2022年度博士学位論文

文の構造知識に基づく自然言語文生成

お茶の水女子大学大学院
人間文化創成科学研究科
理学 専攻

熊谷 香織

2023年　3月

要旨

人間は，日常的に様々な機会で情報を文章化し表現している．日常生活においては，他人と一緒に料理や掃除をするときは自分が依頼したい作業内容を文章化し相手に伝えている．仕事をする上で，複雑な手順を伴う作業を複数人に伝達する必要があるときは，作業手順ごとに行う動作を文章化することで，多くの人に複雑な作業を正確に伝達することができる．人間の持つ文生成の能力は，自分以外の相手と円滑に生活したり，仕事を進めていくうえで必要不可欠な能力といえる．

機械に人間同様の文生成の能力を実装することができれば，我々の生活や仕事を大幅に効率化できると考えられる．例えば，日常生活における料理や掃除の作業について，普段の作業内容を自動で文章化し記録しておくことができれば相手への情報伝達の時間を短縮できる．また，仕事で複雑な手順を伴う作業について，事前に作業内容を撮影し，これを自動で文章化できればマニュアル作成の手間が省けると共に，無意識に行っている作業内容も自動記録されることでより正確に作業を伝達することができる．さらには，目の不自由な人が周囲の環境を把握するために，機械が周囲の環境について説明する文を自動生成することができれば，目の不自由な人がより安心して生活を送れるようになる．人間の伝えたい内容を自動的に文章化する機能は，我々人間の生活や仕事を効率化することができたり，人間の能力拡張に有効活用できると考えられる．

日常の生活や仕事の効率化や，人間の能力拡張に向けて，これまで数多くの文生成に関するタスクが設定され，文生成手法の研究が為されてきた．例えば，画像キャプショニングタスクは画像中の情景や活動を説明する文を生成するタスクである．画像キャプショニングタスクのデータセット MSCOCO は，同じ画像に対して 5 つの正解文が付与されている．正解文から画像中の同じ人物に対して異なる単語や語句を使って説明していることが分かる．例えば，同じ画像中の “二人の女性” に対して “a woman and another woman” や “two women” という表現で説明する．同じ画像中の “人々” に対して “dozens of people” や “a

bunch of people” という表現で説明する．また，それぞれの単語は適切な順番で並び，人間が理解可能な文を構築している．例えば，正解文 “People are walking across the street at an intersection.” は，文頭から “主語（People）” → “述語（are walking）” → “副詞句（across the street at an intersection）” という順番で並び，人間が定めた文法構造規則に従った文を構築している．また，文頭から “動作主格（People）” → “動詞（are walking）” → “対象格（across the street）” → “場所格（at an intersection）” という順番で並び，人間が定めた意味構造的規則に従った文を構築している．上述の通り，画像に対して，正しくかつ柔軟に単語や語句を選択しながら，人間が定めた構造規則に従って単語や語句を並べることで文を生成している．また，画像の視覚的情報を説明するタスクの他にも，長い文書を要約する文書要約タスク，質問文に対して適切な回答文を生成する質問応答タスク，自然な会話文を生成する対話生成タスクなど様々な文生成に関するタスクがある．文書要約タスクでは，文書中の重要な情報を説明する単語を正確にかつ柔軟に選択し，人間が定めた規則に従って選択した単語を適切な語順に並べ，正しい構造を持つ文を生成する必要がある．質問応答タスクや対話生成タスクも同様に，回答文や会話文中で使用され得る単語を正確かつ柔軟に選択し，人間が定めた規則に従って選択した単語を適切な語順に並べ，正しい構造を持つ文を生成する必要がある．上述の通り，人間同様の文生成の能力を機械に実装するために，タスクや状況に応じて正確かつ柔軟に単語や語句を選択し，人間の定めた文法構造的規則や意味構造的規則に基づいた正しい構造の文を生成する必要がある．

しかしながら正確かつ柔軟な単語選択を実現しつつ，人間の定めた規則に基づいた正しい構造の文を生成することは難しい．例えば，テンプレートベースの手法は，予め定めたテンプレート（構造）に従いながら単語選択を行うため正しい構造の文を生成可能だが、柔軟な単語や語句の選択はできない．大量の文書データから尤もらしい単語順を学習し文を生成する ngram ベースの手法は，連続する単語順のみを考慮しており，離れた単語同士の関係を考慮できないことから適切な構造をもつ文の生成が困難である．近年急速な発展を遂げているニューラルネットワークベースの手法は，連続する単語順のみならず離れた単語同士の関係性を学習する Transformer と呼ばれる手法が提案されている．離れた単語同士の関係性を学習することで，尤もらしい文の構造規則を暗に学習することができ，適切な構造をもつ文の生成が可能である．しかしながら大量の文書データを学習したニューラルネットワークは，データ内で頻出する単語を使った文の確率を高くする傾向がある．頻出単語が偏って選択され，柔軟な単語や語句の選択は困難である．上述の通り，正確かつ柔

軟に単語や語句を選択しながら，正しい構造をもつ文の生成を実現することは困難である．

本論文では，人間同様の文生成の能力を機械に実装するために，正確かつ柔軟に単語や語句を選択し，人間が定めた構造的規則に基づき正しい構造を持つ文の生成を実現する方法を検討する．具体的には，文の構造的規則として，文法構造的規則と意味構造的規則の二種類の規則の活用方法についてそれぞれ検討し，各規則の文生成への有効性を検証する．文法構造的規則として文脈自由文法を採用し，文脈自由文法を適用して構築される構文木を生成する．適切な構文木の探索アルゴリズムとして広大な探索範囲を効率的に探索可能なモンテカルロ木探索アルゴリズムを使用する手法を提案する．タスクの設定としては，文の主要な要素となる重要単語のセットを入力として，それらの単語を使用しながら，正確かつ柔軟な単語や語句の選択と，正しい文法構造を備えた文の生成を目指す．次に，意味構造的規則として格構造ラベルを採用し，文生成時に格構造ラベルを条件として与え，与えられた格構造ラベルに対応する単語を選択しつつ，適切な格構造ラベルの順番を同時に推定する手法を提案する．タスクの設定としては，画像を入力として，画像に対応する正しくかつ柔軟な単語や語句を選択しつつ，画像を説明する文として正しい意味構造をもつ文の生成を目指す．画像の特徴抽出と，適切な格構造ラベルの順番推定と，格構造ラベルに対応する正しくかつ柔軟な単語推定とを同時に行う end-to-end 構造の Neural Network を提案する．さらに，Neural Network が適切な文の意味構造を学習するために，離れた単語同士の関係性や格構造ラベル同士の関係性を学習可能な Transformer ベースの画像キャプショニング手法をベースラインとして使用する．各手法の実験では，文法構造的規則や意味構造的規則の文生成への有効性を検証する．

キーワード： 文生成，文法構造的規則，意味構造的規則，文脈自由文法，格構造ラベル，モンテカルロ木探索アルゴリズム，ニューラルネットワーク

Abstract

Humans create sentences to convey information in various situations on a daily life. In our daily life, when we cook or clean with other people, we write down the contents of the work we want to request and tell them to the other person. In the work scene, in order to communicate the complicated procedures to multiple people, we create sentences explaining each work procedure. Creating a manual consisting of sentences for each procedure, we can accurately convey complex procedures to many people. The ability of humans to generate sentences is an indispensable ability to live smoothly with other people and to proceed with work.

By equipping machines with the ability to generate sentences like humans, we can greatly improve the efficiency of our lives and work. For example, by automatically recording the procedures of daily work such as cooking and cleaning, the time required to tell information to the other people is possible to shorten. In the case of work involving complicated procedures, by taking videos of the work in advance and automatically documenting it, the time of creating a manual can be saved. In addition, by automatically recording the details of the procedure done unconsciously, you can convey more accurately. Furthermore, the machine automatically generates sentences explaining the surrounding environment for the visually impaired person. The ability can make the lives of visually impaired people more peace of mind. The ability to automatically generate sentences that explain what people want to convey can make our lives and work more efficient, and can expand human capabilities.

In this paper, we use the structural rules of sentences as external knowledge, and examine the method of generating sentences according to the structural rules of sentences. Sentences with diverse words and phrases can be generated by allowing diversity of words

and phrases within the scope of sentence structure rules. As structural rules of sentences, two types of knowledge, grammatical structural rules and semantic structural rules, are used, and the effectiveness of each rule for sentence generation is examined. We adopt a context-free grammar as the grammatical structural rules of sentences, and propose a method to generate a parse tree constructed by applying the context-free grammar. In order to efficiently search the appropriate parse tree, we employ a Monte Carlo tree search algorithm. We adopt case structure labels as semantic rules for sentences. Based on a set of case structure labels, we propose a neural network that predicts the words corresponding to each given case structure label and at the same time predicts the order of the case structure labels so as to have an appropriate semantic structure. Experiments verify the effectiveness of the proposed method. From our experimental results, we show the effectiveness of grammatical structural rules and semantic structural rules for sentence generation methods.

key words: Sentence generation, Grammatical structural rules, Semantic structural rules, Context-free grammars, Semantic structure labels, Monte Carlo tree search algorithms, Neural networks

目 次

図 目 次

表 目 次

第1章　序論

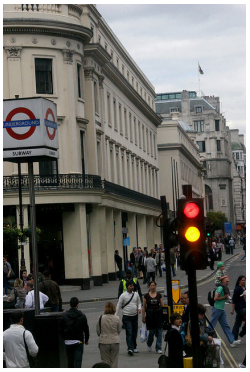
1.1　背景

人間は，日常的に様々な機会で情報を文章化し表現している．日常生活において他人と一緒に料理や掃除をするとき，自分が依頼したい作業内容を文章化し相手に伝えている．仕事中に複雑な手順を伴う作業を複数人に伝達する必要があるときは，作業手順ごとに行う動作を文章化することで，多くの人に複雑な作業を伝達することができる．人間の持つ文生成の能力は，自分以外の相手と円滑に生活したり，仕事を効率的に進めていくうえで必要不可欠な能力といえる．

機械に人間同様の文生成の能力を実装することができれば，我々の生活や仕事を大幅に効率化できると考えられる．例えば，日常生活における料理や掃除の作業について，普段の作業内容を自動で文章化し記録しておくことができれば相手への情報伝達の時間を短縮できる．また，仕事で複雑な手順を伴う作業について，事前に作業内容を撮影し，これを自動で文章化できればマニュアル作成の手間が省けると共に，無意識に行っている作業内容も自動記録されることでより正確に作業を伝達することができる．さらには，目の不自由な人が周囲の環境を把握するために，周辺環境について説明する文を自動生成することができれば，目の不自由な人がより安心して生活を送れるようになる．人間の伝えたい情報や知りたい情報を自動的に文章化する機能は，我々人間の生活や仕事を効率化することができたり，人間の能力拡張に有効活用できると考えられる．

日常の生活や仕事の効率化や，人間の能力拡張に向けて，これまで数多くの文生成に関するタスクが設定され，文生成手法の研究が為されてきた．例えば，画像キャプショニングタスクは画像中の情景や活動を説明する文を生成するタスクである．映像キャプショングタスクは映像中の情景や活動を説明する文を生成するタスクである．図 1.1 は，画像キャプショニングタスクのデータセット MSCOCO [1] 中のサンプルである．同じ画像に対し

A woman sitting on a bench and a woman standing waiting for the bus.
Two women waiting at a bench next to a street.
A woman sitting on a bench and a woman standing waiting for the bus.
A woman sitting on a bench in the middle of the city
A woman sitting on a bench and a woman standing behind the bench at a bus stop
A woman and another woman waiting at a stop.



Dozens of people walking around a metro area.
People walk down a busy city street, with traffic light.
A bunch of people crossing the street in a city.
People cross the street at the corner of luxury buildings.
People are walking across the street at an intersection.

図 1.1: 画像キャプショニングタスクのデータセット [1] 中のサンプル例．（画像は [1] 中の図を引用した．）

て複数のの正解文が付与されている．正解文が示す通り，画像中の同じ人物に対して異なる単語や語句を使って説明している．左のサンプルの画像中の “二人の女性” に対して “a woman and another woman” や “two women” という表現で説明している（図 1.2 中左）．右のサンプルの画像中の “人々” に対して “dozens of people” や “a bunch of people” という表現で説明している（図 1.2 中右）．また，それぞれの単語は適切な順番で並び，人間が理解可能な文を構築している．例えば，右のサンプルの正解文 “People are walking across the street at an intersection.” は，図 1.3 に示す通り，文頭から “主語（People）”→“述語（are walking）”→“副詞句（across the street at an intersection）” という順番で並び，人間が定めた文法構造規則に従った文を構築している．また，図 1.4 に示す通り，文頭から “動作主格（People）”→“動詞（are walking）”→“対象格（across the street）”→“場所格（at an intersection）” という順番で並び，人間が定めた意味構造的規則に従った文を構築している．上述の通り，画像に対して，正しくかつ柔軟に単語や語句を選択しながら，人間が定めた構造規則に従って単語や語句を並べることで文を生成している．映像キャプショニングタスクのデータセット ActivityNet Captions [8] においても，映像中のあるシーンについて正確にかつ柔軟に単語や語句を選択し，人間が定めた構造規則に従って単語や語句を並べて文を生成する必要がある．また，画像や映像などの視覚的情報を説明するタスクの他にも，長い文書を要約する文書要約タスク（図 1.5），質問文に対して適切な回答文を

図 1.2: 画像中の同じ対象について，多様な表現で説明している例.

生成する質問応答タスク（図 1.6），自然な会話文を生成する対話生成タスク（図 1.7）など様々な文生成に関するタスクがある．図 1.5 は，文章要約タスクデータセット CNN/Daily Mail [2] 中のサンプルである．文書要約タスクでは，文書中の重要な情報を説明する単語を正確にかつ柔軟に選択し，人間が定めた規則に従って選択した単語を適切な語順に並べ，正しい構造を持つ文を生成する必要がある．図 1.6 は，質問応答タスクのデータセット [3] 中のサンプルであり，図 1.7 は，対話生成タスクのデータセット [4] 中のサンプルである．質問応答タスクや対話生成タスクも同様に，回答文や会話文中で使用され得る単語を正確かつ柔軟に選択し，人間が定めた規則に従って選択した単語を適切な語順に並べ，正しい構造を持つ文を生成する必要がある．上述の通り，人間同様の文生成の能力を機械に実装するために，タスクや状況に応じて正確かつ柔軟に単語や語句を選択し，人間の定めた文法構造的規則や意味構造的規則に基づいた正しい構造の文を生成する必要がある．

しかしながら正確かつ柔軟な単語選択を実現しつつ，人間の定めた規則に基づいた正しい構造の文を生成することは難しい．例えば，テンプレートベースの手法は，予め定めたテンプレート（構造）に従いながら単語選択を行うため正しい構造の文を生成可能だが、柔軟な単語や語句の選択はできない．大量の文書データから尤もらしい単語順を学習し文を生

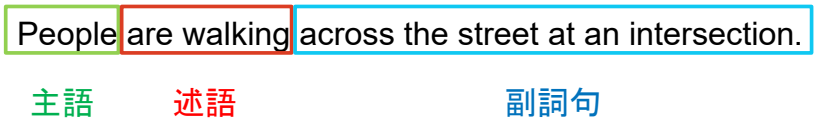


図 1.3: “People are walking across the street at an intersection.” という文に対する文法構造的規則.

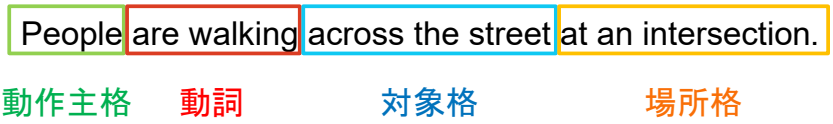


図 1.4: “People are walking across the street at an intersection.” という文に対する意味構造的規則.

成する ngram ベースの手法は，連続する単語順のみを考慮しており，離れた単語同士の関係を考慮できないことから適切な構造をもつ文の生成が困難である．近年急速な発展を遂げているニューラルネットワークベースの手法は，連続する単語順のみならず離れた単語同士の関係性を学習する Transformer と呼ばれる手法 [9] が提案されている．離れた単語同士の関係性を学習することで，尤もらしい文の構造規則を暗に学習することができ，適切な構造をもつ文の生成が可能である．しかしながら大量の文書データを学習したニューラルネットワークは，データ内で頻出する単語を使った文の確率を高くする傾向がある．頻出単語が偏って選択され，柔軟な単語や語句の選択は困難である．具体的に Transformer ベースの対話生成手法と画像キャプショニング手法の生成文例を紹介する．対話生成手法の ChatGPT [1]を使い，指定した単語を使って自由作文した時の生成結果例を図 1.8 に示す．“犬”，“パン”，“食べる” や “人”，“手紙”，“書く” の 3 単語を使った文生成結果例から，文の構造は正しいが，“ が好きだ” という言い回しが頻出しており，単語や語句が限定的であることが確認できる．また，画像キャプショニング手法 [5] を使った生成キャプション例を図 1.9 に示す．図 1.9 に示す通り，“a group of people” という言い回しが頻出しており，一番下のサンプルには “at a table with a” という部分に文法誤りも確認できる．上述の通り，正確かつ柔軟に単語や語句を選択しながら，正しい構造をもつ文の生成を実現することは困難である．

[1]https://openai.com/blog/chatgpt/

元の文章：
lagos, nigeria (cnn) a day after winning nigeria's presidency, muhammadu buhari told cnn's christiane amanpour that he plans to aggressively fight corruption that has long plagued nigeria and go after the root of the nation's unrest. buhari said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, he said his administration is confident it will be able to thwart criminals and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. buhari defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

正解要約文：
british no 1 defeated dominic thiem in miami open quarter finals . andy murray celebrated his 500th career win in the previous round . third seed will play the winner oftomas berdych and juan monaco in the semi finals of the atp masters 1000 event in key biscayne .

図 1.5: 文章要約タスクのデータセット [2] 中のサンプル例

テキスト：
The Norman dynasty had a major political, cultural and military impact on medieval Europe and even the Near East. The Normans were famed for their martial spirit and eventually for their Christian piety, becoming exponents of the Catholic orthodoxy into which they assimilated. They adopted the Gallo-Romance language of the Frankish land they settled, their dialect becoming known as Norman, Normaund or Norman French, an important literary language. The Duchy of Normandy, which they formed by treaty with the French crown, was a great fief of medieval France, and under Richard I of Normandy was forged into a cohesive and formidable principality in feudal tenure. The Normans are noted both for their culture, such as their unique Romanesque architecture and musical traditions, and for their significant military accomplishments and innovations. Norman adventurers founded the Kingdom of Sicily under Roger II after conquering southern Italy on the Saracens and Byzantines, and an expedition on behalf of their duke, William the Conqueror, led to the Norman conquest of England at the Battle of Hastings in 1066. Norman cultural and military influence spread from these new European centres to the Crusader states of the Near East, where their prince Bohemond I founded the Principality of Antioch in the Levant, to Scotland and Wales in Great Britain, to Ireland, and to the coasts of north Africa and the Canary Islands.

