこで，言語モデルの文脈 (非) 依存な振る舞いを精緻に分析するために，以下の条件のいずれかを満たすインスタンスをチャレンジングセットとして抽出した．

- NTC 上と文脈を見せた場合のアノテーションで主題化された文が選ばれた，かつ文脈の有無による選好の変化の大きさが上位 25%に属する (文脈を見ることで主題化された文への選好が強まった)
- NTC 上と文脈を見せた場合のアノテーションで主題化されていない文が選ばれた，かつ文脈の有無による選好の変化の大きさが下位 25%に属する (文脈を見ることで主題化されていない文への選好が強まった)

最終的に 311 インスタンスがチャレンジングセットとして得られた．そのうち 209 インスタンスはもともと項 a が主題化 (TOP) されており，102 インスタンスは非主題化 (NOM) されていた．チャレンジングセットの統計量を表 5 に示す．このチャレンジングセットと全データセットを用いて，言語モデルの分析を行った (5 節).

設定	ラベル数	ラベルの分布		
		TOP	どちらでも良い	NOM
C+S+A	1,853	65.7%	1.73%	32.5%
S+A	1,900	54.9%	8.47%	36.6%
A	1,732	2.02%	97.1%	0.88%

表 5: チャレンジングセットの統計量

5 実験：人間と言語モデルの比較

言語モデルは人間のような主題化の判断に関する選好を示すのだろうか？この問いに答えるために，我々は言語モデルと人間の選好を比較し，対照した．

5.1 実験設定

5.1.1 言語モデル

我々は，left-to-right の 3 つの言語モデルを用いて分析を行った．400M パラメータの Transformer ベースの言語モデル (Trans-l)，55M パラメータの Transformer ベースの言語モデル (Trans-s)，55M パラメータの LSTM ベースの言語モデル (LSTM) [40, 41] の 3 種類を用いた．日本語の新聞記事と Wikipedia から約 3M の文章 (トークン化前は 3.4GB) を用いて，100K 回のパラメータ更新を行いながら学習した．言語モデルへの入力は JUMAN [35] で形態素に分割され，さらに sentencepiece [42] でサブワードに分割した．ここでは unigram model [43] を用いた[7]．各言語モデルの学習時のハイパーパラメータを表 6 に示す．

5.1.2 言語モデルの選好

式 (1) と同様に，言語モデルにおける主題化の選好を定量的に評価を行う．例えば，設定 S+A では，各インスタンス $(c, s, a)_d$ に対して，$s_{\texttt{TOP}}$ と $s_{\texttt{NOM}}$ の文ペア

[7] coverage=0.9995, vocab size=100,000 とした．

パラメータ		TRANS-L	TRANS-S	LSTM
Fairseq model	architecture	gpt2	gpt_small	lstm_lm
	share-decoder-input-output-embed	True	True	True
	embed_dim	1,024	384	400
	ffn_embed_dim	4,096	2048	-
	hidden_size	-	-	1,024
	layers	24	8	2
	heads	16	6	-
	dropout	0.1	0.1	0.1
	attention_dropout	0.1	0.1	-
Optimizer	algorithm	AdamW	AdamW	AdamW
	learning rates	5e-4	5e-4	1e-3
	betas	(0.9, 0.98)	(0.9, 0.98)	(0.9, 0.98)
	weight decay	0.01	0.01	0.01
	clip norm	0.0	0.0	0.0
Learning rate scheduler	type	inverse_sqrt	inverse_sqrt	inverse_sqrt
	warmup updates	4,000	4,000	4,000
	warmup init lrarning rate	1e-7	1e-7	1e-7
Training	batch size	61,440 tokens	61,440 tokens	20,480 tokens
	sample-break-mode	none	none	none