質問文：
Who was the duke in the battle of Hastings?

正解回答文：
William the Conqueror

質問文：
Who ruled the duchy of Normandy

正解回答文：
Richard

図 1.6: 質問応答タスクのデータセット [3] 中のサンプル例

1.2 研究目的

本論文では，人間同様の文生成の能力を機械に実装するために，正確かつ柔軟に単語や語句を選択し，人間が定めた構造的規則に基づき正しい構造を持つ文の生成を実現する方法を検討する．具体的には，文の構造的規則として，文法構造的規則と意味構造的規則の二種類の規則の活用方法についてそれぞれ検討し，各規則の文生成への有効性を検証する．文法構造的規則として文脈自由文法を採用し，文脈自由文法を適用して構築される構文木を生成する．適切な構文木の探索アルゴリズムとして広大な探索範囲を効率的に探索可能なモンテカルロ木探索アルゴリズムを使用する手法を提案する．タスクの設定としては，文の主要な要素となる重要単語のセットを入力として，それらの単語を使用しながら，正確か

A: I'm worried about something.
B: What's that?
A: Well, I have to drive to school for a meeting this morning, and I'm going to end up getting stuck in rush-hour traffic.
B: That's annoying, but nothing to worry about. Just breathe deeply when you feel yourself getting upset.
A: Ok, I'll try that.
B: Is there anything else bothering you?
A: Just one more thing. A school called me this morning to see if I could teach a few classes this weekend and I don't know what to do.
B: Do you have any other plans this weekend?
A: I'm supposed to work on a paper that'd due on Monday.
B: Try not to take on more than you can handle.
A: You're right. I probably should just work on my paper. Thanks!

図 1.7: 雑談対話タスクのデータセット [4] 中のサンプル例

図 1.8: ChatGPT の生成結果サンプル例

つ柔軟な単語や語句の選択と，正しい文法構造を備えた文の生成を目指す．次に，意味構造的規則として格構造ラベルを採用し，文生成時に格構造ラベルを条件として与え，与えられた格構造ラベルに対応する単語を選択しつつ，適切な格構造ラベルの順番を同時に推定する手法を提案する．タスクの設定としては，画像を入力として，画像に対応する正しくかつ柔軟な単語や語句を選択しつつ，画像を説明する文として正しい意味構造をもつ文の生成を目指す．画像の特徴抽出と，適切な格構造ラベルの順番推定と，格構造ラベルに対応する正しくかつ柔軟な単語推定とを同時に行う end-to-end 構造の Neural Network を提案する．さらに，Neural Network が適切な文の意味構造を学習するために，離れた単語同士の関係性や格構造ラベル同士の関係性を学習可能な Transformer ベースの画像キャプショニング手法 [5] をベースラインとして使用する．各手法の実験では，文法構造的規則や

図 1.9: 画像キャプショニング手法の [5] を使った生成キャプション例

意味構造的規則の文生成への有効性を検証する．

1.3 本論文の構成

本論文は，下記の 5 つの章から構成される．本章は序論であり，本論文の研究背景や従来技術の概要，研究の目的について述べた．第 2 章では，本論文で取り扱う研究テーマに関連する先行研究について述べる．具体的には，文の文法構造的規則と意味構造的規則について説明した後，既存の文生成手法の研究について概観する．さらに，画像キャプショニングタスクに限定した近年の文生成手法について説明する．また，提案手法の中で採用したモンテカルロ木探索，Latent Dirichlet Allocation，ngram モデルと Kneser-Ney スムージングについて説明する．第 3 章では，文の文法構造的規則に基づく文生成手法を検討する．第 4 章では，文の意味構造的規則に基づく文生成手法を検討する．第 5 章は本論文の結論として，各章で得られた知見をまとめ，今後の課題と展望について述べる．

第2章　関連手法

本論文で取り扱う研究テーマに関連する手法について述べる．具体的には，2.1 節で文の文法構造的規則と意味構造的規則について説明し，2.2 節で提案手法の中で採用したモンテカルロ木探索，2.3 節で Latent Dirichlet Allocation，2.4 節で ngram モデルと Kneser-Ney スムージングについて説明する．

2.1　文の構造的規則

本節では，文の文法構造的規則と意味構造的規則の二種類の構造規則について紹介する．2.1.1 で文の文法構造的規則として文脈自由文法，2.1.2 で文の意味構造的規則として述語項構造について述べる．また，本節は『自然言語処理 [10]』と『確率的言語モデル [11]』の構造的規則に関する説明を基に述べる．

2.1.1　文法構造的規則

文法とは，文の構造に関する法則を体系的に記述したものである．形式言語理論によれば，文法規則は 0 型文法，文脈依存文法（1 型文法），文脈自由文法（2 型文法），正規文法（3 型文法）の四つのクラスに分類され，この順に文法記述に関する制限が強い [12]．正規文法には，非常に効率の良い解析アルゴリズムが存在するが，自然言語がもつ階層構造や長距離依存性を記述できない．文脈依存文法は，さまざまな言語現象を記述するのに十分な記述能力をもっているが，効率的な解析アルゴリズムが見つかっていない．このため，自然言語の文法記述には，階層構造や長距離依存性を記述することができ，かつ，効率の良い解析アルゴリズムが存在する文脈自由文法を用いることが多い．本論文における文法構造的としても文脈自由文法を採用する．

文脈自由文法

文脈自由文法は，以下の 4 項組 $G = \{V_N, V_T, P, S\}$ により定義される．

1. V_N：非終端記号 (Nonterminal Symbol) の有限集合.
2. V_T：終端記号 (Terminal Symbol) の有限集合.
3. P：生成規則 (Production) の有限集合.
   生成規則は $A \to \alpha$ の形をしている．ここで $A \in V_N$, $\alpha \in (V_N \cup V_T)$ である．
4. $S\,(\in V_N)$：開始記号 (Start Symbol)

非終端記号 V_N は，「名詞句」や「動詞句」などの抽象的な文法カテゴリーを表し，終端記号 V_T は「私」や「投げる」などの個々の単語を表す．特に，非終端記号 V_N 内において，終端記号 V_T を生成するものを特別に前終端記号とも呼び， これは品詞（Parts Of Speech）と同等である．生成規則 P（句構造規則）は文法カテゴリーあるいは単語間の階層的な関係を記述するものであり，A を非終端記号，α を非終端記号あるいは終端記号から成る記号列として，$A \to \alpha$ の形をしている．生成規則 $A \to \alpha$ は，左辺にある A という記号を右辺にある α という記号列に書き換えることを意味しており，書換え規則 (rewriting rule) と呼ばれることもある．

生成規則 r の適用により記号列 α が記号列 β に書き換えられるとき，

$$\alpha \overset{r}{\Rightarrow} \beta \tag{2.1}$$

と書く．また，複数の生成規則を順次適用することにより α が β に書き換えられるときは

$$\alpha \overset{*}{\Rightarrow} \beta \tag{2.2}$$

と書く．このとき, α は β を導出する, あるいは β は α から導出されるという．生成規則適用の際に最も左側の非終端記号を書き換えるような導出を最左導出 (leftmost derivation) と呼ぶ．逆に, 最も右側の非終端記号を書き換えるような導出は最右導出 (rightmost derivation) と呼ぶ.

開始記号 $S\,(S \in V_N)$ は特殊な非終端記号であり，この記号から書き換えが始まる．開始記号 S から導出される記号列を文形式 (sentential form) と呼ぶ．文脈自由文法によって定義される言語は，開始記号 S から導出される終端記号列の集合である．

文脈自由文法の生成規則は $A \to \alpha$ の形をしている．これは左辺の非終端記号 A が右辺の文法記号列 α に書き換えられることを意味するが，この生成規則に A が α に書き換えられる条件付き確率 $P(\alpha|A)$ を与えたものを確率文脈自由文法 (Probabilistic Context-Free Grammar，以下 PCFG) と呼ぶ．左辺に非終端記号 A をもつ生成規則すべての条件付き確率 $P(\alpha|A)$ を足し合わせると 1 にならなければならない (式 (2.3))．

$$\sum_{\alpha} p(\alpha|A) = 1 \tag{2.3}$$

確率文脈自由文法の例を表 2.1 に示す．

表 2.1: 確率文脈自由文法の例

S	→	NP	VP	1.0
NP	→	DT	NN	0.3
NP	→	DT	NNS	0.7
VP	→	VBZ	NP	0.6
VP	→	VB		0.4
DT	→	a		1.0
NN	→	dog		0.6
NN	→	bread		0.4
NNS	→	dogs		1.0
VBZ	→	eats		1.0
VB	→	run		1.0

本論文では Penn Treebank[13] で採用されている品詞を用いる．Penn Treebank における品詞を表 2.2 に示す．

文 S が m 個の文法規則 $r_1, ..., r_m$ を適用することにより開始記号 S_0 から導出されるとする．ここで α_i は規則 r_i を適用した後の文形式である．

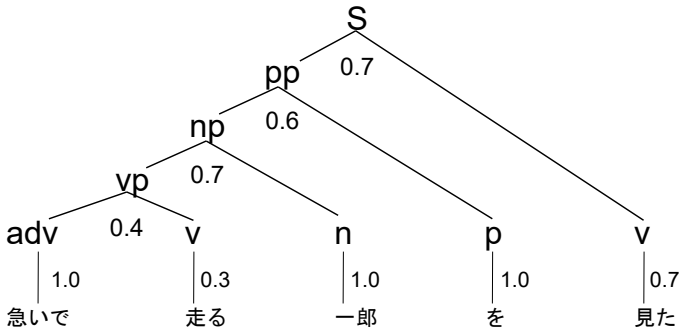
$$S_0 \overset{r_1}{\Longrightarrow} \alpha_1 \overset{r_2}{\Longrightarrow} \alpha_2 \cdots \overset{r_m}{\Longrightarrow} \alpha_m = S \tag{2.4}$$

生成規則 r の確率を $P(r)$ で表すことにすると，文脈自由とは，その定義より，導出において $p(r_i|r_1^{(i-1)}) = p(r_i)$ が成り立つことである．

表 2.2: Penn Treebank の品詞

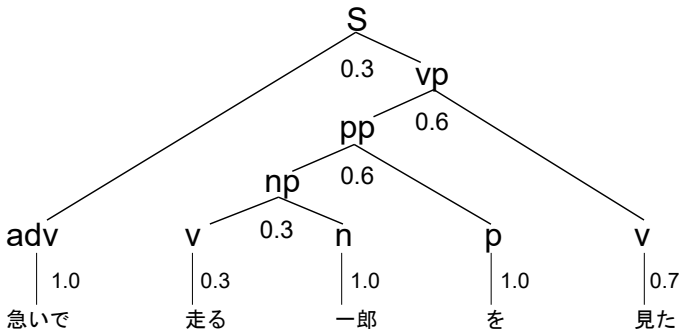
CC	Coordinating conjunction	PPS	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential there	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition	SYM	Symbol
JJ	Adjective	TO	infinitival to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund/present pple
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd ps. sg. present
NNP	Proper noun, singular	VBZ	Verb, 3rd ps. sg. present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WPS	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

確率文脈自由文法では，構文木の確率を計算できるので，文の構造的曖昧性を定量的に扱うことができる．例えば，「急いで走る一郎を見た」という文からは図 2.1, 2.2 に示す二つの構文木が得られる．左側の構文木は「一郎が走っているところを（話者が）見ていた」という解釈，右側の構文木は「一郎が走っているところを（話者が）急いで見た」という解釈を表している．それぞれの構文木の確率は 0.025 及び 0.007 であるので，左側の構文木のゆう度が右側の構文木より約 3 倍大きい．



p(t1) = 1.0 x 0.3 x 0.4 x 1.0 x 0.7 x 1.0 x 0.6 x 0.7 x 0.7 = 0.025

図 2.1: 「急いで走る一郎を見た」という文に対する構文木の例 1.



p(t2) = 0.3 x 1.0 x 0.3 x 1.0 x 0.6 x 0.7 x 0.6 x 1.0 x 0.3 = 0.007

図 2.2: 「急いで走る一郎を見た」という文に対する構文木の例 2.

2.1.2 意味構造的規則

文の意味とは，ある表現形式によって示される内容であって，自然言語における意味的なまとまりとして，語，句，節，文，文の集合などのいくつかのレベルが考えられる．従来の自然言語理解の研究では，これらいくつかのレベルの内容を適切に表現できる形式が模索され，論理式，意味ネットワーク，概念依存構造などの意味構造が研究されてきた．語は，意味構造の基本的な構成要素として位置付けることができる．語の意味は，語義として定義されるが，その設定にあたっては，他の語との関係や他の言語との対応関係といった言語学的な観点や，機械翻訳などの工学的な観点から研究されている．意味構造は，基本的に，語の意味と，語や句や節などのまとまりの間の意味関係によって表現される．単純なモデルでは，文の意味構造は，文に存在する単語の意味と単語間の意味関係の 2 種類

によって表現される．意味解析とは，与えられた文から，妥当な語の意味と意味関係を明らかにし，意味構造を生成することといえる．カナ漢字変換，質問応答，機械翻訳，音声理解など自然言語処理を用いた応用システムにおいては，意味解析が多かれ少なかれ必要であり，既に，実用システムの中に採り入れられている技術も多い．

古典的な自然言語処理システムにおいては，構文解析後の構文構造を対象として意味解析をする場合が多い．したがって，意味構造と構文構造には密接な関係がある．構文構造と意味構造の関係で考えると，意味構造は，構文構造をベースとして意味情報が反映されているもの，意味構造をベースとして構文構造が反映されているもの，構文構造に依存しない概念レベルのものの 3 種類に大別できる．意味構造を表現する枠組みとしては，入力となる文や構文構造，表現される意味の内容，意味構造を用いて行う処理など，システムの構成に応じて，木構造，論理式，ネットワーク，フレーム表現などが用いられる [14]．意味構造の表現形式は，知識工学における知識表現としても研究されてきた．本論文では，代表的な意味構造として，格構造を採用する．

格構造

文を構成する要素（語や句）は，文の中で一定の役割を担っている．文の構成要素を核要素，核要素が述語に対して果たす役割（機能）を格（case）という．格の種類として，構文的なもの（表層格）と意味的なもの（深層格）に分けられる．意味解析に関係の深い構造は，深層格を用いた格構造である．深層格の概念を中心とする文法理論は，Fillmore によって唱えられ，格文法（case grammar）と呼ばれる．Fillmore が初期に設定した深層格を表 2.3 に示す．

格構造を用いると，次の四つの文は，図 2.3 のように表現される．

1. John opened the door.
2. The key opened the door.
3. The door opened.
4. John opened the door with the key.

格構造によると，“The door was opened by John.” という例文（１）の受動態が，基本的に例文（１）と同一の構造で表現できるだけでなく，(1) (4) の例文に存在する意味的共通性も，その構造のなかに自然に表現される．また，格による意味の表現は，知識工学にお

表 2.3: Fillmore の深層格

格の名称	説明
動作主格	ある動作を引き起こす者の役割
経験者格	ある心理事象を体験する者の役割
道具格	ある出来事の直接原因となったり，ある心理事象と関係して反応を起こさせる刺激となる役割
対象格	移動する対象物や変化する対象物，あるいは，判断，創造のような心理事象内容を表す役割
源泉格	対象物の移動における起点，および状態変化と形状変化における最初の状態や形状を表す役割
目標格	対象物の移動における終点，および状態変化や形状変化における最終的な状態，結果を表す役割
場所格	ある出来事が起こる場所及び位置を表す役割
時間格	ある出来事が起こる時間を表す役割

けるフレームや意味ネットワークの表現手法と共通する部分が多く，実際の自然言語処理システムにおいて多く用いられている．述語の語彙項目として格フレームと選択条件を記述して，意味的曖昧性解消や構文解析の曖昧性解消に用いられている．しかしながら，格の設定の基準や原理が明確ではないので，実際の設定は容易ではない．また，実用的観点からは，選択できないほど細分化された格は，いたずらに曖昧性を増加し，システムを複雑にするので好ましくない．機械翻訳システム PIVOT では，実際的な翻訳の実現を目指し，格関係を拡張し，中間言語における意味的関係として関係概念を定義している [15]．表 2.4 に，機械翻訳システム PIVOT の上位レベルの関係概念の一覧を示す．

表 2.4: PIVOT における関係概念

略号	名称	説明	例文
OBJ	object	述語の対象	John hit **the desk** with his fist.
AGT	agent	動作主	**He** gave me his books.

CAU	causer	原因者，使役者	**She** let me leave.
EXP	experiencer	経験者，知覚の主体	They suspect that he is a murder.
INS	instrument	道具	The computer solved the problem.
MEA	means, method	手段，方法	She persuaded him to stay with a kiss.
BEN	beneficiary	受益者	They are working for me.
LOC	location	場所	I saw it under the table.
TIM	time	時間	We have not seen land **in 20 days**.
SOR	source	源泉，起点	John left **town**.
TAR	target	目標，着点	John spent all his money **on clothes**.
PRT	participant	随伴者，随伴物	They married Taro **to Hanako**.
CAP	capacity	役割，機能	He attended meeting **as a leader**.
FCS	focus	焦点	He wrote a book **about Japan**.
MAT	material	材料	Their house is built **of wood**.
ELM	element	要素	Japan consists **of four islands**.
POS	possessor	所有者	This is **my** house.
POF	part of	全体の中の部分	**The front** of the car was destroyed.
NUM	number	数量	It is five meter.
NAM	name	名称	The yamanote line is ...
NMOD	noun modifier	名詞連続	The magnetic discs are ...
ATT	attribute	属性	What is your shoe size?
VAL	value	属性値	The length of the axis is 30 meter.
MOD	modifier	副詞的修飾	People struggle to live better.
QUNT	quantity	量	She bought two bedding sets.
EQ	equality	同一性	That must be a whale.
APP	apposition	同格	Memory devices, such as CD ...
REF	reference	参照における限定	Which man did John say saw him?
CPQT	quantity of comparative	比較における量	He is 5 cm taller than I.
MDQT	quantity of modification	変化における量	He walks 5 km.

RLBS	base of relativity	相対概念における基準	He looks at the outside of the window.
DCMP	dummy case for comparative	比較の相手	He is taller than I.
MODS	sentential modifier	文修飾句	He is, in a word, an idiot.
PAR	parallel	等位接続	John and Mary heard the news.
GOA	goal	目標，目的	They stopped in order to rest.
CON	connection	時間的に連続な出来事	Returning to my office, I slept.
REA	reason	理由	The rain made us seek shelter.
CAS	case	事象の発生	In case of the accident, call me.
CASA	case assumption	仮定	If he comes, call me.