表 6: 言語モデルの学習時のハイパーパラメータ

を作成した．そして，式 (1) に沿い，言語モデルの**主題化率** r_{LM} を以下のように算出した．

$$r_{\mathrm{LM}} = \frac{n(s_{\texttt{TOP}})}{n(s_{\texttt{TOP}}) + n(s_{\texttt{NOM}})} \ ,$$
$$n(s) := \prod_{i=1}^{|s|} p(w_i|w_{<i})^{\frac{1}{|s|}} \ , \tag{3}$$

ここで，$n(s)$ は，言語モデルから得られる各文のパープレキシティである．

各インスタンスについて, 2 種類の主題化率 r_{LM} を計算した．先行文脈がある場合の $r_{\mathrm{LM}}^{\mathrm{C+S+A}}$ と，先行文脈がない場合の $r_{\mathrm{LM}}^{\mathrm{S+A}}$ に該当する．先行文脈ありの設定での言語モデルの主題化率 $r_{\mathrm{LM}}^{\mathrm{S+A}}$ については，式 (3) の $n(s)$, の代わりに，条件付き確率 $n(s|c) = \prod_{i=1}^{|s|} p(w_i|c, w_{<i})^{\frac{1}{|s|}}$, を用いて計算した．

5.1.3 評価指標

言語モデルの TOP を選択する際の確信度が人間に近いかどうかという観点での分析を行う．実験結果では，$r_{\mathrm{human}}^{\mathrm{C+S+A}}$ と $r_{\mathrm{LM}}^{\mathrm{C+S+A}}$ の順位相関係数 (以下，ρ_r) を報告する．人間が TOP を好むほど，言語モデルもそれを好むと予想される．

さらに，言語モデルの先行文脈の考慮の仕方が人間らしいかどうかを分析するために，式 (2) に類似した言語モデルの主題化率の変化量を以下のように計算した．

$$\Delta_{\mathrm{LM}} = r_{\mathrm{LM}}^{\mathrm{C+S+A}} - r_{\mathrm{LM}}^{\mathrm{S+A}}.$$

言語モデルと人間の文脈に依存した主題化の判断に関する選好の変化がどれだけ似ているかを定量的に評価するために，Δ_{human} と Δ_{LM} の順位相関係数を報告した (以下，ρ_Δ と呼ぶ).

さらに，ゴールドラベルとして，コーパス (NTC) での助詞の選択 (すなわち，コーパス上である項 a に TOP と NOM のどちらが用いられていたか) に対する人

モデル	設定	ρ_r	ρ_Δ	Macro F1	TOP F1	NOM F1
Trans-l	C+S+A	0.67	−0.12	83.5	88.8	78.3
	S+A	0.60		81.7	87.6	75.8
Trans-s	C+S+A	0.72	−0.07	85.3	89.5	81.1
	S+A	0.61		83.7	88.1	79.3
LSTM	C+S+A	0.69	−0.20	81.9	86.9	77.0
	S+A	0.62		82.3	87.1	77.5
Human	C+S+A	-	-	(100)	(100)	(100)
	S+A	-		81.1	86.5	75.7

表 7: チャレンジングセットでの選択の傾向 (主格)

間と言語モデルの F1 スコアを報告する．ここで，人間／言語モデルは主題化率 $r_{\text{human/LM}}$ が 0.5 を超えた場合， TOP を選択し，それ以外は NOM を選択したと見なした．このコーパスは新聞記事のデータが元となっているため，これらの判断は多かれ少なかれ日本語の主題マーカーの適切な使い方を反映していると考えられる．

5.2 結果

表 7 にチャレンジングセットでの人間と言語モデルの主格における主題化の選択の傾向を示す．すべての設定において，主題化の選好の相関 (ρ_r 列) に関して，人間と言語モデルの間に中程度の相関が観察された．

5.2.1 先行文脈の考慮の仕方

主題化率 ρ_r の相関はやや高いものの，文脈の有無による選好の変化の度合いの相関 ρ_Δ は負であり，選好文脈の有無による主題化選好の変化は，人間と言語

モデル	設定	ρ_r	ρ_Δ	Macro F1	TOP F1	NOM F1
Trans-l	C+S+A	0.80	−0.04	88.1	89.3	86.8
	S+A	0.78		87.5	88.8	86.2
Trans-s	C+S+A	0.80	−0.02	87.7	88.8	86.5
	S+A	0.78		87.3	88.3	86.3
LSTM	C+S+A	0.79	−0.01	85.2	86.4	84.1
	S+A	0.78		85.3	86.4	84.2
Human	C+S+A	-	-	89.7	91.0	88.4
	S+A	-		89.1	90.4	87.8

表 8: データセット全体での選択の傾向 (主格)

モデルで異なることが確認された．このことから，言語モデルと人間の文脈の考慮の仕方には大きな乖離があることが示された．

5.2.2 文内文脈の考慮の仕方

設定 C+S+A と設定 S+A での F1 スコアの違いに注目すると，言語モデルの結果にはほとんど差がないことがわかる．また，設定 S+A での F1 スコアについてみると，言語モデルは人間よりも高いスコアを報告している．さらには，LSTM については，先行文脈を考慮しない方がよりコーパスと一貫した主題化の判断がなされるという，やや奇妙な傾向が確認された．これらの結果は，言語モデルが何らかの非文脈的な手がかり (特に，文内文脈の情報) を用いて主題化に関する判断を行い，人間とは異なる汎化を行っている可能性を示唆する．

5.2.3 人間と言語モデルの選好の例

表 1 に，主題化に関する人間と言語モデルの選好の例をいくつか示す．表 1 の 3 番目の例はチャレンジングセットに属し，主題化に関する文脈に依存した判断に関して，人間と言語モデルの間の不一致を表している．人間は先行文脈を考慮

して初め TOP を選択するが，言語モデルの選好は先行文脈の情報を提供しても変化しないことがわかった．この点で人間と言語モデルの主題化に関する判断には乖離があることがわかる．

5.2.4　データセット全体での結果

また，チャレンジングセットに限定せず，クラウドソーシングで作成したデータセット全体においても，人間と言語モデルの主題化に関する判断の傾向を報告する．表 8 にその結果を示す．注目すべきは，データセット全体においても，人間と言語モデルの間で文脈の影響度合い ρ_{Δ} の相関がほとんどないことである．主題化の判断における確信度の強さの相関 ρ_r については，中程度の相関があることを踏まえると，言語モデルは「は」か「が」かというレベルの判断では人間と同様の選好を示すが，選好文脈の考慮の仕方という点では人間と乖離した傾向を示した．

6　分析

6.1　文脈の情報を極端に制限した場合の振る舞い

実験結果から，言語モデルの主題化に関する判断は人間と乖離しているとともに，非文脈依存であることが示された．このような言語モデルの非人間的な行動をさらに理解するために，文脈情報が極端に少ない状況での言語モデルの行動を調査した．

3.4 節で紹介したように，人間は主格の項のみを表示する設定 (設定 A; 表 2) で主題化に関する判定に苦労することが観察されている．そこで，言語モデルがデータセットを設定 A の条件で扱った際にも，人間のような困難を示すかどうかを検証した．なお，この分析の目的は，言語モデルが人間から乖離する状況を見つけることである．

モデル	ρ_r	F1		
		Macro	TOP	NOM
Trans-l	0.18	63.8	71.1	56.5
Trans-s	0.18	65.9	72.3	59.4
LSTM	0.17	65.1	71.9	58.4
Human	-	36.8	14.0	59.3

表 9: 設定 A(主格の項のみを提示) での選択の傾向

6.1.1 言語モデルの選好

データセット中の主格の項 a について，a のみを言語モデルに見せた場合の主題化率を計算した．具体的には，言語モデルに文脈を見せずにある項とそれに伴う助詞のみ，例えば 花瓶が と 花瓶はを見せ，各文のパープレキシティを計算し，その後，式 3 に従ってこれらの主題化率を計算した.