2.2 モンテカルロ木探索

モンテカルロ木探索のアルゴリズムを説明する．モンテカルロ木探索（MCTS）は，ランダムシミュレーションと木構造に対する正確な探索を組み合わせたアルゴリズムである．コンピュータ囲碁における MCTS の成功により，ゲームに対する課題だけではなく，状態と行動の対のデータを有する様々なドメインに適用され，シミュレーションによってその出力を予測することに用いられている．

2.2.1 木構造の用語

まず，MCTS の説明に用いる木構造の用語について述べる．

- **ノード**：一つの要素．
- **エッジ**：ノード間を結ぶもの．

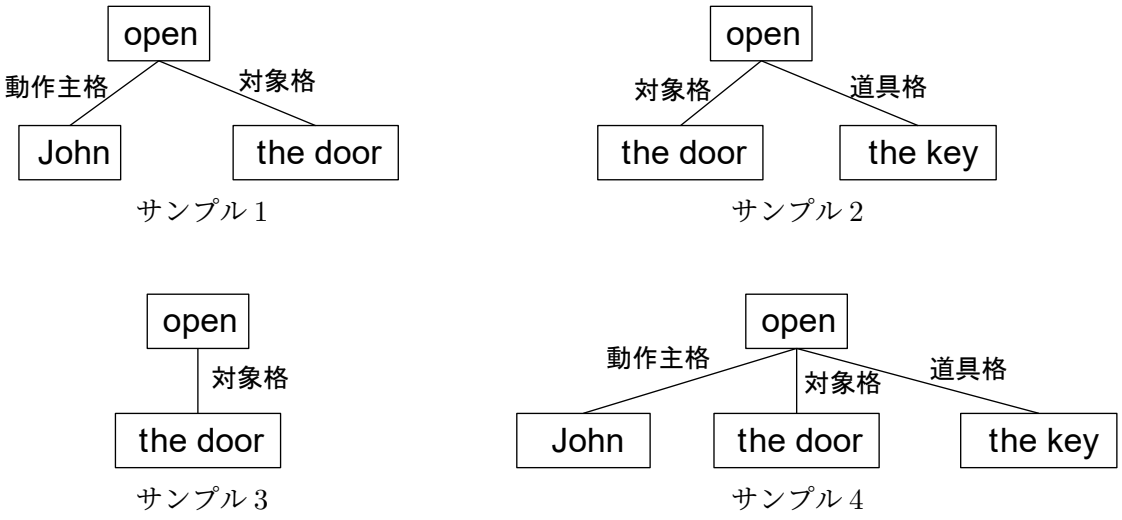


図 2.3: 格構造を用いた文の表現例.

- **子ノード**：各ノードは 0 個以上の子ノードを持つ.
- **親ノード**：子ノードから見た，子ノードを持つノード.
- **子孫ノード**：あるノードの，子ノードやそれらから先の子ノード全てのいずれか.
- **先祖ノード**：あるノードの，親ノードやそこから先の親ノードの全てのいずれか.
- **根ノード**：親ノードを持たないノード.
- **葉ノード**：子ノードを持たないノード.
- **内部ノード**：子ノードを持つノード．すなわち葉ノード以外のノード.
- **部分木**：木構造の一部．それ自信も完全な木構造となっている.
- **高さ**：あるノードについて，そのノードの子孫である葉ノードへのエッジ数の最大値.
  すなわち根ノードの高さはその木構造の高さ.
- **深さ**：あるノードについて，そのノードからルートノードまでのエッジ数.

2.2.2 基本アルゴリズム

MCTS の基本アルゴリズムの概要を図 2.4 に示す[1].

MCTS は,「選択 (Selection)」,「拡張 (Expansion)」,「シミュレーション (Simulation)」「逆伝搬（Backpropagation)」4 つの基本処理から構成されている.

[1]図 2.4 は文献 [16] より引用.

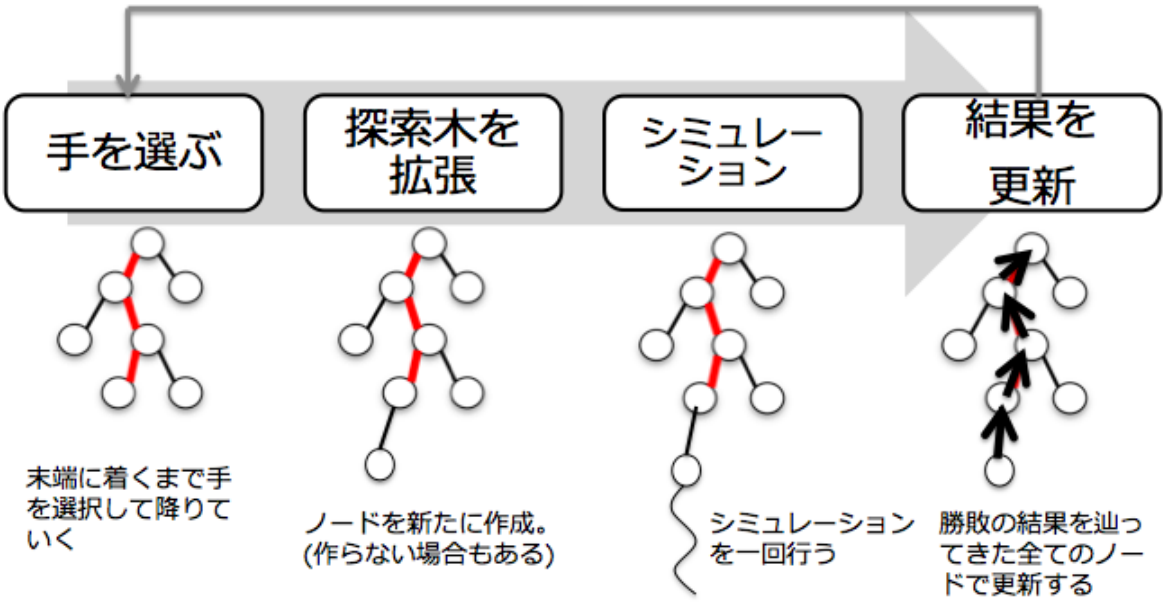


図 2.4: MCTS アルゴリズムの概要

Step 1:（選択）： 探索木の末端に辿り着くまで手を選択していく．

Step 2:（拡張）： ノードを新たに作成し，探索木を拡張する．

Step 3:（シミュレーション）： 新たに作成したノードから 1 回シミュレーション（プレイアウト）を行う．

Step 4:（逆伝搬）： シミュレーションによって得られた報酬をその時のルートノードまでの全てのノードへ逆伝搬し，それらノードの評価値を更新する．

2.2.3 UCB1 値

Multi-Armed Bandit 問題に対処するアルゴリズムとして，Auer ら [17] によって UCB1 というアルゴリズムが提案された．そこでは，スロットマシンを選択する指標として，従来の報酬の平均から UCB1 値が代わりに用いられた．UCB（Upper Confidence Bounds）1 値は，勝率の項、および探索が不十分なノードに対して選択の可能性を考慮した項から構成される（式 2.5）．

$$v_i + C\sqrt{\frac{logN}{n}} \tag{2.5}$$

v_i はそのノードの勝率，C は調整係数，N は全試行回数，n はそのノードを選択した回

数を示す．UCB1 値における第 1 項が「知識の適用（exploitation）」を，第 2 項が「探査（exploration）」を考慮している．それによりバランスをとった探索が実行される．

2.3 Latent Dirichlet Allocation

大量の文書データから話題になっているトピックを知るための潜在的意味解析手法として，Latent Dirichlet Allocation(以下，LDA) を紹介する．LDA における文書の扱いは Bag Of Words(BOW) であり，単語の頻度のみを考慮するため，人手の作業を必要としない．また，文書特有の特徴は考えないため，文書データだけでなく，画像処理，音楽情報処理など様々な分野で応用されている．

また，LDA は文書のための確率モデルであるが，確率モデルの研究は以下の 2 つの部分から成り立っている．

1. モデリング：文書をどのようにモデリングするか
2. 学習アルゴリズム：あるモデルをどのように学習するか

2.3.1 モデリング

LDA における，文書のモデリング方法について述べる．LDA とは，一つの文書が複数のトピックを持つと仮定したモデルである．LDA におけるグラフィカルモデルを図 2.5 に示す．グラフィカルモデルとは，確率変数やパラメータを頂点とし，それらの依存関係を有向グラフで表現したものである．

網掛けの頂点は観測変数を示し，それ以外の頂点は潜在変数や未知パラメータを示す．矩形部分は，その隅にしめされた回数だけサンプリングが繰り返されることを表わす．図 2.5 は，以下のようなことを示している．各文書は，トピック分布 θ を持ち，文書上の各単語の位置について，θ に従ってまずトピック z が選ばれ，そのトピック z に対応する単語分布 ϕ に従って，その位置の単語 w が生成される．K はトピック数，D は文書数，N_d は文書 d 上の単語の出現回数を表わしており，トピック分布 θ は各文書ごとに生成され，単語分布 ϕ は各トピックごとに生成され，単語 w とその単語のトピックを表わす z は各単語の出現する位置ごとに生成される．また，α と β はハイパーパラメータであり，それぞれ，パラ

メータ θ が従うディリクレ分布のパラメータ，パラメータ ϕ が従うディリクレ分布のパラメータを示す．これらの変数の中で，実際に観測される変数は文書上に現れている単語 w であり，実用的には，この観測変数を用いて潜在変数の推定を行っている．LDA における文書の生成過程は，以下のような手順である．

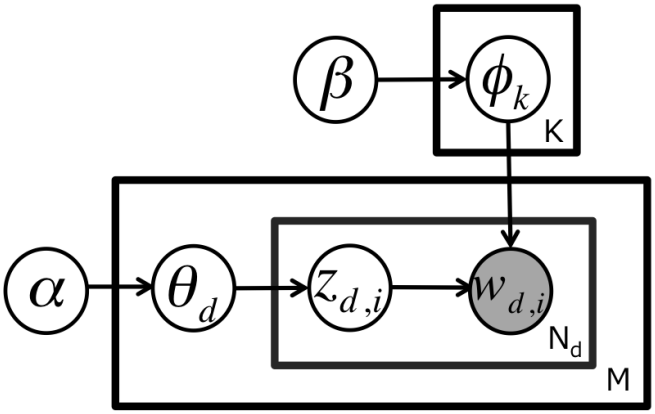


図 2.5: LDA のグラフィカルモデル

1. 各トピック $k = 1,\cdots,K$ について：
   (a) ディリクレ分布に従って単語分布 ϕ_k を生成
   $\phi_k \sim Dir(\beta)$
2. 各文書 $d = 1,\cdots,D$ について：
   (a) ディリクレ分布に従ってトピック分布 θ_d を生成
   $\theta_d \sim Dir(\alpha)$
   (b) 文書 d における各単語 $n = 1,\cdots,N_d$ について：
      i. 多項分布に従ってトピックを生成
      $z_{dn} \sim Multi(\theta_d)$
      ii. 多項分布に従って単語を生成
      $w_{dn} \sim Multi(\phi_{z_{dn}})$

なお，ϕ_k はトピック k の単語分布，θ_d は文書 d のトピック分布，z_{dn} は文書 d の n 番目の単語の潜在的トピック，w_{dn} は文書 d の n 番目の単語を表わし，$Dir(\cdot)$ はディリクレ分布，$Multi(\cdot)$ は多項分布を表わす．

2.3.2 学習アルゴリズム

LDA の学習アルゴリズムの主要な手法の一つとして，サンプリング近似法がある．本研究では，サンプリング近似法の一種である，Collapsed Gibbs Sampling 法を用いた．これは，パラメータ θ，ϕ を積分消去し，ある単語のトピック zd, i をハイパーパラメータ α，β と，単語の頻度情報から直接的にサンプリングする手法である．そのため，Gibbs Sampling 法と比較すると，実装が容易で，収束時間も早いことが利点である．

LDA における全ての潜在変数の結合分布を，グラフィカルモデルの依存関係に基づき，ベイズの定理と条件付き独立を用いて展開する．式 (2.6) で示す．

$$p(z, w, \theta, \phi|\alpha, \beta) = p(w|z, \phi)p(z|\theta)p(\theta|\alpha)p(\phi|\beta). \tag{2.6}$$

式 (3.3.3) は，Collapsed Gibbs Sampling 法によって求めた，潜在的トピック $z_{d,i}$ のサンプリング更新式である．

$$\begin{aligned}
p(z_{d,i} = k|w_{d,i} = v, w^{\backslash d,i}, z^{\backslash d,i}, \alpha, \beta) &\propto p(z_{d,i} = k, w_{d,i} = v, w^{\backslash d,i}, z^{\backslash d,i}|\alpha, \beta) \\
&= \int p(z_{d,i} = k, w_{d,i} = v, w^{\backslash d,i}, z^{\backslash d,i}, \theta, \phi|\alpha, \beta)d\theta d\phi \\
&\propto \int p(w_{d,i} = v|z_{d,i} = k, \phi_k)p(\phi_k|w^{\backslash d,i}, z^{\backslash d,i}, \beta)d\phi \\
&\quad \times \int p(z_{d,i} = k|\theta_d)p(\theta_d|z^{\backslash d,i}, \alpha)d\theta \\
&= \frac{(n_{k,v}^{\backslash d,i} + \beta_v)}{(n_{k,\cdot}^{\backslash d,i} + \sum_v \beta_v)} \frac{(n_{d,k}^{\backslash d,i} + \alpha_k)}{(n_d^{\backslash d,i} + \sum_k \alpha_k)}.
\end{aligned} \tag{2.7}$$

ここで，$z^{\backslash d,i}$ は潜在的トピック集合 $\mathbf{z}$ から $z_{d,i}$ を取り除いた潜在的トピック集合を表す．$n_{k,v}^{\backslash d,i}$ は潜在的トピック k に割り当てられた単語 v の頻度回数を，$n_{d,k}^{\backslash d,i}$ は文書 d に割り当てられた潜在的トピック k の頻度回数を，$n_{k,\cdot}^{\backslash d,i}$ はコーパス中で割り当てられた潜在的トピック k の頻度回数を, そして，$n_d^{\backslash d,i}$ は文書 d で生成される単語の頻度回数を表し，いずれの変数においても文書 d における単語位置 i の頻度回数が除かれる．

式 (2.8)，式 (2.9) は，θ 並びに ϕ の確率算出式であり，式 (2.10)，式 (2.11) は，それぞれ，Collapsed Gibbs Sampler における θ と ϕ に関する更新式である．

$$p(\theta_d|z_d,\alpha) = \frac{\Gamma(\sum_k n_{d,k}+\alpha_k)}{\prod_k \Gamma(n_{d,k}+\alpha_k)} \prod_{k=1}^{K} \theta_{d,k}^{n_{d,k}+\alpha_k-1}, \tag{2.8}$$

$$p(\phi_k|w,z,\beta) = \frac{\Gamma(\sum_v n_{k,v}+\beta_v)}{\prod_v \Gamma(n_{k,v}+\beta_v)} \prod_{v=1}^{V} \phi_{k,v}^{n_{k,v}+\beta_v-1}, \tag{2.9}$$

$$\hat{\theta}_{d,k} = \frac{n_{d,k}^{\backslash d,i}+\alpha_k}{n_d^{\backslash d,i}+K\alpha_k}, \tag{2.10}$$

$$\hat{\phi}_{k,v} = \frac{n_{k,v}^{\backslash d,i}+\beta_v}{n_{k,\cdot}^{\backslash d,i}+V\beta_v}. \tag{2.11}$$

ここで，θ_d は, 文書 d の潜在的トピック分布を, $z_{d,i}$ は, 文書 d において出現する n 番目の単語に対応する潜在的トピックを, $w_{d,i}$ は, 文書 d において出現する i 番目の単語を, ϕ_k は，潜在的トピック k $(z_{d,i}=k)$ に対する単語分布を表す．また，α,β は *Dirichlet* 分布のハイパーパラメータである．

2.4 ngram モデル

言語モデルの基本的な役割は，与えられた単語列 $W=w_1^I=w_1...w_I$ に対し，その生成確率 $p(W)$ を計算することである．

$$p(W)=\prod_{i=1}^{I} p(w_i|w_1^{i-1}) \tag{2.12}$$

式 (2.12) における条件付き確率 $p(w_i|w_1^{i-1})$ を求めることができれば，単語列全体の確率を計算することができる．しかし，すべての単語の組み合わせに対し，$p(w_i|w_1^{i-1})$ を求めることは現実的に不可能である．

ngram モデルは，単語の生起が直前の $(n-1)$ 単語にのみ依存すると考えられた言語モデルである．つまり，式 (2.12) における $p(w_i|w_1^{i-1})$ は，以下式 (2.13) で計算できる．

$$p(w_i|w_1^{i-1})=p(w_i|w_{i-n+1}^{i-1}) \tag{2.13}$$

また，$n = 1, 2, 3$ のとき，それぞれ $unigram$，$bigram$，$trigram$ と呼ばれる．例として，$bigram$ のときの単語列 W の生起確率は式 (2.14) のように計算できる．

$$p(W) = \prod_{i=1}^{I} p(w_i|w_{i-1}) \tag{2.14}$$

また，文頭，文末の扱いについては，文頭には文頭を表す特別な記号を，文末には文末を表す特別な記号を，それぞれ $n-1$ 個ずつ単語列に付け足す．

また，ngram モデルの確率の推定方法であるが，単語列 w_1^i が学習データ中に出現する回数を $C(w_1^i)$ と表すことにすると，ngram の確率は以下式 (2.15) のように求められる．

$$p(w_i|w_{i-n+1}^{n-1}) = \frac{C(w_{i-n+1}^{i})}{C(w_{i-n+1}i-1)} \tag{2.15}$$

2.4.1 Kneser-Ney スムージング

式 (2.15) において，例えば，$C(w_{i-n+1}^{i})$ が 0 のとき，$p(w_i|w_{i-n+1}^{n-1})$ が 0 となり，これによって文全体の確率も 0 となってしまう．このような，コーパス中に出現しない単語や単語列に対しどのように対応するかという問題をゼロ頻度問題と呼び，ongram モデルにおける大きな問題となっている．適切な推定値を得るためには確率値のスムージングを行う必要がある．スムージング方法は様々存在するが，ここでは，実験的に最も性能が良いとされた Kneser-Ney スムージング [18] について説明する．

Kneser-Ney スムージングは，直前の単語の種類数を重視するスムージング方法である．以下に，trigram の場合の式を示す (式 2.16)．

$$p(z|xy) = \frac{(c(xyz) - d)}{c(xy*)} + \frac{d|xy*|}{c(xy*)}p(z|y) \tag{2.16}$$

c はカウント数, d は種類数である．$(d|xy*|)$ が直前の単語の種類数を示す．例えば，“San Fransisco” が頻出のコーパスにおいて，“Fransisco” の頻度は高いが，ほとんどの場合それは”San” の後に続く．しかし，“Fransisco” の unigram 確率が高いため，“Tokyo Fransisco” なども高い確率となってしまう．この問題を回避するため，Kneser-Ney スムージングでは直前の単語の種類数を重視している．

第3章　文法構造的規則に基づく文生成

第 1 章で述べた通り，正確かつ柔軟に単語を選択し，人間が定めた構造的規則に基づき正しい構造を持つ文の生成手法を検討するために，文の文法構造的規則に基づく文生成手法を検討し，文法構造的規則の文生成への有効性について検証する．本章では，文法構造的規則として文脈自由文法を採用し，文脈自由文法を適用して構築される構文木を生成することで文法構造的規則に基づく文を生成する．タスクの設定としては，文の主要な要素となる重要単語のセットを入力として，それらの単語を使用しながら，正確かつ柔軟な単語や語句の選択と，正しい文法構造を備えた文の生成を目指す．