6.1.2 結果

表 9 に，言語モデル/人間に主格の項 a のみを示した場合の主題化の判断の傾向を示す．まず，言語モデルの F1 スコアは人間よりも予想以上に高い値となっていた．これは，主題化は談話レベルの現象であるにもかかわらず，言語モデルが文脈情報に依存せずに助詞 (TOP か NOM のどちらを選ぶのが良いか) を何らかの形で予測できることを示唆している．次に，言語モデルと人間との主題化率の相関は非常に低かった．この結果は，言語モデルの主格の項に対する主題化の選好が人間の選好と乖離していることを示唆している.

6.2 主格以外の格での主題化に関する振る舞い

3 節で述べたように，主題化は主格のみに限らず，与格や対格でも生じる．そこで，与格・対格などの主題化，すなわち語順レベルでの変化が起こる文脈依存の言語現象についても分析を行う．まず，主格の場合と同様の手順で，与格・対

格でのデータセットを作成し，人手評価を行った．その後，同一のデータを用いて人間と言語モデルの選好の比較を行った．

6.2.1 与格・対格での主題化に関するデータセットの作成

3 節で示した条件を満たしたある項 a を含む文 s について，主題構造の観点でのみ異なる文ペアを作成した．文 s で対象とする項 a が主題化されていた場合 (とりたて助詞「は」を伴う場合) は，とりたて助詞を格助詞に変え，項を基本語順位置に移動させることで，主題化が生じていない無標の文を作成した．ここで，与格・対格について無標の文を作成する際には述語の直前に項を移動するようにした．なお，与格と対格が同時に出現する文は省いているため，基本語順において与格と対格のどちらが先かという議論は生じない．逆に，項が主題化されていなかった場合は，とりたて助詞「は」を付与して文頭に移動することで，主題化が生じている有標の文を作成した．この操作の結果，ある項 a について，(文脈 c，項が主題化された文 $s^{\texttt{TOP}}$，項が主題化されていない文 $s^{\texttt{NOT}}$) の 3 つ組が収集された．得られたデータの例を表 10 に示す．

作成したデータセットを用いて主格の時と同様の方法でアノテーションを行った．得られたデータセットの統計量を表 11 に示す．3.2 節で作成したチャレンジングセットと同様に，与格・対格についてもチャレンジングセットを作成した．以降の結果では，チャレンジングセットおよびデータセット全体での人間と言語モデルの与格・対格での主題化における振る舞いを報告する．

6.2.2 チャレンジングセットでの結果

チャレンジングセットでの人間と言語モデルの選択の傾向を表 12 に示す．文脈ありの設定では，人間と言語モデルの判断について，与格，対格とも正の相関が観察された ($\rho_r \geq 0.48$)．したがって，ある文脈に続く文として，ある要素を主題化するか非主題化するかというレベルの判断ではある程度の相関があった．一方，文脈の影響度の相関 ρ_Δ は，与格，対格のいずれにおいても非常に弱く，主

表 10: 与格・対格でのデータセットの例 (機械的に作成した文の文頭には*を付与した)

文脈 c	文 $s^{\texttt{TOP}}$	文 $s^{\texttt{NOT}}$
二月に公開が予定されている映画「フランケンシュタイン」の監督・主演をしたケネス・ブラナーが来日した．ブラナーは「から騒ぎ」などのシェークスピア原作映画を製作するなど，活発な創作活動を続けている．	**「フランケンシュタイン」は**フランシス・フォード・コッポラの製作，ロバート・デ・ニーロの主演という面でも見逃せない．	*フランシス・フォード・コッポラの製作，ロバート・デ・ニーロの主演という面でも**「フランケンシュタイン」を**見逃せない．
千畳敷付近は，星空の名所.「降るような星空でした」と星野さんが言うように，この夜も満天の星空が楽しめたという．星野さんらは三日午前中に木曽駒ケ岳に登頂.	***昼食は**，昼ごろに宝剣山荘に着き，全員で済ませた．	昼ごろに宝剣山荘に着き，全員で**昼食を**済ませた．
日米科学技術協力協定に基づく日米合同高級委員会が十二日，東京で開かれ，オゾン層破壊の解明につながる北極圏上空の大気観測を日米共同で行うことを新たに決めた.	**委員会には**日本から田中真紀子科学技術庁長官，米側からジョン・ギボンズ大統領補佐官らが出席した.	*日本から田中真紀子科学技術庁長官，米側からジョン・ギボンズ大統領補佐官らが**委員会に**出席した.

題化に対する文脈の影響について，人間と言語モデルで乖離があることが示唆された.

また，人間の F1 スコアは文脈の有無によって大きく変化しているのに対し，言語モデルの F1 スコアはほとんど変わらず，言語モデルが不当に文脈非依存な選択を行っていることが示唆された．特に与格や対格などコーパスと比較して人間の選択が著しく変わる設定において，なぜ言語モデルが文脈なしにコーパスと一貫した選択が行えているのかは興味深い．なお，文脈依存セットの作成では，人間とコーパスの判断が一致することを条件に課したため，文脈ありの設定で人間の F1 スコアは 100 となる．また文脈の有無で主題化における選好が大きく変わ

表 11: データセットの統計量 (主格・与格・対格)

格	インスタンス数	ラベル数
主格	1,355	16,133
与格	256	2,861
対格	543	6,116

表 12: チャレンジングセットでの人間と言語モデルの選択の傾向 (与格・対格)

モデル	文脈	与格			対格		
		F1	ρ_r	ρ_Δ	F1	ρ_r	ρ_Δ
TRANS-L	○	88.4	0.68	−0.19	86.2	0.62	−0.01
		89.9	−0.56		86.9	−0.65	
TRANS-S	○	88.2	0.71	−0.16	82.3	0.53	0.01
		86.9	−0.51		80.8	−0.59	
LSTM	○	88.2	0.69	−0.40	80.6	0.48	−0.25
		89.8	−0.50		79.8	−0.58	
HUMAN	○	(100)	-	-	(100)	(1.0)	-
		19.8	-		18.8	-	

るデータを収集したため，文脈がないときに人間は積極的にコーパスと逆の選択をとる可能性があり，F1 スコアがチャンスレートを下回ることは不自然でないと考える．

6.2.3 データセット全体での結果

表 13 に 6.2 節で作成したデータセット全体での人間と言語モデルの選択の傾向を示す．データセット全体においても与格，対格いずれの文脈の影響度合いの相関 ρ_Δ はほとんどなく， 5 節の結果と一貫して，人間と言語モデルの文脈の考慮の仕方には乖離があることが確認された．

F1 スコアについては，主格における主題化の判断では，人間も言語モデルも 90 ポイント前後と高い値を示した．また，データセット全体では人間も言語モデル

表 13: データセット全体での人間と言語モデルの選択の傾向 (与格・対格)

モデル	文脈	与格			対格		
		F1	ρ_r	ρ_Δ	F1	ρ_r	ρ_Δ
TRANS-L	○	89.3	0.26	−0.12	83.6	0.25	0.03
		88.5	−0.06		81.6	−0.06	
TRANS-S	○	85.0	0.26	−0.10	84.9	0.21	0.04
		83.6	−0.08		82.9	−0.01	
LSTM	○	81.9	0.24	−0.11	79.3	0.21	−0.05
		82.3	−0.03		78.9	−0.02	
HUMAN	○	62.5	-	-	59.6	-	-
		43.4	-		49.0	-	