3.1　はじめに

本節ではまず，文法構造規則として文脈自由文法を採用し構文木を構築する際の課題を説明する．次に，前述した課題を解決するため，我々が提案する構文木の評価方法と，モンテカルロ木探索アルゴリズムを使用した文脈自由文法に基づく文生成手法について述べる．さらに，モンテカルロ木探索アルゴリズムで効率的に探索シミュレーションを行うために，探索方針の設定方法と探索範囲の絞り込み方法について説明する．

文脈自由文法を使用して構文木を構築する方法について説明する．2 章で説明した通り，表 2.1 で記載した複数の文法から適用可能な文法を次々に選択することで構文木を構築し，文法構造規則に従った文を生成する．具体的には，文の開始記号 S から適用可能な文法は，S $\rightarrow$ NP VP のみであり，これを選択する．次に，非終端記号 NP から適用可能な文法は，NP $\rightarrow$ DT NN，もしくは，NP $\rightarrow$ DT NNS であり，これら 2 種類の文法規則から NP $\rightarrow$ DT NN を選択する．さらに，非終端記号 NN から選択可能な文法は，NN $\rightarrow$ dog，もしくは，NN $\rightarrow$ bread であり，これら 2 種類から 1 つ選択する．このように，文の開始記号 S から選択可能な文法を次々と選択することで，図 3.1 のような 12 種類の構文木を構築するこ

とができ，状況に応じて適切な構文木を選択すればよい．しかしながら，状況に応じた適切な構文木を機械的に自動選択するためには 2 つの問題を解決する必要がある．1 つ目に，状況に応じた適切な構文木であることを機械的に自動判定することはできない．人間が適切な構文木であると判定する際と同等な構文木の評価指標を用意する必要がある．2 つ目に，文脈自由文法の文法種類数が増えるほど構築され得る構文木は増大し探索範囲が膨大になる．例えば，非終端記号 NNS から生成可能な終端記号が 1 種類から 100 種類の単語に増えると，構築され得る構文木種類数は 1,506 個に増大する．文脈自由文法を使用して構文木を構築する際，文法規則が増えるほど構築され得る構文木は増大し，膨大な種類の構文木から状況に応じた適切な構文木を探索することが必要である．

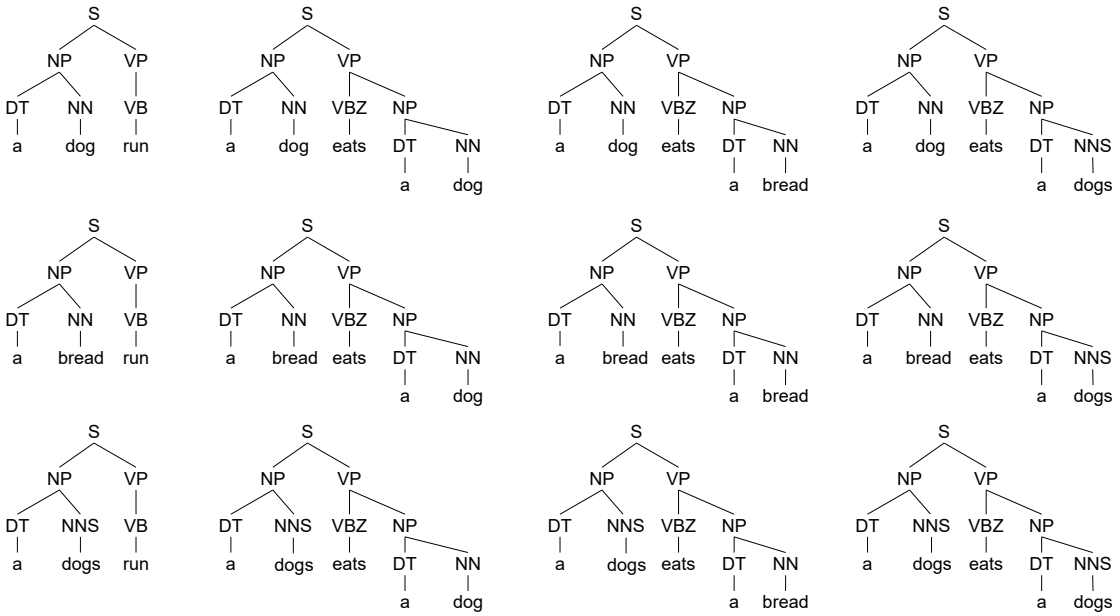


図 3.1: 表 2.2 に示す文脈自由文法から構築され得る 12 種の構文木.

本論文では，1 つ目の課題に対し，適切な構文木を判定するための複数の構文木の評価値を提案する．本章ではタスクの設定として，状況を説明する際に主要な要素となる複数の単語（以降，“Situational Input” とする．例えば，{dog,eat,bread} である．）を入力情報として与え，これらの単語に基づき適切な文が生成されているか評価する．具体的には，生成した構文木を評価する視点として，文法構造の適切さ，単語の繋がりの正しさ，生成内容の適切さの 3 つの点から評価する．文法構造の適切さを評価するために，入力情報である “Situational Input” の各単語の品詞がすべて含まれているか判定する．単語の繋がりの

正しさを評価するために，ngram に基づく評価値を採用し，生成内容の適切さを評価するために，“Situational Input” との意味の類似度に基づく評価値を採用する．2 つ目の課題に対し，構築され得る構文木種類が膨大であっても効率的に適切な構文木を探索するためにモンテカルロ木探索アルゴリズムを使用する．モンテカルロ木探索とは，2 章で説明した通り，UCB1 値に基づいて「知識の適用（exploitation）」と，「探査（exploration）」のバランスを保ちながら効率的に探索するアルゴリズムである．前述した，すべての構文木の評価値が収束するように探索を進めるため，構文木の評価値に基づく UCB1 値の設定方法を提案する．さらなる工夫として，提案手法ではモンテカルロ木探索を効率的に行うために，探索方針の設定と探索範囲の絞り込みを行う．探索方針の設定方法として，文の主要な要素から優先的に探索する方法を提案する．具体的には，文の述語や主語が文の主要な要素となるという考えのもと，述語を優先的に探索した後，述語よりも左の文法規則を適用することで優先的に主語を探索し，最後に述語よりも右の文法規則を適用する．探索範囲の絞り込み方法として，文の主要な要素となる重要単語のセット “Situational Input” への関連強さに基づいて文法規則のサンプリング確率を設定し，設定したサンプリング確率に基づいて語彙の絞り込みをおこなう．

以降は，3.2 章で文生成手法の先行研究を紹介し，3.2 章でモンテカルロ木探索を用いた文生成の基本アルゴリズムを紹介した後，生成文がある状況について適切に説明しているか評価するための構文木の評価値と，それらの評価値全てを収束させるための UCB1 値の設定方法を説明する．さらに，探索を効率的に行うための探索方針の設定と探索範囲の絞り込み方法について説明する．3.3 章で提案手法の評価実験について述べ，3.4 章で本研究のまとめと今後の課題について述べる．

3.2 先行研究

文生成手法の先行研究として，大きく 2 種類のアプローチがある．1 つ目は，過剰生成と順位付けによって成される手法であり，中間段階で候補となる文を多数生成し，その後文生成のゴールとの一致度によって順位付けするというものである．具体的には，チャートパーザー上に構築されたシステム [19],[20]，森林構造を使用した HALogen/Nitrogen などのシステム [21]，tree conditional random fields を使用したシステム [22] などがあり，近

年数多く研究されているニューラルネットワークを用いた手法 [23],[24],[25] も，このアプローチを採用している．しかしながら，これらの手法は各アルゴリズムが求めた尤度などによって順位付けするため，上位の文で用いられる言い回しや語彙に偏りがうまれ，多様な表現の文の生成は期待できない．また，リカレントニューラルネットワークを用いた手法は単語の予測をするものであり，統語情報を考慮していないため誤った文法構造を持つ文を生成する可能性がある．

もう 1 つのアプローチは，文生成を目標指向計画問題と捉え，自動でプランニング処理を行うものである．このプランニング処理の例として，最初に意味的な処理を行って文の生成内容を調整し，次に自然言語表現に変換する (surface realization) といった，パイプライン生成処理 [26] という手法がある．

次にこのパイプライン生成に代わるものとして, 文の内容を調整するプランニングと自然言語表現に変換する処理を同時に行う研究 (CRISP[27], PCRISP[28]) がある．Graphplan[29] のような，多くの文生成プランニング手法で採用されている形式を入力として受け取り，意味的な要素や文法をエンコードする．プランニング中において，文生成と共に意味的な解析を行うため，非文法的な文の生成を防ぐ．

また，我々の提案手法と関連が強い研究として，UCT アルゴリズムを使用した統語構造を考慮した文生成の研究に，Sentence Tree Realization with UCT(STRUCT[30], S-STRUCT[31]) がある．しかしながら，これらの手法は，入力情報が意味的に限定され，ある特定の文を生成することを目的としており，多様な単語や語句の使用が制限されている．

3.3 提案手法

本節では，はじめに，モンテカルロ木探索を用いた文脈自由文法に基づく文生成手法の基本アルゴリズムを述べた後，適切な構文木を効率的に探索するための，生成文の評価方法，探索方針の設定方法，探索範囲の絞り込み方法について説明する．

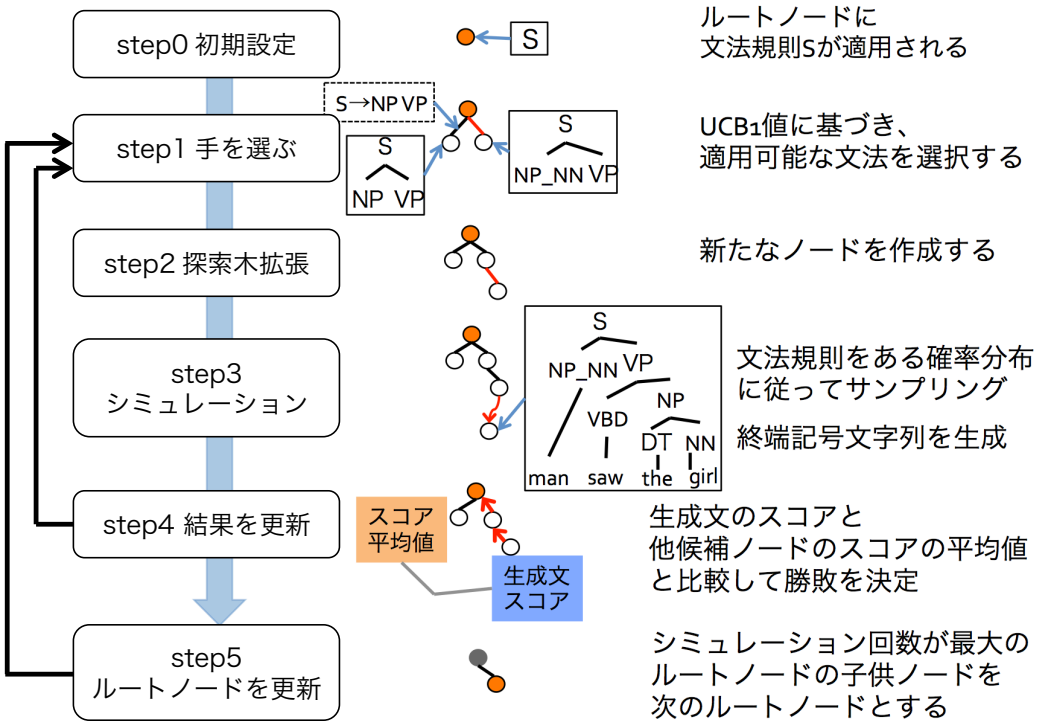


図 3.2: MCTS を用いた文生成

3.3.1 基本アルゴリズム

モンテカルロ木探索を用いた，文脈自由文法に基づく構文木構築のアルゴリズムを図 3.2 に示す．提案手法では，探索木の**エッジ**が，文脈自由文法の生成規則であり，**ノード**は構文木となる．**葉ノード**は文が完成した構文木，**内部ノード**は未完成の構文木となる．
最初の探索木の**ルートノード**に，文の開始記号 **S** が適用され (step0)，ある回数試行 (step1 から step4) を繰り返す．試行の後，**ルートノード**の**子ノード**の中から有望な**ノード**を決定し，次の**ルートノード**とする (step5)．その**ルートノード**から新たな探索木が作られ，再び試行 (step1 から step4) を繰り返す．以上のような手順を繰り返すことで最後に完成した構文木を生成する．

以降，適切な構文木を構築するための，下記の 2 点における具体的な手法を記載する．

- step4 における構文木評価値と勝敗決定方法 (3.3.2 節)
- step1 における適用可能な文法の絞り込み方法 (3.3.3 節)

3.3.2 構文木評価値と UCB1 値

ゲームの勝敗と同様に生成文に基づいた勝敗を決定するために，生成文のスコアを適切に設定し，スコアに基づいて勝敗を決定する．本節では，提案する複数の生成文のスコアと，全てのスコアを収束させるための UCB1 値の設定方法と，勝敗決定方法について述べる．

生成文のスコア

図 3.2 中 step4 における生成文のスコアについて説明する．我々は単語の繋がりの正しさと文の内容の適切さについての二つの視点から評価を行った．詳細は下記に示す．

1. 文の内容の適切さ
   Situational input として与えた単語群と生成文の SVO の単語について，word2vec による分散表現を求め，それらの cos 類似度を使用した．
2. 単語の繋がりの正しさ
   ngram language model に基づく値を用いた． 以下の 2 種類のどちらかを使用し、生成文に用いられる単語の傾向を比較した．
   - *Perplexity* (以降，PP)
     Kneser-Ney スムージング [18] による trigram の Perplexity の値を使用した．
   - *Acceptability* (以降，AP)
     英語母語話者にとっての許容可能性 (acceptability) を測る *Acceptability* [32] スコアを使用した。*Acceptability* スコアは下記式 3.1 の通り定義される．

$$Acceptability(s) = \log \left(\frac{p_{model}(s)}{p_{uni}(s)} \right)^{\frac{1}{|s|}} \tag{3.1}$$

     ここで，$p_{model}(s)$ は n-gram 言語モデルを示し，本論文では Kneser-Ney スムージング [18] による trigram を採用する．$p_{uni}(s)$ は unigram 確率を示す．

構文木評価値に基づく UCB1 値

図 3.2 中 step4 における UCB1 値の設定方法を述べる．上記で述べた複数の評価値を同時に収束させるために，各評価値における勝敗の項を持つ UCB1 値を式 (3.2) のように設定

した．

$$v_i^{cond.} + v_i^S + v_i^V + v_i^O + v_i^{PPorAP} + C\sqrt{\frac{2\log N}{n_i}} \tag{3.2}$$

勝率の期待値 $v_i^{cond.}$,v_i^S,v_i^V,v_i^O,v_i^{PPorAP} の値域が $[0,5]$ であるため，「知識の適用（exploitation)」と「探査(explanation)」のバランスを保つために C を 5 とした． $v_i^{cond.}, v_i^{\{S,V,O\}}, v_i^{AP}$ について下記に説明する．

- $v_i^{cond.}$

  探索初期において，SVO が含まれない文が多く生成されるため，生成文のスコアのみを考慮したとき候補ノード間で UCB1 値の差が生まれにくく探索の収束が難しい．探索初期における収束を早めるために，生成文の最低条件を設定し，最低条件を満たすときに勝ちとする． 生成文の最低条件とは，文の要素として主語-述語-目的語 (SVO) を持ち，文長の制約を満たすこととする． SVO を持ち，かつ文長制約を満たすとき，$v_i^{cond.} = 1$ を勝率とし， SVO を持たない，もしくは文長制約を満たさないとき，$v_i^{cond.} = 0$ を返す．
- v_i^S, v_i^V, v_i^O

  3.3.2 の 1 で述べた，文の内容の適切さを評価するための cos 類似度を用いた勝率である．$v_i^{cond.}$ が勝ちの時のみ S, V, O における cos 類似度を計算し，その値が他候補ノードにおけるそれらの平均値より大きい時，勝ち点 1 を返す．それ以外のとき 0 を返す．
- v_i^{PP} or v_i^{AP}

  3.3.2 の 2 で述べた，単語の繋がりの正しさを示す AP もしくは PP を用いた勝率である．$v_i^{\{S,V,O\}}$ のすべてにおいて勝ちのときのみ AP を計算し，その値が他候補ノードにおける AP の平均値より大きい時，勝ち点 1 を返す．それ以外のとき 0 を返す．

3.3.3 適用可能な文法の絞り込み方法

本章では効率的に探索を行うための 2 つのアプローチについて説明する．

探索の方針と SVO 判定

文生成時，推論済みの単語に依存してその他の単語を選択をすることになる．例えば，最左の単語が前置詞の a である場合 a の右隣の単語は a とのつながりを考慮して選択するこ

ととなる．我々は，文中の主要な要素となる situational input と関連の強い要素から構文木を構築することを考えた．　具体的には，動詞 → 名詞 → 形容詞 or 副詞 → ストップワードの順に木を展開する．動詞である述語は，主語と目的語の両方の語彙選択に関連するため，situational input の中で最も主要な要素と判断し，最初に推論することとする．　ここで，PCFG によって構築される構文木から直接 SVO を判定することはできないため，以下に示す順に木を展開し，SVO を決定した．

1. V から始まる品詞を展開
   初めて展開された終端記号を文の述語とする．
2. 述語より左の N から始まる品詞を展開
   初めて展開された終端記号を文の主語とする．
3. 述語より右の N から始まる品詞を展開
   初めて展開された終端記号を文の目的語とする．
4. J と RB から始まる品詞を展開
5. その他の品詞 (ストップワード) を展開

ここで，1，2，3 中において，初めて展開された終端記号を SVO とするのは，構文木の浅いところにある単語ほど文の主要な要素となる可能性が高いことを考慮している．

サンプリング対象の確率分布の設定

MCTS を用いて文生成をする際，膨大な文法規則と語彙により探索範囲が広大になってしまうことが問題となる．文法を一様分布からサンプリングするときを考える．例えば situational input の述語として eat が与えらるとき，eat と run が同じ確率で選択される．我々は，効率的に探索するために，situational input や 前後の単語とのつながりや，PCFG の確率を考慮した確率分布に基づき文法をサンプリングする方法を提案する．非終端記号，SVO，SVO 以外の自立語，ストップワードでそれぞれ選択する際の確率分布を設定した．詳細を以下に示す．

- 非終端記号
  PCFG の確率を使用する．

- SVO

  situational input と意味が似ている単語の選択される確率を高くするため，word2vec により、situational input として与えられた単語との cos 類似度を使用する．cos 類似度の正規化式を式 (3.3) に示す.

$$\frac{\exp(\beta r)}{\sum(\exp(\beta r))} \tag{3.3}$$

  ここで，r は cos 類似度であり，$\beta = 2.0$ とした.

- SVO 以外の自立語

  situational input words と共起しやすい単語の選択される確率を高くするため，situational input を W(e.g.{dog,eat,bread}) とした時の単語分布 $p(v|W)$ を求めることを考える．我々は LDA(Latent Dirichlet Allocation [33]) を用いた．[33]. LDA で学習済みのモデルから新規文書 W のトピック分布 θ_{new} を求め，これを $p(\theta_{new}|W)$ とする．また，$p(v|\theta_{new})$ はトピック毎の単語分布 ϕ を用い，式 (3.4) より求める．$p(v|W)$ は式 (3.5) より求める.

$$p(v|\theta_{new}) = \sum_{k} \phi_{k,v}\theta_{new,k} \tag{3.4}$$

$$p(v|W) = \int p(v|\theta_{new})p(\theta_{new}|W)d\theta_{new} \tag{3.5}$$

- ストップワード

  単語の前後の繋がりを考慮するため，bigram と逆 bigram 確率を求め，それらの平均値を使用した.

上記の場合分けについてのイメージを図 3.3 に示す.

3.3.4 語彙の絞り込み

MCTS を用いて文生成をする際，コーパスから収集された語彙が多量である場合，探索範囲が広大になり，シミュレーション回数が増大する．本論文では，situational input の各単語と共起しやすい語彙に絞り込むことで探索範囲を縮小する方法を検討する．situational input の全ての単語と同時に共起しやすい単語を収集することを考え，トピックモデルにより求めた単語の確率分布を採用した．具体的な語彙絞り込みの手続きを以下に示す.

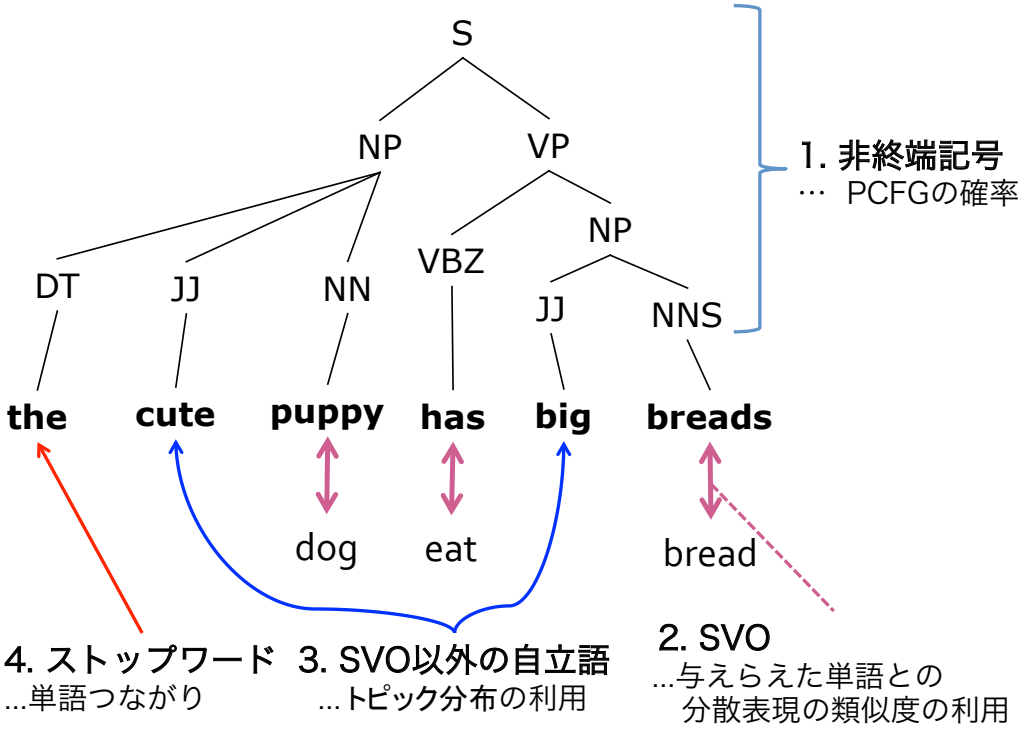


図 3.3: 確率分布の設定方法の全体像

1. コーパス中での出現回数が 5 回以上の単語を使用.
2. 3.3.3 で設定した分布を考慮

   - SVO

     cos 類似度が 0.3 以上の単語を使用.
   - SVO 以外の自立語

     それぞれの品詞において，$p(v|W)$(5.2 節で記載) が上位 20 単語のみ使用.

3.4 実験

3.4.1 実験設定

使用したモデルの学習コーパスについて説明する．LDA と AP について，Microsoft Research Video Description Corpus(MVDC) [1] 中の英文全て 85, 413 文を用いた．word2vec モデルと n-gram モデルは，Wikipedia English Corpus [2]を用いた．次に，探索範囲となる PCFG と語彙情報に用いたデータについて述べる．PCFG については，Penn TreeBank [13] の WSJ00 における 0000 ～ 0009 から作成した．文法数は 717 である．語彙情報は, Microsoft Research Video Description CorpusC 中の英文全てに出現した語彙の内，出現回数が 5 回以上の 4, 464 単語を用いた．また，文生成の situational input として式 () のように 3 種類について検証を行った．

$$W = \{dog, eat, bread\}, \{boy, play, soccer\} \{man, write, letter\}$$

文長の制約を 3 以上 5 以下の時，もしくは 6 以上 8 以下とした．また，シミュレーション回数は候補ノード数の 100 倍とした．

3.4.2 結果と考察

生成文例を表 3.2 に示す．表 3.2 に示すように同じ重要単語を与えた条件下で様々な単語や語句を使用した文が生成された．例えば，“situational input” として dog,eat,bread を使用したとき，dog について “another dog” や “dogs” という語句で表現していたり，boy, play, basketball を使用したとき、boy について，“boys” や “powerful boys” という語句で説明する文を確認した．文法構造については，文頭から “主語”→“述語”→“目的語” の順番で正しく単語選択されていることが確認された．一方で，“主語”，“述語”，“目的語” の周辺で不自然な意味を持つ単語が選択され，文全体として意味が通らない例も確認された．例えば，“situational input” として boy, play, basketball を使用したとき，“another boy play underneath soccer vaccination” という文が生成された．“underneath” や “vaccination” は文脈的に不自然な意味を持つ単語であり，文全体として意味が通らない文が確認された．

[1] https://www.microsoft.com/en-us/download/details.aspx?id=52422
[2] http://dumps.wikimedia.org/enwiki/

また，動詞や名詞の変形が不自然な例も確認された．例えば，“situational input” として dog,eat,bread を使用したとき，“another dog eaten smallest breads” という文が生成された．このとき，過去分詞形の “eaten” という時制を使用するのは不自然である．生成文の評価値として PP と AP を使用した時を比較すると，AP を使用した時の方が比較的多様な自立語を使用していることを確認したが，文生成の精度としては大きな違いは見られなかった．上述の通り，文法構造としては文頭から正しく “主語”→“述語”→“目的語” の順番で単語選択され，様々な単語や語句を使用する文が生成されたが，“主語”，“述語”，“目的語” の周辺で不自然な意味の単語が選択されたり，動詞や名詞の変形が不自然な例も確認された．次に，構文木を探索中の生成文の評価値について分析する．図 3.4 は {dog, eat, bread} を situational input としたときの文の最低条件を満たす確率の推移を示している．文の条件を満たす確率は，探索が進むとともに 1 に収束していることから探索初期において SVO を含み、文長の制約を満たす文を生成することが困難であり、探索後半では大抵の場合、文の条件を満たしていること確認した．また図 3.5 は，ルートノードが更新されるごとの探索中の各評価値の平均と分散の推移を示す．分散値はすべての値が小さく収束している．また，平均値は大きな変動はないが、探索初期よりも大きな値で探索が終わっていることが確認できる．このことからすべての値についてより良いスコアに収束するように探索が進んだことがわかる．

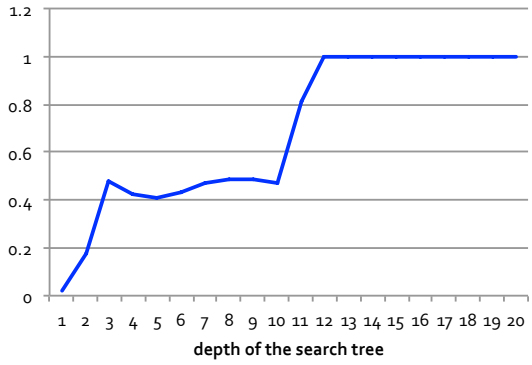


図 3.4: {dog, eat, bread} を Situational input としたときの文の最低条件を満たす確率の推移．

表 3.1: AP を使用した時の生成文例

$Situationalinput$	文長	AP
$\{dog, eat, bread\}$	$3 \sim 5$	· another dog eaten smallest breads
		· another dog ate connecticut breads
		· another dog eats smallest breads
		· another dog eating smallest breads
		· dogs eaten bread connecticut
	$6 \sim 8$	· either dog ate underneath jeff connecticut breads
		· another dog eaten automatically recumbent breads automatically
		· every dog eat underneath another outstretched bread breads
		· another dog eats automatically recumbent breads
		· another dog eats automatically battering breads
$\{boy, play, basketball\}$	$3 \sim 5$	· another boy play overweight soccer
		· another boy play eight soccer
		· boys play another soccer 's
		· another boy played overweight soccer
		· another boy play eight soccer
	$6 \sim 8$	· another boy play underneath soccer vaccination
		· half boy play underneath soccer connecticut
		· these boy play underneath soccer vaccination
		· another boy play recumbent soccer horizontally fistfighting soccer
		· the boy play underneath soccer vaccination proffestionals
$\{man, write, letter\}$	$3 \sim 5$	· man learn letter vaccination
		· another man read letter
		· overweight men read letter
		· men wrote letter vaccination
		· man wrote another smallest letter
	$6 \sim 8$	· that man wrote underneath letter connecticut
		· another man read together letter connecticut
		· another man read together letter vaccination letters binoculars
		· overweight men read another overweight letter
		· another man read together letter connecticut

表 3.2: PP を使用した時の生成文例

$Situationalinput$	文長	PP
$\{dog, eat, bread\}$	$3 \sim 5$	· both dog eats quick bread
		· both dog gather bread washington
		· dog eating eight bread
		· dog ate 1 bread
		· the dog eaten either breads
	$6 \sim 8$	· both dog ate with the breads and passengers
		· both dog eat on combining breads
		· a dog eat with italian quick breads poles
		· the dog eaten bread ' aggressively
		· an dog eat lighter breads overboard
$\{boy, play, basketball\}$	$3 \sim 5$	· refreshing boys played soccer
		· the boy play an soccer
		· powerful boys played soccer
		· every boy play five soccer
		· the boy play their soccer
	$6 \sim 8$	· any boy play for any soccer
		· every boy play till olympic soccer
		· a boy play between olympic soccer
		· the boy play soccer episode ritually
		· any boy play for its soccer
$\{man, write, letter\}$	$3 \sim 5$	· an man learn 10 letter
		· these man read letter washington
		· men read both italian letter
		· man learn your letter
		· these woman wrote letter washington
	$6 \sim 8$	· the man read in no smallest letter
		· the man read open letter episode
		· the man read together italian letter
		· the man read in both letter '
		· the man read in an lower letter

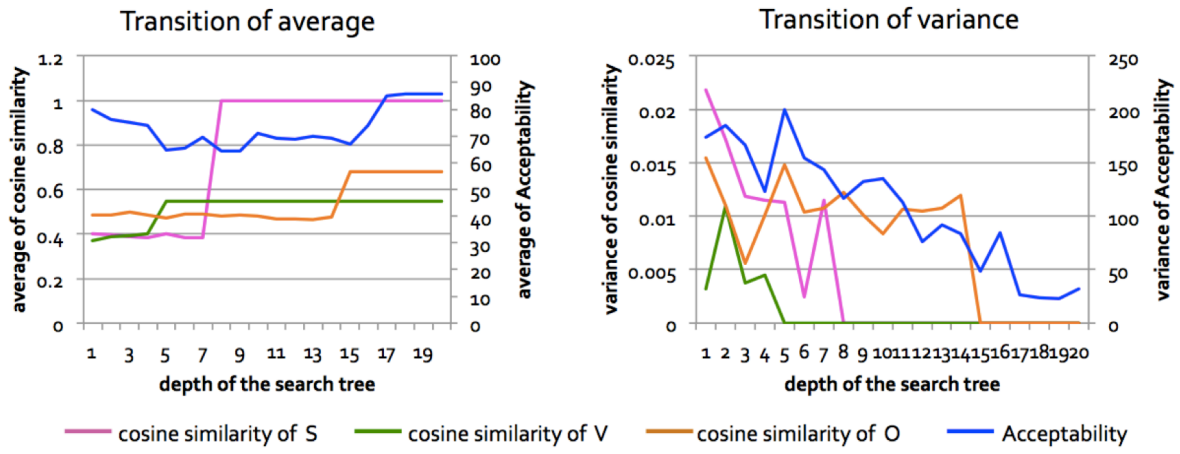


図 3.5: ルートノードが更新されるごとの探索中の各評価値の平均と分散の推移.

3.5 まとめ

文の文法構造的規則として文脈自由文法を採用し，文脈自由文法を適用して構築される構文木を生成する手法を提案し，文脈自由文法の文生成への有効性について検証した．タスクの設定として，文の主要な要素となる重要単語のセットを入力として，それらの単語を使用しながら，正確かつ柔軟な単語や語句の選択と，正しい文法構造を備えた文の生成を目指した．文脈自由文法を使用して構文木を構築する際，適切な構文木を機械的に自動判定するために構文木の評価指標を用意する必要がある．また，文法種類数が増えるほど構築され得る構文木は増大し，膨大な種類の構文木から状況に応じた適切な構文木を探索することが必要となる．我々は，適切な構文木を判定するための評価指標として，文法構造の適切さ，単語の繋がりの正しさ，生成内容の適切さの 3 つの視点に基づく構文木の評価値を提案した．また，効率的に適切な構文木を探索するために，モンテカルロ木探索アルゴリズムを使用した文脈自由文法に基づく文生成手法を提案した．また，効率的にモンテカルロ木探索のシミュレーションを行うために，生成文の評価値と探索方針の設定と探索範囲の絞り込み方法を提案した．実験では，与えた文の主要な要素となる重要単語のセットに基づいた様々な文が生成されることを確認した．一方で，与えられた重要単語以外の単語選択が意味的に不自然な例や，単語の変形を誤る例を確認した．今後の課題として，重要単

語のみならず，文全体としての意味を考慮した手法を検討し，文全体として意味的に自然な文を生成することが挙げられる．また，長い文の生成に向けて，事前に大量の文書データから学習する深層学習ベースの手法など，他の探索アルゴリズムの検討が挙げられる．

第4章　意味構造的規則に基づく文生成

第 1 章で述べた通り，正確かつ柔軟に単語を選択し，人間が定めた構造的規則に基づき正しい構造を持つ文の生成手法を検討するために，文の意味構造的規則に基づく文生成手法を検討し，意味構造的規則の文生成への有効性について検証する．本章では意味構造的規則として格構造ラベルを採用し，文生成時に格構造ラベルを条件として与え，与えられた格構造ラベルに対応する単語を選択しつつ，適切な格構造ラベルの順番を同時に推定する手法を提案する．タスクの設定として，画像を入力として，画像に対応する正しくかつ柔軟な単語や語句を選択しつつ，画像を説明する文として正しい意味構造をもつ文の生成を目指す．画像の特徴抽出と，適切な格構造ラベルの順番推定と，格構造ラベルに対応する正しくかつ柔軟な単語推定とを同時に行うため，end-to-end 構造の Neural Network を提案する．さらに，Neural Network が適切な文の意味構造を学習するために，離れた単語同士の関係性や格構造ラベル同士の関係性を学習可能な Transformer ベースの画像キャプショニング手法 [5] をベースラインとして使用する．

4.1　はじめに

本節ではまず画像キャプショニング分野を概観し，分野における課題を説明する．その後，画像キャプショニング分野における格構造ラベルを使用する手法の位置づけを整理する．格構造ラベルを使用する画像キャプショニング手法における課題を述べ，提案手法の概要を述べる．また，本章では，画像キャプショニングの分野において格構造ラベルを “意味役割ラベル（Semantic Role labels）” と呼ぶことから，以降，格構造ラベルを “意味役割ラベル” と記載する．

画像キャプショニングは画像の説明文を生成するタスクである．工場内を撮影した画像から作業者の行動ログを生成したり，小売店舗内を撮影した画像から店員の販売行動や

陳列行動のログを生成するなど，様々な場面で行動を記録することが期待され，多くの研究 [34, 35, 36, 37, 9, 38, 39, 40, 41, 42, 5] が報告されている．これらの手法の多くはエンコーダ部とデコーダ部で構成されており，エンコーダ部は画像を入力として画像特徴量を抽出し，デコーダ部は画像特徴量を入力として文の先頭から一単語ずつ逐次的に推定し生成文を出力する．更なる発展として，画像の内容を効果的に表現した画像特徴量を抽出するために，エンコーダ部に入力する画像中の各領域間の関係性を計算するアテンションモジュールを使用した手法 [9, 38, 39, 40, 41, 42] が報告されている．また，文生成の際に次の単語の推定に関係する画像特徴量により注目するために，エンコーダ部の出力である画像特徴量とデコーダ部の入力である直前まで推定された単語との接続をアテンションモジュールで計算する [5] が提案されている．特に Meshed-memory transformer [5] は，様々な抽象度の画像特徴量と単語間の関係性を計算するために，複数層からなるエンコーダ部とデコーダ部の各層間の関係の強さを学習することで高精度な画像キャプショニングを実現している．しかし，画像キャプショニング手法は入力情報が画像のみであり，説明対象を指定する機能は持たない．例えば小売店舗内を撮影した画像について，ユーザが店員の行動ログを必要とするときであっても，店員と顧客の行動を区別できず，顧客の行動に関する説明文を生成する可能性がある．画像キャプショニング手法は，様々な人物が様々な行動を同時に行っている実世界において，ユーザーが必要とする情報を指定して説明することができない．

制御可能画像キャプショニング（Controllable Image Captioning，以降 CIC とする）は，画像に加えて制御信号を入力として使うことで，ユーザが指定した情報に基づいて画像の説明文を生成するタスクである．制御信号として，“fly” などの「動詞 (V)」とその動詞に関連する “agent”，“directional”，“patient” などの「意味役割ラベル (Semantic Role labels. 以降，SR とする．)」のセット（動詞固有意味役割ラベル: Verb-specific Semantic Roles，以降 VSR とする.）を使用する方法（VSR-guided CIC）が提案され，CIC タスクにおいて高精度な制御性を実現している [7]．図 4.1 に，VSR-guided CIC の概要を示す．入力の VSR が “eat” と “agent1, agent2, patient1, patient2, verb” のセットの時，生成された説明文中において “agent1” は “a boy”，“agent2” は “in pajamas”，“eat” は “eating”，“patient1” は “a cake”，“patient2” は “with frosting” に相当する．VSR-guided CIC では，ユーザが指定した VSR に基づいて，VSR の各ラベルに相当するフレーズから構成された説明文を生成する．