も文脈の有無による F1 スコアの変化が小さく，主格における主題化の判断，すなわち「は」と「が」の選択は文内の情報のみで十分可能であることが示唆された．一方で，与格と対格における主題化の判断では，言語モデルは 80 から 90 ポイント前後の値となっているのに対し，人間は 50 から 60 ポイント前後の値となっていた．この値は 2 値分類のチャンスレートに近いことから，人間にとって与格と対格における主題化判断は一方に選好が偏るものではないことが考えられる．しかしながら，文脈を見ることでコーパスと一貫した判断がわずかではあるができるようになるという点で，人間は与格や対格における主題化の判断を文脈に依存して行っていることが改めて確認された．人間と比べて言語モデルは F1 スコアで高いスコアとなっているが，依然として文脈に依存した判断を行っていないことから，過剰にコーパスにフィットしていたり，N-gram レベルの情報で判断をしていたりする可能性が考えられる．人間にとっては困難な判断を言語モデルがどんな情報を頼りに行っているのかということについては今後さらに調査したい．

7　議論

7.1　言語学への貢献

言語学への貢献という観点では，本研究で作成したデータセット自体が価値を持つものであると考える．日本語の主題化は半世紀以上前から言語学の分野で注目されている [31, 32, 38]．しかし，大規模コーパスやクラウドソーシングを利用して作成されたリソースは限られている．主題化に関して大規模な人手評価を行った本データセットが今後の更なる研究に役立つことを期待したい．

事実として本データセットからは以下のような興味深い主題化における人間の選好が観察された．例えば，言及頻度と主題化選好の関係 (表 4) は，情報の新旧と主題化に関する選好の関係を実証的に支持すると見なすことができるだろう [38]．また，設定 S+A においては，比較的高い一致率が得られ，データセット全体から，文内文脈が主題化判断のための有益な手がかりを提供することが示唆された．このことは，主題マーカである TOP の使用が文脈に依存した現象となっていないことが多いことを意味し，この知見は [44] と一致する．なお，このような文脈に依存しないインスタンスはチャレンジングセットから除外されていることに注意されたい．

7.2　コヒーレンスモデルの検証

今回は標準的な言語モデルのみを評価したが，文脈を考慮できているかという点で，コヒーレンスモデルので分析も行い歌いと考える．さらに，コヒーレンスモデルには，エンティティベースのコヒーレンスモデル [45] や，文章の一貫性に関する目的関数をもとに明示的にコヒーレンスについて学習させたニューラル言語モデル [46] など，コヒーレンスモデルにはいくつかの選択肢が存在する．このようなモデルが，主題化の判断に関する選好において，より人間に近い振る舞いを示すかどうかを調べることは興味深いことである．

8　おわりに

本研究では，談話において必要不可欠な側面である主題化に関して，言語モデルと人間の選好の比較を行った．実験結果から，主題化に関する振る舞いにおいて言語モデルは人間と乖離していることが示された．この結果は，人間らしい談話レベルの言語的な知識を言語モデルに獲得させるためには，モデルのアーキテクチャや学習において更なる帰納バイアスの必要性を示唆している．

謝辞

本研究を進めるにあたり，多くの方々のご協力，ご助言をいただきました．お忙しいなか，研究活動だけでなく進路や学生生活に関してもご指導やご助言をいただいた主指導教員である乾 健太郎教授，鈴木 潤教授に心から感謝申し上げます．また，本論文の審査をお引き受けくださいました，本学の塩入 諭教授に深く感謝申し上げます．東北大学の栗林 樹生さん，徳久 良子さん，阿部 香央莉さんには，日々の研究活動の際のご助言やご協力に加え，学生生活を送るうえでの多くのご助言をいただきましたことに，心から感謝申し上げます．最後に，研究会や日々の議論のなかで多くのご助言をいただきました乾研究室，鈴木 (潤) 研究室，松林研究室の皆様，これまで学生生活や研究活動においてお世話になったすべての皆様に感謝申し上げます．

参考文献

[1] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 521–535, 2016.