しかしながら，VSR-guided CIC の既存手法 [7] は，構造上発生する推論誤りが説明文の精度を劣化させる．VSR-guided CIC では VSR に基づいた正確な説明文を生成するために，説明対象が存在する物体領域の推定と，説明文の意味構造を表す SR の順序を推定する．例えば図 4.1 左の例では，説明文を生成するために，説明対象である “agent1,2” や “patient1,2” に相当する “boy” や “cake” が存在する物体領域を推定する必要がある．さらに，説明文の意味構造を示す SR の順序を推定する必要がある．具体的には “eat” の動作主が “agent1,2” であり “eat” の行為対象が “patient1,2” であるという文の意味構造を示す，SR の順序 “agent1-agent2-verb-patient1-patient2” を推定する必要がある．説明対象が存在する物体領域の推定と SR の順序推定は，説明文の生成精度に強く関連する．既存手法 [7] は，これらの推定について，複数の独立したモデルによって逐次的に処理している．一つ目のモデルは，説明対象である各 VSR に相当する物体が存在する物体領域を推定し，二つ目は説明文の意味構造を示す SR の順序を推定する．三つ目は推定された物体領域と順序付けられた SR から説明文を生成する．これらの推定は独立して逐次的に処理されるため，各モデルにおける推論誤りは説明文の精度劣化の要因となる．具体的に，既存手法 [7] における実験では，物体領域の推論誤りが説明文の精度劣化の主要因であると報告している．

上記問題に対応するため，本論文では，説明文生成と共に物体領域の推定と SR の順序推定を同時に解決する End-to-End VSR-guided CIC モデルを提案する．提案手法はエンコーダ部とデコーダ部で構成され，エンコーダ部は物体領域と VSR を入力として中間特徴を出力する．エンコーダ部として transformer encoder [9] を採用することで，各入力情報間の関係性を計算する self-attention モジュール [9] により，VSR と関係の強い物体領域に重みづけされた中間特徴が抽出される．デコーダ部は直前までに推定された単語と中間特徴との両方を入力として、次の単語とその単語の SR を推定する．デコーダ部として transformer decoder [9] を採用することで，直前まで推定された単語同士の関係性を計算する self-attention モジュール [9] により直前までの文脈を考慮しながら，エンコーダとデコーダの接続の関係性を計算する source-target-attention モジュール [9] により，次の単語に関連する物体領域に注目して次の単語とその単語の SR を推定する．本論文ではさらに，SR の順序を正しく推定し，正しい意味構造の文を高精度に生成するために，直前までに推定された単語に加えて，直前までに推定された SR を入力し，次の単語と次の SR の推定を行う方法（SR-guided captioning decoder．以降，SR-dec.）を提案する．SR-dec により，直前まで推定された単語に加えて SR 間の関係性を考慮することで，直前までの文脈と同時

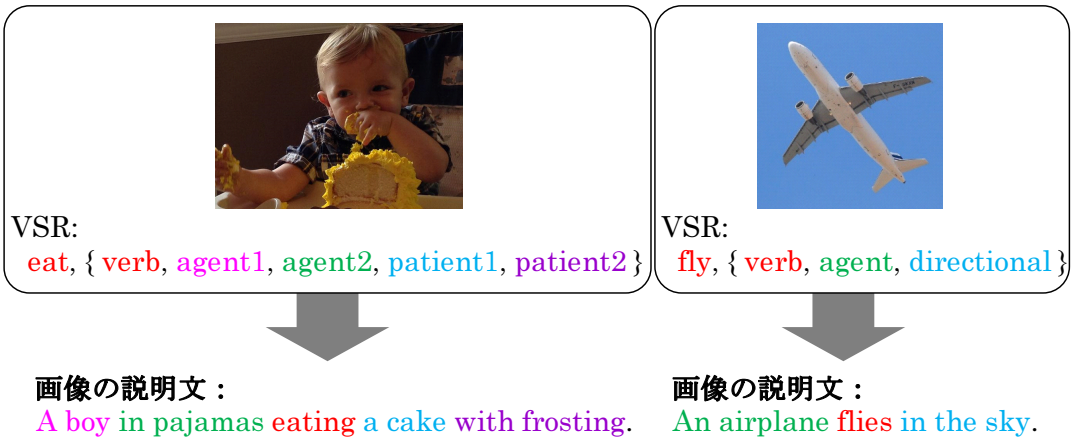


図 4.1: VSR-guided CIC の概要．画像と VSR のセットを入力として，指定された VSR に基づいて画像の説明文を生成する．

に文の意味構造を考慮しながら次の単語と SR を推定することができる．

実験では，CIC 手法の評価時に広く使用される 2 つのデータセット（COCO エンティティ [6] と Flickr30K エンティティ [43]）を使用し, 評価を行う．結果より，提案手法がベースラインと比較して最高精度を達成することを示す．さらに，SR-dec のアブレーション評価では，従来のキャプショニングデコーダと比較して SR-dec を使用する場合，より正しく SR の順序を推定でき，正しい意味構造を持つ説明文生成に寄与することを示す．

本章では下記 3 つの点で貢献している．

- VSR-guided CIC の既存手法における推論誤りを削減するための，End-to-End VSR-guided CIC モデルを提案する．
- 正しい意味構造を持つ説明文を高精度に生成するための SR-dec を提案する．
- ２種類のデータセットを使った実験により，我々の提案手法がベースラインと比較して最高精度であることを示す．アブレーション評価では，SR-dec が正しい意味構造の説明文生成に寄与することを示す．

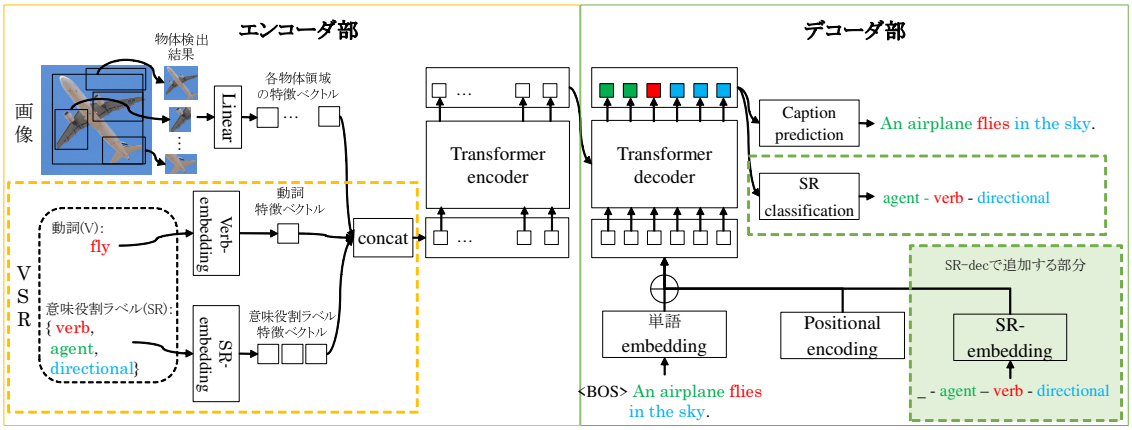


図 4.2: VSR を使用した End-to-End 制御可能画像キャプショニングモデルの全体像．点線枠内は，提案する End-to-End VSR-guided CIC を実現するためにベースラインの meshed-memory transformer [5] に追加した部分．

4.2 先行研究

本節では，画像キャプショニング手法について概観した後，制御可能画像キャプショニング手法について説明する．

画像キャプショニングは画像の説明文を生成するタスクである．近年，多くの DNN ベースの手法 [34, 35, 36, 37] が提案されている．これらの手法の多くはエンコーダ部とデコーダ部で構成されており，エンコーダ部は画像を入力として画像特徴量を抽出し，デコーダ部は画像特徴量を入力として文の先頭から一単語ずつ逐次的に推定し生成文を出力する．最初の単語の推定時は，画像特徴量に加え，文の先頭を意味するラベル “BOS” をデコーダ部に入力し，二単語目以降の推定では，画像特徴量に加え，直前まで推定された単語列をデコーダ部の入力とする．文末を意味するラベル “EOS” を出力するまで繰り返し単語を推定することで，画像の説明文を生成する．また，学習時においては，直前まで推定された単語列を入力する代わりに，正解文の直前までの単語列を入力する teacher forcing という学習方法が広く採用されている．更なる発展として，エンコーダ部に入力する画像中の各領域間の関係性を計算する attention モジュールを使用した手法 [39, 40, 41, 42] が報告されている．また，文生成の際に次の単語の推定に関係する画像特徴量に注目するために，エンコーダ部の出力である画像特徴量とデコーダ部の入力である直前まで推定された単語との接続

を attention モジュールで計算する手法 [9, 38, 5] が提案されている．特に meshed-memory transformer [5] は，様々な抽象度の画像特徴量と単語間の関係性を計算するために，複数層からなるエンコーダ部とデコーダ部の各層間の関係の強さを学習することで高精度な画像キャプショニングを実現している．

制御可能画像キャプショニング（CIC）は，制御信号として，画像中の物体領域 [6, 44] や文長 [45] を使用する研究が報告されている．物体領域を制御信号とするとき，指定されたすべての物体領域について言及した文を生成する．文長を制御信号とするとき，指定された文長の文を生成する．これらの制御信号はユーザが必要とする情報に絞って説明文を生成することはできない．例えば，ユーザが小売店舗内の店員の販売行動について知りたいとき，販売行動に関連する単語（現金，レジ，請求書など）を選択するように文生成を制御することができない．文長を制御信号とするとき，指定した文長が長すぎると，販売行動に無関係の単語（店舗内に置かれた掃除道具など）を使用して強制的に指定文長の生成文を出力してしまう．物体領域を制御信号とするとき，指定した物体領域に販売行動に無関係の物体（商品棚など）が映り込んでいると，強制的にその物体を示す単語を使って生成文を出力してしまう．この問題に対し，新しい制御信号として，“fly” など画像中の行動を表す「動詞」と，その動詞に関連する “agent” や “directional” などの「意味役割ラベル」とのセット（Verb-Specific Semantic Roles，以降 VSR とする）が提案されている [7]．VSR を CIC の制御信号として使用する（VSR-guided CIC） [7] ことで，ユーザが必要とする情報を正確に説明する文の生成を実現し CIC タスクにおける最高精度を達成した．

本稿では，高い制御性を実現する制御信号として VSR を採用し，VSR-guided CIC における既存手法 [7] の推論誤りを削減するための End-to-End VSR-guided CIC モデルを提案する．提案手法のニューラルネットワーク構造のベースラインとして，End-to-End 画像キャプショニング手法の内，高精度な文生成を実現する meshed-memory transformer [5] のネットワーク構造を採用する．

4.3 提案手法

既存手法の推論誤りを削減するため，説明対象が存在する物体領域の推定と，SR の順序推定と，推定された物体領域と順序付けられた SR からの説明文生成を同時に実現する

End-to-End VSR-guided CIC モデルを提案する．提案手法はエンコーダ部とデコーダ部で構成されている（全体概要を図 4.2 に示す）．エンコーダ部は，画像の各物体領域と VSR のセットを入力とし，入力された各物体領域と，各 VSR との間の関係性を計算しながら中間特徴を抽出する．デコーダ部は，単語系列とエンコーダの出力である中間特徴との両方を入力とし，入力の単語系列の次の単語とその単語の SR を推定する．提案手法のエンコーダ部の transformer encoder とデコーダ部の transformer decoder のネットワーク構造のベースラインとして，様々な抽象度の画像特徴と単語間の関係性をモデル化し，高い生成精度を実現した meshed-memory transformer [5] を採用する．図 4.2 中，エンコーダ部における橙点線とデコーダ部における緑点線枠内が本論文で提案する機能にかかる部分である．

4.3.1 手法概要

エンコーダ部 は，説明対象である各 VSR と関係の強い物体領域に重みづけされた中間特徴の抽出を担う部分である．具体的には，エンコーダ部は，画像と VSR（「動詞 (V)」と「意味役割ラベル (SR)」のセット）を入力とし，動詞や各意味役割ラベルに対応する物体領域に重みづけされた中間特徴を抽出する．最初に，画像は，学習済みの物体検出モデルに入力され，検出結果である各物体領域ごとの画像特徴を取得する．VSR の「動詞 (V)」については，動詞種類数のサイズの one-hot ベクトルを作成する．入力する動詞が “fly” のときは fly に相当するインデックスのみ 1 で，その他は 0 である．「意味役割ラベル (SR)」については，サイズが意味役割ラベル種類数の one-hot ベクトルを作成する．入力する意味役割ラベルが “agent” のとき，“agent” に相当するインデックスのみ 1 で，その他は 0 である．各物体領域の画像特徴と動詞と意味役割ラベルの one-hot ベクトルについて，それぞれ用意された Linear 層（図 4.2 中の Linear，Verb embedding，SR embedding）に入力し，同じサイズの特徴ベクトルに変換する．これらの特徴ベクトルは系列方向に連結し，図 4.2 中の transformer encoder に入力する．transformer encoder は，self-attention モジュールにより各入力特徴の相互の関係性の強さ（self-attention マップ）を計算し，self-attention マップによって各入力特徴が重みづけされ抽出された中間特徴を出力する．本手法においては，VSR と画像特徴間の関係の強さが計算され，VSR と関係の強い画像特徴が重みづけされた中間特徴が出力される．VSR と関係の強い画像特徴とは，説明対象が存在する物体領域の特徴と捉えられることから，エンコーダ部の出力である中間特徴は，説明対象が存

在する物体領域が重みづけされた特徴量であると考えられる．

デコーダ部 は，SR の順序推定と，説明文生成を担う部分である．具体的には，デコーダ部は直前まで推定された単語とエンコーダの出力である中間特徴を入力とし，次の単語とその単語の SR を推定する．最初に，直前まで推定された各単語から語彙数サイズの one-hot ベクトルを作成する．入力する単語が “I” のときは “I” に相当するインデックスのみ 1 で，その他は 0 である．これらの one-hot ベクトルを図 4.2 中の単語 embedding に入力し単語特徴ベクトルに変換する．また，Positional encoding [9] によって，各単語の文中における位置を表す位置特徴ベクトルを作成する．単語特徴ベクトルと位置特徴ベクトルを足し合わせた後，transformer decoder に入力する．transformer decoder 内では，エンコーダ部から受け取った中間特徴と直前まで推定された単語の特徴ベクトル間の相互の関係性の強さ（source-target-attention マップ）を計算する．transformer decoder は，source-target-attention マップによって重み付けされ抽出された特徴ベクトルを出力する．最後に，この特徴ベクトルを Caption prediction モデルと SR classification モデルに入力し，各モデルによって次の単語とその単語の SR を推定する．Caption prediction モデルは，特徴ベクトルを語彙数長のベクトルに変換することで，各単語の確率を推定する．SR classification モデルは，特徴ベクトルを SR ラベル種類数長のベクトルに変換することで，各 SR ラベルの確率を推定する．

SR-guided captioning decoder (SR-dec) さらに本論文では，SR の正しい順序を推定し，正しい意味構造の文を高精度に生成するために，直前までに推定された単語に加えて，直前までに推定された SR をデコーダーに入力する方法（SR-guided captioning decoder. 以降，SR-dec. 図 4.2 中，緑網掛け部分.）を提案する．最初に，直前まで推定された各 SR ラベルからサイズが意味役割ラベル種類数の one-hot ベクトルを作成する．これらの one-hot ベクトルを，エンコーダ部と同様の SR embedding に入力し，SR の特徴ベクトルに変換する．単語特徴ベクトルと位置特徴ベクトルと SR 特徴ベクトルを足し合わせた後，transformer decoder に入力する．直前までの単語と SR をデコーダに入力することで，直前までの文脈と意味構造を同時に考慮でき，より正確な意味構造を持つ説明文の生成を実現する．

4.3.2 損失関数

説明文生成の学習をするために，既存の画像キャプショニング手法で採用されているキャプションロスを踏襲する．同時に，SR の推定を学習するために，SR 分類ロスを使用する．キャプションロスと SR 分類ロスを同時に減衰させるように学習する．

キャプションロス は，Caption prediction モデルによって出力された各単語の確率について計算される．多くの画像キャプショニング手法 [46, 47, 48] で一般的な，事前学習とファインチューニングの 2 段階の学習方法を踏襲する．事前学習では，推定された単語と正解文における単語の間で，下記式で定義されるクロスエントロピー損失（L_{cap}^{ce}）を算出する．

$$L_{cap}^{ce} = -\frac{1}{n}\sum_{i=1}^{n} y \log p(x), \tag{4.1}$$

ここで，n は文長，x は単語を示し，$p(x)$ は推定した単語の確率，y は正解ラベルであり 0 か 1 の値である．ファインチューニングでは，ビームサーチ [48] で探索された文に対し，予め設定した報酬関数について，以前より報酬が向上するように学習を進める self-critical sequence training（SCST）[47] を行う．既存手法 [48] を踏襲し，SCST における報酬関数は CIDEr-D [49] を使用し，報酬の平均をベースラインとする．サンプルされた文についての損失関数は下記式のように定義する．

$$L_{cap}^{rl} = -\frac{1}{k}\sum_{i=1}^{k}(r(w^i) - b)\log p(w^i), \tag{4.2}$$

ここで，w^i はサンプルされた候補の内，i 番目の文であり，$r()$ は報酬関数，b はベースラインであり，候補文の報酬の平均として算出される．推論時は，ビームサーチを使用して文を生成し，候補文の中で最も高い確率の文を出力する．

SR 分類ロス は，SR classification モデルによって出力された各単語の SR の確率について計算される．各単語は複数の意味役割ラベルを持つことができるため SR classification はマルチラベル分類を行うため，下記式のように定義されるバイナリクロスエントロピー損失を使用する．

$$L_{sem} = -y\log(q) - ((1-y)\log(1-q)), \tag{4.3}$$

ここで，q は SR classification の出力値であり，y は 0 または 1 の正解ラベルである．

全体ロス 事前学習時における全体ロスはキャプションロスと SR 分類ロスの線形和とし，下記式のように定義する．

$$L_{all} = aL_{cap}^{ce} + (1-a)L_{sem}. \tag{4.4}$$

ファインチューニング時は，キャプションロスのみを採用し，全体ロスは L_{cap}^{rl} のみとする．

4.4 実験

実験では下記３つの観点について評価する．一つ目は，提案手法を VSR-guided CIC の既存手法 [7] と比較することで，VSR-guided CIC において提案手法の End-to-End 構造が説明文生成の精度向上に寄与するか評価する．二つ目は，提案手法のネットワーク構造のベースラインとして採用した meshed-memory transformer [5] と比較することで，End-to-End のキャプショニング手法における VSR の制御性を評価する．三つ目に，アブレーション評価では，従来のキャプショニングデコーダと提案する SR-dec を比較することで，SR-dec が SR の順序の正確性向上に寄与するか評価する．