[2] Jey Han Lau, Alexander Clark, and Shalom Lappin. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, Vol. 41, No. 5, pp. 1202–1241, 2017.

[3] Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of EMNLP*, pp. 1192–1202, 2018.

[4] Yoav Goldberg. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.

[5] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 625–641, 2019.

[6] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 377–392, 2020.

[7] Jiwei Li and Dan Jurafsky. Neural net models of open-domain discourse coherence. In *Proceedings of EMNLP*, pp. 198–209, September 2017.

[8] Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. Do massively pretrained language models make better storytellers? In *Proceedings of CoNLL*, pp. 843–861, 2019.

[9] Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, Vol. 21, No. 2, pp. 203–225, 1995.

[10] Michael Halliday, Christian MIM Matthiessen, and Christian Matthiessen. *An Introduction to Functional Grammar*. Routledge, 2014.

[11] Eva Hajicová and Jiří Mírovskỳ. Discourse coherence through the lens of an annotated text corpus: A case study. In *Proceedings of LREC*, pp. 1637–1642, 2018.

[12] David E Rumelhart and James L McClelland. On learning the past tenses of english verbs. Technical report, California Univ., San Diego, La Jolla. Inst. for Cognitive Science, 1985.

[13] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of ACL*, pp. 1725–1744, 2020.

[14] Enric Vallduví. *The Informational Component*. Ph.D. thesis, University of Pennsylvania, Philadelphia, 1990.

[15] Eleni Miltsakaki. Locating topics in text processing. In *CLIN*, pp. 127–138, 1999.

[16] Wallace L. Chafe. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. 1976.

[17] T Givón. *Topic Continuity in Discourse: A quantitative cross-language study*. John Benjamins, 1983.

[18] John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*, pp. 4129–4138, June 2019.

[19] Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of EMNLP*, pp. 6862–6868, November 2020.

[20] Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. Evaluating commonsense in pre-trained language models. In *Proceedings of AAAI*, Vol. 34, pp. 9733–9740, 2020.

[21] Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. Probing for referential information in language models. In *Proceedings of ACL*, pp. 4177–4189, July 2020.

[22] Shiva Upadhye, Leon Bergen, and Andrew Kehler. Predicting reference: What do language models learn about discourse models? In *Proceedings of EMNLP*, pp. 977–982, November 2020.

[23] Lalchand Pandia, Yan Cong, and Allyson Ettinger. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of CoNLL*, pp. 367–379, November 2021.

[24] Murathan Kurfalı and Robert Östling. Probing multilingual language models for discourse. In *Proceedings of RepL4NLP*, pp. 8–19, August 2021.

[25] Fajri Koto, Jey Han Lau, and Timothy Baldwin. Discourse probing of pretrained language models. In *Proceedings of NAACL-HLT*, pp. 3849–3864, June 2021.

[26] Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. Syntaxgym: An online platform for targeted evaluation of language models. In *Proceedings of ACL (System Demonstrations)*.

[27] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *Proceedings of ICLR Workshop*. OpenReview.net, 2017.

[28] Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. Pareto probing: Trading off accuracy for complexity. In *Proceedings of EMNLP*, pp. 3138–3153, November 2020.

[29] Kazuhiro Teruya. Metafunctional profile of the grammar of japanese. *Language typology. A functional perspective*, pp. 185–254, 2004.

[30] Kazuhiro Teruya. *A systemic functional grammar of Japanese*. Bloomsbury Academic, 2007.

[31] Susumu Kuno. *Nihon bunpō kenkyū (Study of Japanese Grammar)*. Taishukan Shoten, 1973.

[32] Hisashi Noda. *Wa to ga (Wa and ga)*. Kurosio Publishers, 1996.

[33] Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the linguistic annotation workshop*, pp. 132–139, 2007.