4.4.1 実験設定

評価値

生成された説明文の精度を定量的に評価するため，画像キャプショニング手法で広く使用される評価値である，BLEU-4 (B4) [50]，METEOR (M) [51]，ROUGE (R) [52]，CIDEr-D (C) [49]，SPICE (S) [53] を使用した．アブレーション評価において，SR-dec を使用して推定された SR の順序の正確さを評価するため，二種の評価値を使用した．一つは，SR のセットとしての recall 値，二つ目は SR の系列としての recall 値であり，それぞれ R_{SR1} および R_{SR2} と記載する．R_{SR1}，R_{SR2} を下記式のように定義する．

$$R_{SR1} = \frac{\sum_{i=1}^{k} N(OUT_{set}^{i} \cap GT_{set}^{i})}{\sum_{i=1}^{k} N(GT_{set}^{i})}. \tag{4.5}$$

$$R_{SR2} = \frac{\sum_{i=1}^{k} c_i}{k}. c_i = \begin{cases} 1 & (OUT_{seq}^{i} = GT_{seq}^{i}) \\ 0 & (OUT_{seq}^{i} \neq GT_{seq}^{i}) \end{cases} \tag{4.6}$$

ここで，$N()$ は () 内の集合に含まれるラベル数，k は全サンプル数を示す．OUT^i_{set}，GT^i_{set}，OUT^i_{seq}，GT^i_{seq} は，それぞれ i 番目のサンプルにおける，出力 SR 系列中の SR ラベルの集合，正解 SR 系列中の SR ラベルの集合，出力 SR 系列，正解 SR 系列を示す．上述した全ての評価値は，単一の生成文について単一の正解文に対して算出する．

比較手法

提案手法について 3 つの観点で評価するため下記の 3 種類の比較手法を採用した．

VSR-guided CIC における既存手法（既存手法） [7]．
VSR-guided CIC において，提案手法の End-to-End 構造の説明文生成精度への影響を評価するために既存手法 [7] を採用した．

Meshed memory transformer (Meshed)． End-to-End ネットワーク構造における VSR の制御性を評価するために，提案手法のニューラルネットワーク構造のベースラインである Meshed [5] を採用した．

提案手法 w/o SR-dec． SR-dec のアブレーション評価として，提案手法から SR-dec 部分のみ省いた手法を用いた．

実験設定

実験に使用したデータセットと，提案手法の実装時の詳細な設定について説明する．

データセット． 既存手法 [7] の実験を踏襲し，31,000 枚の画像で構成される Flickr30K Entities [43] と，120,000 枚の画像で構成される COCO Entities [6] を使用した．すべての画像には，5 つのキャプションが付与されている．既存手法 [7] と同様に，画像中の行動を説明しているキャプションに絞り込むため，キャプション中に動詞を含むサンプルを収集した．また，VSR アノテーションの作成方法も既存手法 [7] を踏襲した．具体的には，事前学習済みの意味役割ラベリングツール [54] を使用し，各キャプションの各単語に対して動詞と意味役割ラベルを付与し，それらのラベルアノテーションを正解アノテーションとみなした．COCO と Flickr30K の動詞種類はそれぞれ $2,662$ と $2,926$ であり，意味役割ラベル種類数は 25 である．使用した意味役割ラベルの種類を表 4.1 に示す．

表 4.1: 使用した意味役割ラベル

略称	意味役割ラベル	和訳
Arg0	agent	動作主
Arg1	patient	述語の対象
Arg2	instrument	道具
Arg3	starting point	起点
Arg4	ending point	終点
COM	comitative	共格
LOC	locative	場所
DIR	directional	方向
GOL	goal	動作の目的
MNR	manner	様態
TMP	temporal	時間
EXT	extent	程度
REC	reciprocals	相互
PRD	secondary predication	二次述語
PRP	purpose	目的
PNC	purpose not cause	行動の動機
CAU	cause	原因
DIS	discourse	接続
ADV	adverbials	副詞
ADJ	adjectival	形容詞
MOD	modal	法助動詞
NEG	negation	否定
LVB	light verb	軽動詞

表 4.2: キャプショニング評価指標による生成された説明文の精度比較 (%)

	COCO Entities [6]					Flickr30K Entities [43]				
Method	B4	M	R	C	S	B4	M	R	C	S
Meshed [5]	12.8	18.9	40.4	126.5	27.3	6.8	12.2	29.1	41.3	18.6
既存手法 [7]	16.0	23.2	47.1	162.8	35.7	7.9	14.7	32.6	71.6	18.2
提案手法 (w/o SR-dec)	24.7	27.0	53.0	214.0	**40.4**	**13.7**	**18.8**	**39.8**	**112.0**	23.5
提案手法 (w/ SR-dec)	**24.8**	**27.1**	**53.3**	**216.2**	**40.4**	13.6	18.7	39.6	109.7	**23.6**

実装 提案手法のニューラルネットワーク構造のベースラインとして採用した Meshed [5] の実装の詳細を踏襲した．具体的には，入力画像から物体検出結果を取得するために，VisualGenome データセット [55] でファインチューニングされた FasterR-CNN [56] を使用した．物体検出結果の内，出力確率が上位 50 件の物体領域について 2048 次元の特徴ベクトルを取得した．Transformer のパラメータ設定は，各層 512 次元とし，head の数を 8，メモリベクトルのサイズを 40 とした．各 attention 層と feedforward 層の後，0.9 の確率でドロップアウトを採用した．デコーダーの入力の単語系列は，事前学習時は，正解文における直前までの単語系列を，ファインチューニング時は，直前までに推定された単語系列を採用した．事前学習時の学習率は [9] を踏襲し，ファインチューニング時は学習率を 5×10^6 に固定した．最適化手法は Adam [57]、バッチサイズは 50，ビームサイズは 5 として，すべてのモデルを学習した．

また，入力の動詞 (V) と意味役割ラベル (SR) について，サンプルごとに数が変動するため最大種類数をそれぞれ 5, 10 と設定した．動詞種類数が 5 未満のとき，余剰分は “none-V” ラベルを設定した．“none-V” ラベルについては，transformer encoder 内での計算に影響しない様に，“none-V” に該当するインデックスのみ 0 を，それ以外は 1 を掛け合わせるマスク処理を行った．意味役割ラベル種類数が 10 未満の時，余剰分は “none-SR” ラベルを設定し，“none-SR” ラベルについても同様に transformer encoder 内でマスク処理を行った．式 (4.4) のパラメータ a は，$[0.1, 0.9]$ の範囲を 0.1 毎に変化させて実験した．

表 4.3: recall ベースの SR の評価指標による，SR のセットと系列に関する精度比較．(%)

	COCO [6]		Flickr30K [43]	
Method	R_{SR1}	R_{SR2}	R_{SR1}	R_{SR2}
既存手法 [7]	49.6	5.84	51.3	5.83
提案手法 (w/o SR-dec)	98.4	66.8	95.9	53.6
提案手法 (w/ SR-dec)	**99.7**	**81.5**	**97.9**	**60.2**

4.4.2 結果と考察

定量評価． 表 4.2 は，式 4.4 のパラメータ $a = 0.5$ のときのキャプショニング評価値の値を示す．既存手法 [7] や Meshed [5] に対し，両データセットにおけるすべての評価値について，提案手法（w/ SR-dec，w/o SR-dec）が最高精度を達成している．単語の並びに関する評価値 (B4, M, R, C) に加え，文の構文木構造に関する評価値 (S) についても精度が向上している．提案手法は VSR を使用し文の意味構造を制御しながら画像説明文を生成することで，単語の並びに加えて文の構造の正確性も考慮した説明文が生成可能であると考えられる．また，各比較手法と比べると，既存手法 [7] を上回ったことから，提案手法の End-to-End 構造が，VSR-guided CIC における説明文の生成精度向上に寄与していることが考えられる．Meshed [5] よりも精度が向上していることから，End-to-End 構造のキャプショニング手法において，VSR による制御性が有効に働くことが考えられる．表 4.3 は，SR についての recall ベースの評価結果を示す．提案手法（w/ SR-dec，w/o SR-dec）は，R_{SR1} と R_{SR2} 共に，既存手法 [7] より大幅に精度が向上していることが確認できる．提案手法の End-to-End 構造が，既存手法の SR 順序の推論誤りの削減に効果があることが考えられる．また，SR-dec を使用した場合の方が SR-dec を使用しない場合よりも精度が向上していることが確認できる．特に，SR 系列に関する評価値 R_{SR2} が大幅に改善されている．

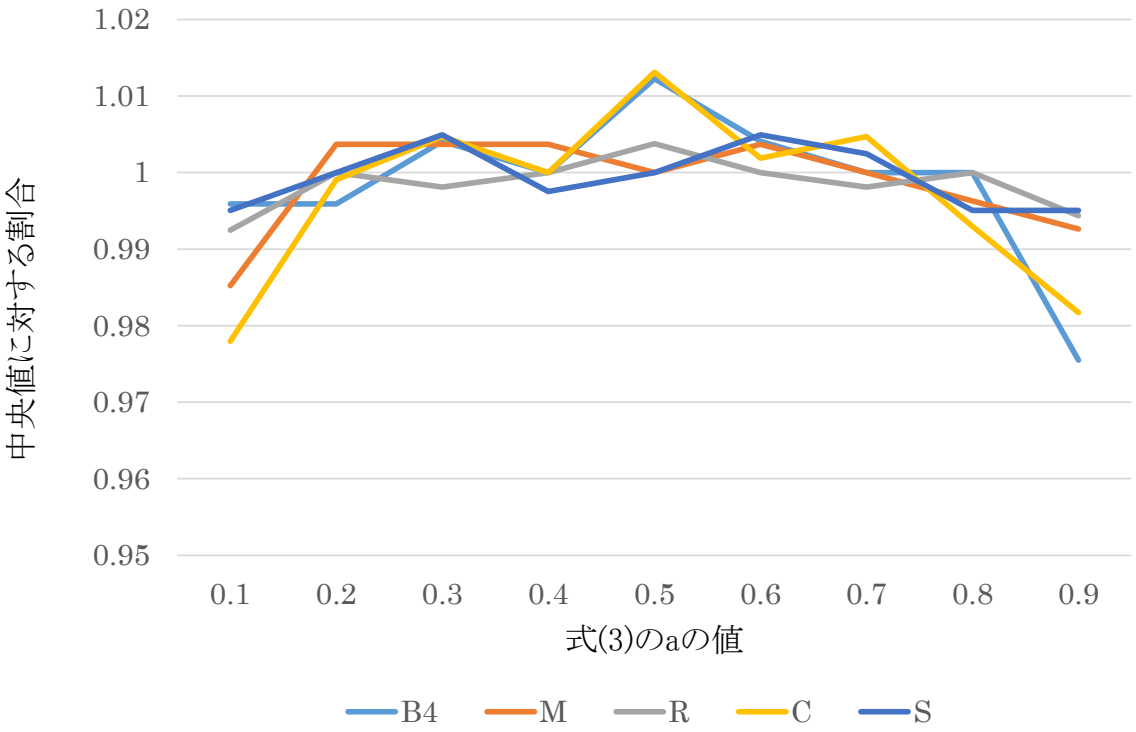
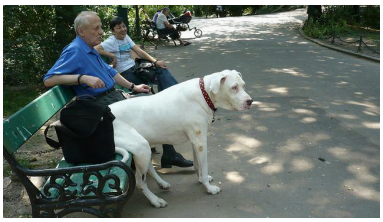


図 4.3: COCO Entities [6] を使用したときの提案手法 (w/ SR-dec) における各キャプショニング評価値について，式 (3) の a の値を $[0.1, 0.9]$ の範囲で 0.1 毎に変化させたときの各キャプショニング評価値の中央値に対する割合．

これらの結果より，SR-dec が，正しい意味役割ラベルの順序，つまり正しい意味構造を持つ説明文の生成に寄与することが考えられる．

次に，提案手法の式 (4.4) のパラメータ a を $[0.1, 0.9]$ の範囲で 0.1 毎に変化させたときの説明文生成精度への影響を確認する．COCO Entities [6] を使用して実験したときの各キャプショニング評価値の中央値に対する，各キャプショニング評価値の割合を図 4.3 に示す．パラメータ a の値は $[0.3, 0.7]$ の範囲であれば説明文の生成精度を高水準で維持できることが確認できた．キャプションロスと SR 分類ロスの両方を同程度に考慮することにより，説明文生成精度向上に寄与することが考えられる．

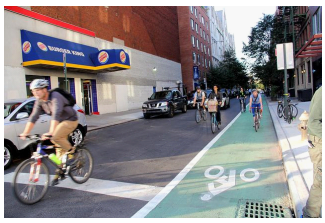
可視化による定性評価.図 4.4 は，サンプル画像と，正解の説明文と SR 系列，各比較手



GT：patient-verb(sit)-instrument-prediction
既存手法：agent-verb(sit)-patient
提案手法 (w/o SR-dec)：patient-verb(sit)-instrument-prediction
提案手法 (w/ SR-dec)：patient-verb(sit)-instrument-prediction

GT：**a dog sitting on a bench next to an old man**
Meshed：**a white dog and a man on a bench with a**
既存手法：**a man sitting on a bench with a dog**
提案手法 (w/o SR-dec)：**a man sitting on a bench next to a white dog**
提案手法 (w/ SR-dec)：**a man sitting on a bench next to a white dog**

サンプル a



GT：agent-verb(pass)-patient-locative
既存手法：agent-verb(pass)-patient
提案手法 (w/o SR-dec)：agent-verb(pass)-patient-locative
提案手法 (w/ SR-dec)：agent-verb(pass)-patient-locative

GT：**bike riders passing burger king in city street**
Meshed：**a group of people riding bikes down a street**
既存手法：**a man passes a bike down a street**
提案手法 (w/o SR-dec)：**a motorcyclist passes a corner on a city street**
提案手法 (w/ SR-dec)：**a man passing a bicycle on a street**

サンプル b

GT：patient-patient-patient-verb(hang)
既存手法：agent-verb(hang)-patient
提案手法 (w/o SR-dec)：patient-patient-verb(hang)-patient
提案手法 (w/ SR-dec)：patient-patient-verb(hang)

GT：**a utility table in a kitchen with pots pans and utensils hanging nearby**
Meshed：**a kitchen with a wooden chair and a wooden table**
既存手法：**a cart with pots and pots hanging on a rack**
提案手法 (w/o SR-dec)：**a kitchen stove with pots and pans hanging on it**
提案手法 (w/ SR-dec)：**a kitchen with pots and pans hanging up**

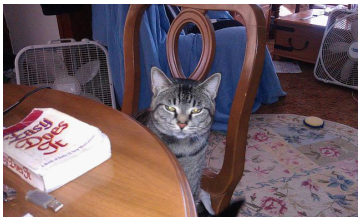
サンプル c

図 4.4: 入力画像と，正解（GT）と既存手法 [7]，提案手法 (w/o SR-dec，w/ SR-dec) により推定された SR 系列（赤字で記載）と，正解（GT）と Meshed [5]，既存手法 [7]，提案手法 (w/o SR-dec，w/o SR-dec) により生成された説明文（黒太字で記載）の例.

GT：patient-verb(swim)-locative
既存手法：patient-verb(swim)-locative
提案手法 (w/o SR-dec)：patient-verb(swim)-locative
提案手法 (w/ SR-dec)：patient-verb(swim)-locative

GT：**two ducks are swimming in some brown water**
Meshed：**two ducks swimming in a body of water**
既存手法：**a water swimming in the water**
提案手法 (w/o SR-dec)：**two ducks are swimming in the water**
提案手法 (w/ SR-dec)：**two young geese are swimming in the water**

サンプル d



GT：patient-verb(sit)-instrument-locative-manner
既存手法：patient-verb(sit)-instrument-locative
提案手法 (w/o SR-dec)：patient-verb(sit)-instrument-manner
提案手法 (w/ SR-dec)：patient-verb(sit)-instrument-locative-manner

GT：**a cat sitting in a chair at a table with a book on it**
Meshed：**a cat sitting on a chair next to a table**
既存手法：**a cat sitting next to a cat sitting on a chair**
提案手法(w/o SR-dec)：**a cat sitting on a chair with a newspaper next to it**
提案手法(w/ SR-dec)：**a cat sitting in a chair at a table with a book**

サンプル e

GT：agent-verb(fly)-directional-manner-manner
既存手法：agent-verb(fly)-locative
提案手法(w/o SR-dec)：agent-verb(fly)-directional-manner
提案手法(w/ SR-dec)：agent-verb(fly)-directional-manner-manner

GT：**white jet plane flying in the sky with engines in the wings**
Meshed：**an airplane is flying in the blue sky**
既存手法：**an airplane flying in the sky with a blue sky**
提案手法(w/o SR-dec)：**an airplane flying in the sky with the landing gear down**
提案手法(w/ SR-dec)：**an airplane flying through a clear blue sky with propellers in the sky**

サンプル f

図 4.4: 入力画像と，正解（GT）と既存手法 [7]，提案手法 (w/o SR-dec，w/ SR-dec) により推定された SR 系列（赤字で記載）と，正解（GT）と Meshed [5]，既存手法 [7]，提案手法 (w/o SR-dec，w/o SR-dec) により生成された説明文（黒太字で記載）の例.

GT：patient-verb(lather)-instrument
既存手法：agent-verb(brush)-patient
提案手法 (w/o SR-dec)：_none_
提案手法 (w/ SR-dec)：patient-verb(surprise)-instrument

GT：**a young girl is lathered up with toothpaste**
Meshed：**a woman is brushing her teeth with a spoon**
既存手法：**a woman brushing a bite of a fork**
提案手法 (w/o SR-dec)：**a woman is <unk> a piece of pizza**
提案手法 (w/ SR-dec)：**a close up of a person is surprised by a toothbrush**

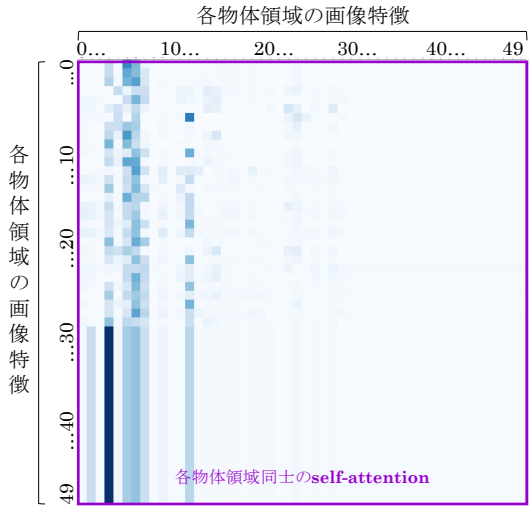
サンプル g

GT：patient-verb(look)
既存手法：agent-verb(look)-patient
提案手法 (w/o SR-dec)：agent-verb(look)-patient
提案手法 (w/ SR-dec)：agent-verb(look)-patient

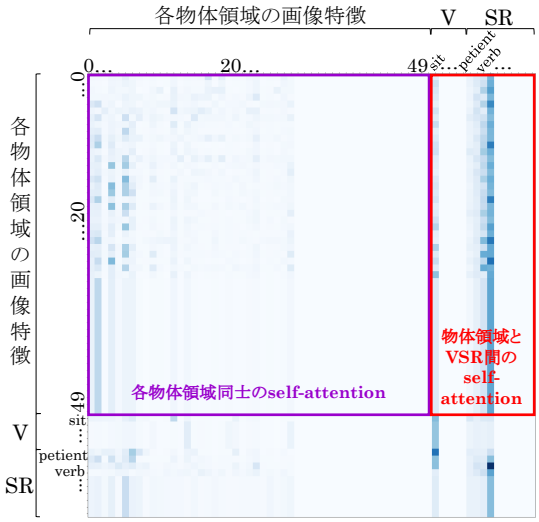
GT：**a cake that has been made to look like a cup**
Meshed：**a piece of cake on a white plate with a**
既存手法：**a cake looks on a plate**
提案手法 (w/o SR-dec)：**we are looking at a piece of cake**
提案手法 (w/ SR-dec)：**we are looking at a piece of cake**

サンプル h

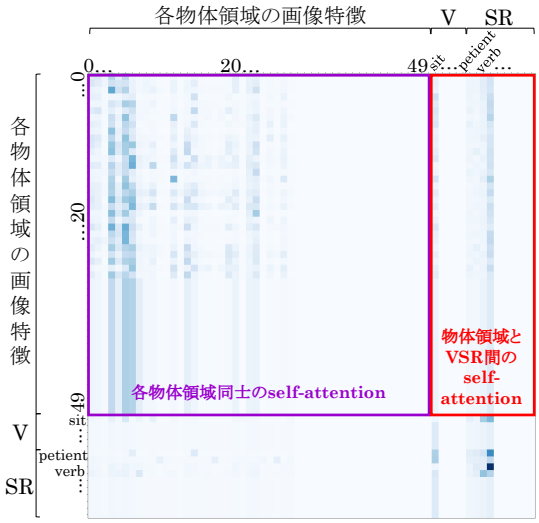
図 4.4: 入力画像と，正解（GT）と既存手法 [7]，提案手法 (w/o SR-dec，w/ SR-dec) により推定された SR 系列（赤字で記載）と，正解（GT）と Meshed [5]，既存手法 [7]，提案手法 (w/o SR-dec，w/o SR-dec) により生成された説明文（黒太字で記載）の例.