[34] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with mace. In *Proceedings of NAACL-HLT*, pp. 1120–1130, 2013.

[35] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. In *Proceedings of LREC*, pp. 1344–1347, 2006.

[36] Klaus Krippendorff. *Content Analysis: an Introduction to its Methodology*. Sage publications, 2004.

[37] László A. Jeni, Jeffrey F. Cohn, and Fernando De la Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *In Proceedings of ACII*, pp. 245–251, 2013.

[38] Daizaburo Matsushita. *Hyōjun nihon kougohou (Standard Japanese Spoken Method)*. Chūbunkan Shoten, 1930.

[39] 日本語記述文法研究会. 現代日本語文法 5 とりたて・主題. くろしお出版, 2009.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NeurIPS*, pp. 5998–6008, 2017.

[41] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[42] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP*, pp. 66–71, 2018.

[43] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of ACL*, pp. 66–75, 2018.

[44] Satoshi Imamura, Yohei Sato, and Masatoshi Koizumi. Influence of Information Structure on Word Order Change and Topic Marker WA in Japanese. In *Proceedings of PACLIC*, 2014.

[45] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, Vol. 34, pp. 1–34, 2008.

[46] Prathyusha Jwalapuram, Shafiq Joty, and Xiang Lin. Rethinking self-supervision objectives for generalizable coherence modeling. In *Proceedings of ACL*, pp. 6044–6059, 2022.

発表文献一覧

受賞一覧

1. 人工知能学会言語・音声理解と対話処理研究会 (SLUD) 第 96 回研究会 (第 13 回対話システムシンポジウム) 第 5 回対話システムライブコンペティション 優秀賞

2. 人工知能学会言語・音声理解と対話処理研究会 (SLUD) 第 90 回研究会 (第 11 回対話システムシンポジウム) 第 3 回対話システムライブコンペティション 優秀賞

国際会議論文

1. Riki Fujihara, Tatsuki Kuribayashi, Kaori Abe, Ryoko Tokuhisa, Kentaro Inui. Topicalization in Language Models: A Case Study on Japanese. In Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022), pp.851–862, Oct 2022.

2. Riki Fujihara, Tatsuki Kuribayashi, Kaori Abe, Kentaro Inui. Topicalization in Language Models: A Case Study on Japanese. Non-archival submission for the 2021 ACL Student Research Workshop (ACL SRW 2021), August 2021.

国内会議・研究会論文

1. 藤原吏生，栗林樹生, 徳久良子，乾健太郎．主題化における人間と言語モデルの対照．言語処理学会第 29 回年次大会 (NLP2023)，March 2023.

2. 守屋彰二, 塩野大輝, 岸波洋介, 藤原吏生, 木村昴, 松本悠太, 曽根周作, 赤間怜奈, 鈴木潤, 乾健太郎. aoba_v3 bot: 多様な応答生成モデルとルールベースを統合したマルチモーダル雑談対話システム. 第 96 回 人工知能学会 言

語・音声理解と対話処理研究会 (SLUD) 第 13 回対話システムシンポジウム, December 2022.

3. 藤原吏生，栗林樹生, 乾健太郎．人と言語モデルが捉える文の主題．言語処理学会第 27 回年次大会 (NLP2021)，March 2021.

4. 藤原吏生，岸波洋介，今野颯人，佐藤志貴，佐藤汰亮，宮脇峻平， 加藤拓真，鈴木潤，乾健太郎．ILYS aoba bot: 大規模ニューラル応答生成モデルとルールベースを統合した雑談対話システム． 人工知能学会 言語・音声理解と対話処理研究会 (SLUD) 第 90 回研究会， November 2020.

5. 藤原吏生，栗林樹生, 乾健太郎．日本語言語モデルが選択する文形式の傾向と文脈の影響 ー 主題化・有標語順について．NLP 若手の会第 15 回シンポジウム (YANS2020)，September 2020.