Meshed [5] の transformer encoder における self-attention マップ

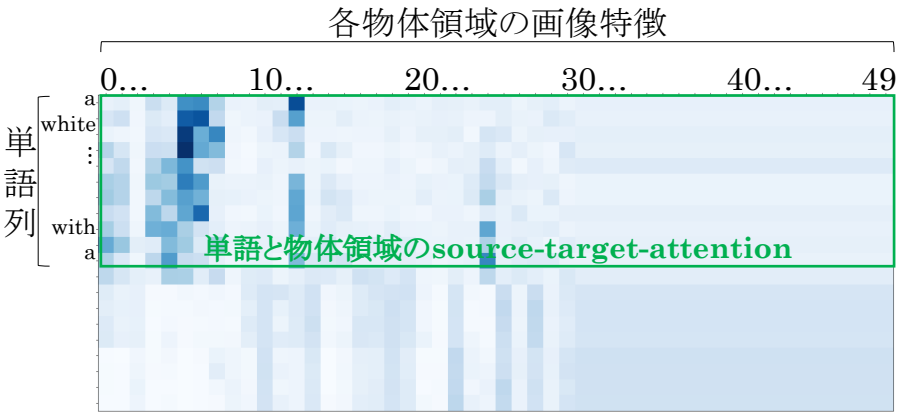


提案手法 (w/o SR-dec) における self-attention マップ

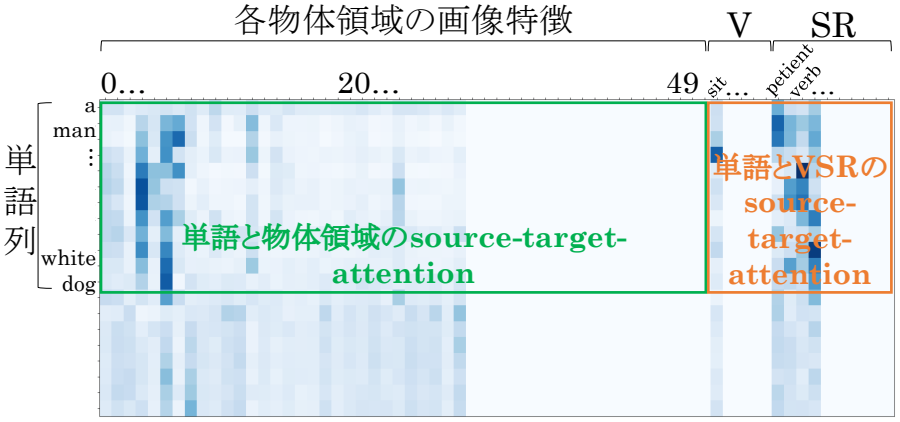


提案手法 (w/ SR-dec) における self-attention マップ

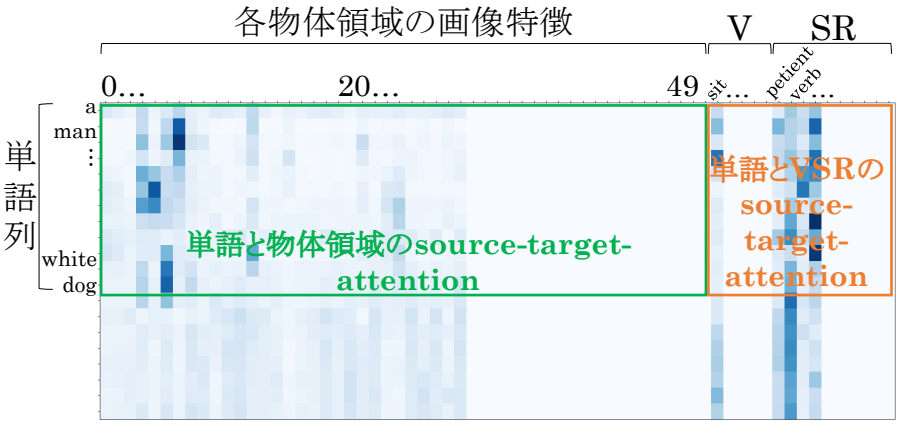
図 4.5: 図 4 のサンプル a について，self-attention マップの可視化例.



Meshed [5] における source-target-attention マップ

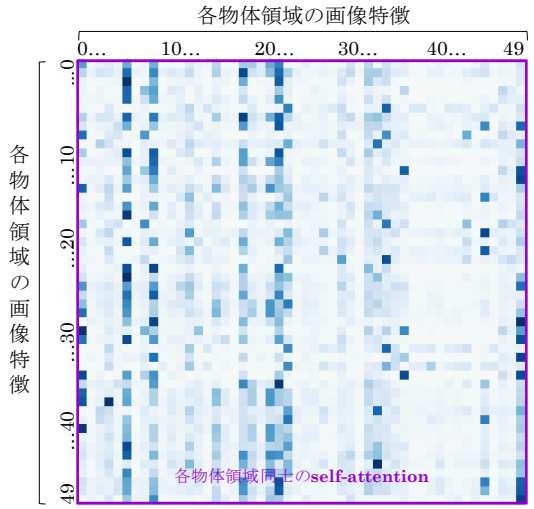


提案手法 (w/o SR-dec) における source-target-attention マップ

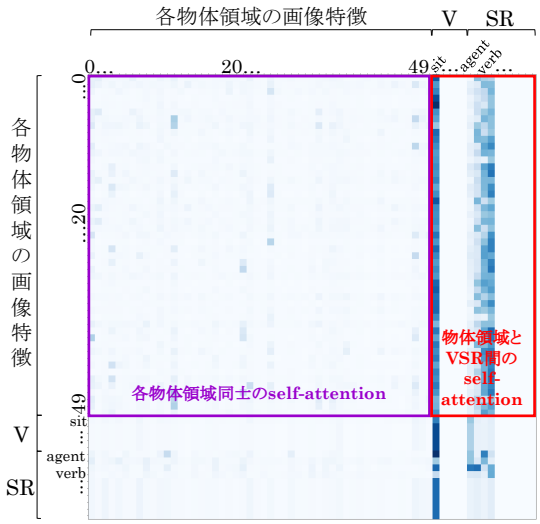


提案手法 (w/ SR-dec) における source-target-attention マップ

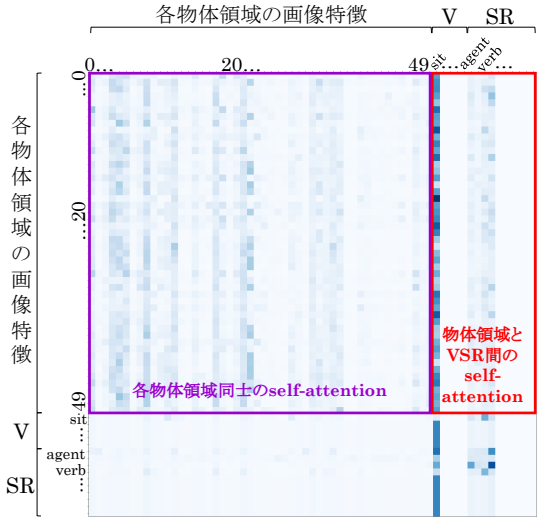
図 4.6: 図 4 のサンプル a について，source-target-attention マップの可視化例.



Meshed [5] における self-attention マップ

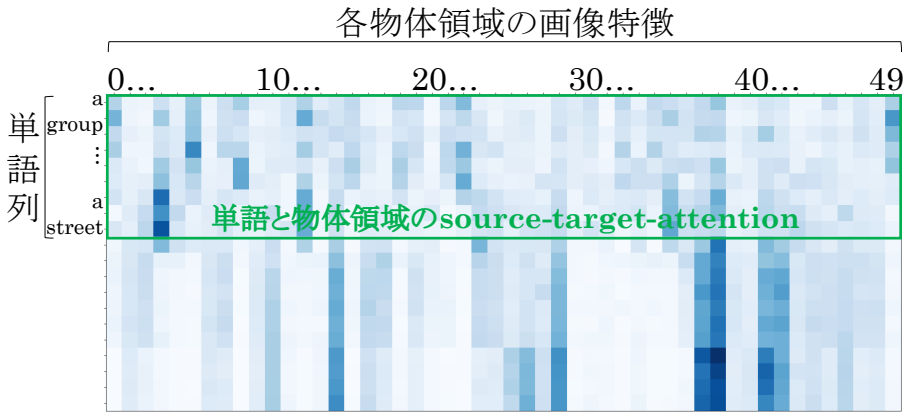


提案手法 (w/o SR-dec) における self-attention マップ

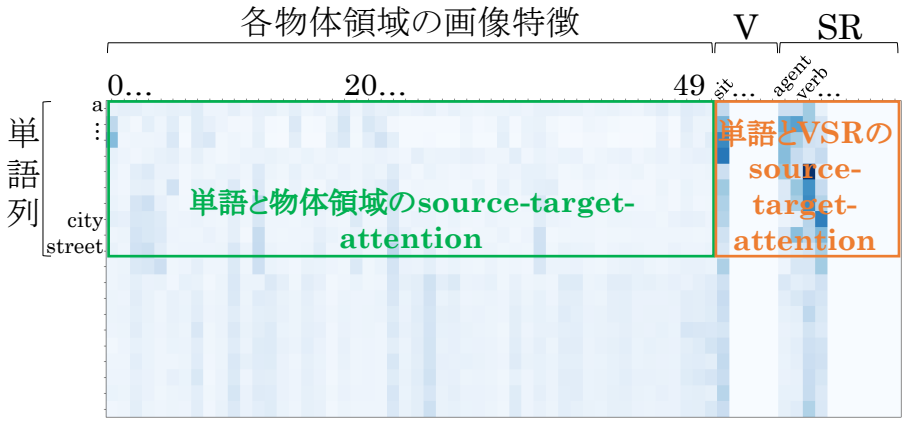


提案手法 (w/ SR-dec) における self-attention マップ

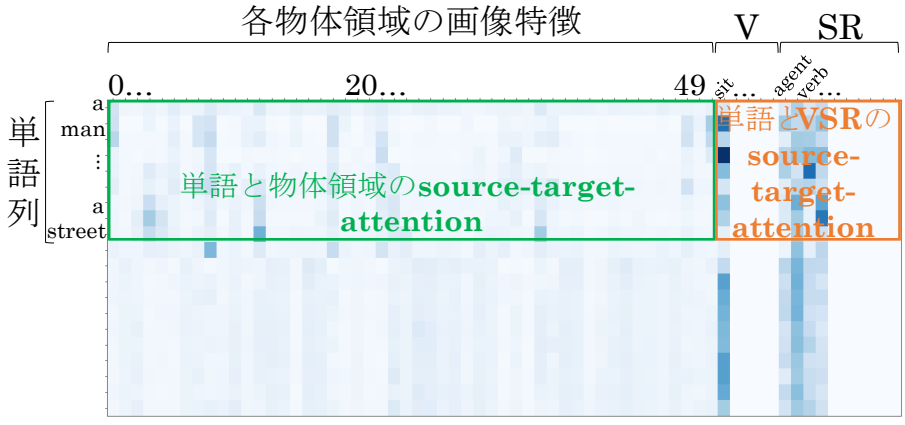
図 4.7: 図 4 のサンプル b について，self-attention マップの可視化例.



Meshed [5] における source-target-attention マップ



提案手法 (w/o SR-dec) における source-target-attention マップ



提案手法 (w/ SR-dec) における source-target-attention マップ

図 4.8: 図 4 のサンプル b について，source-target-attention マップの可視化例.

法と提案手法（w/ SR-dec と w/o SR-dec）により生成された説明文と推定された SR 系列の例を示す．SR 系列をサンプル画像の右側に赤字で記載し，説明文をサンプル画像下部に黒字で記載している．Meshed [5] は，入力が画像のみであり，説明対象を指定することができないため，異なる動作について説明する文を生成している．例えば，図 4.4 中のサンプル a では，“sit” という動作を，サンプル b では，“pass” という動作を説明していない．VSR-guided CIC の既存手法 [7] は，VSR により説明対象を指定しているため，正確に説明している例もあるが，異なる物体領域について説明する文を生成している例も確認できる．例えば，図 4.4 中サンプル c より，中央の kitchen table については触れず，局所的な “car” や “pots” についてのみ言及している．また，図 4.4 中サンプル d では，“ducks” について言及していない．提案手法 w/o SR-dec は，正確に説明文を生成している例も多いが，正解の SR 系列と異なる SR 系列を持つ説明文を生成している例が確認できる．例えば，図 4.4 中サンプル e では，推定した SR 系列に “locative” が無く，説明文中で “table” という場所について言及していない．提案手法 w/ SR-dec は，正確に説明文を生成している例は多いが，意味構造は正確だが正解文と異なる単語を使用する例が確認できる．例えば，図 4.4 中サンプル f では，SR 系列は正確だが，正解文中の “in the sky” を “through a clear blue sky” と言い換えている．提案手法では，説明文の意味構造を制御しながら，単語の多様性を許容していることが確認できた．一方で，画像中の物体や動作を誤認識してしまい，全ての手法において正確な説明文の生成が困難な例も確認できた．例えば，図 4.4 中サンプル g では，歯磨き粉（“toothepaste”）や泡立てている（“lather”）様子の認識が困難なため，すべての手法で正確な説明文を生成できていない．また，入力した動詞の使い方が複数考えられる場合，特定の使い方を指定できず正確な文を生成できない例も確認した．例えば，図 4.4 中サンプル h では，“look” という動詞を指定している．正解文では “look like (ように見える)” という使い方をしているが，提案手法の生成文では “look at (を見る)” という使い方をしている．上述の通り，提案手法においても，画像中の物体や動作を認識困難な場合や，複数の動詞の使い方が考えられる場合，正確な文を生成することが困難な場合がある．

次に，図 4.6 と図 4.8 は，図 4.4 中のサンプル a とサンプル b における，transformer encoder 内の入力特徴ベクトル間の関係の強さを算出した self-attention マップ（上段）と，transformer decoder 内においてエンコーダ部から受け取った中間特徴と直前まで推定された単語の特徴ベクトル間の相互の関係の強さを算出した source-target-attention マップ（下段）の可視化例を示す．self-attention マップは，縦横軸ともに，transformer encoder の入

力特徴ベクトルを示す．マップの左上を $(0,0)$ とするとき，Meshed [5] と提案手法（w/o SR-dec, w/ SR-dec）の self-attention マップについて，縦横軸の $0 \sim 49$ メモリが入力の各物体領域の特徴ベクトルを示す．提案手法（w/o SR-dec, w/ SR-dec）の self-attention マップについて，縦横軸の $50 \sim 54$ メモリが動詞 (V) の特徴ベクトル，縦横軸の $55 \sim 64$ メモリが意味役割ラベル (SR) の特徴ベクトルを示す．紫枠は各物体領域同士の関係の強さを，赤枠は各物体領域と VSR の関係の強さを可視化している．source-target-attention マップにおいて，縦軸は直前まで推定した単語列を，横軸は transformer encoder 内で重みづけされ抽出された中間特徴を示す．マップの左上を $(0,0)$ とするとき，Meshed [5] と提案手法（w/o SR-dec, w/ SR-dec）の source-target-attention マップについて，横軸の $0 \sim 49$ メモリは，各物体領域の特徴ベクトルが transformer encoder 内で重みづけされ抽出された中間特徴である．提案手法（w/o SR-dec, w/ SR-dec）の source-target-attention マップにおける $50 \sim 54$ が動詞 (V) の特徴ベクトル，$55 \sim 64$ メモリが意味役割ラベル (SR) の特徴ベクトルがそれぞれ transformer encoder 内で重みづけされ抽出された中間特徴である．source-target-attention マップの縦軸のメモリは上から順に直前までに推定した単語であり，文末以降はメモリを記載していない．緑枠は直前まで推定した単語と transformer encoder 内で重みづけされ抽出された各物体領域との関係の強さを，橙枠は直前まで推定した単語と transformer encoder 内で重みづけされ抽出された VSR との関係の強さを可視化している．transformer encoder 内の self-attention マップについて観察する．赤枠部分より，提案手法では物体領域と，制御信号の VSR との関係性が濃くなっている．このことから，提案手法の transformer encoder 内で，VSR と関係の強い物体領域が重みづけされていることが確認できる．紫枠部分において，提案手法の w/o SR-dec と w/ SR-dec を比較すると，w/ SR-dec の方が比較的濃くなっている．SR-dec を使用することで，transformer encoder 内において，物体領域同士の関係性も考慮していることが確認できる．source-target-attention マップについては，橙枠部分より，提案手法では，直前まで推定した単語と，制御信号の VSR との関係性が特に濃くなっている．次の単語を推定する際に，VSR を強く考慮していることが考えられる．緑枠部分より，単語と画像特徴間の関係性を見ると，Meshed [5] より提案手法の方が濃い部分が少なく，疎になっている．提案手法の transformer decoder 内で，次の単語を推定する際に注目する物体領域が絞られていると考えられる．上記の通り，transformer 内の attention マップを可視化することで，VSR との関係が強い物体領域に注目したり，VSR に基づいて次の単語に関係する物体領域を絞り込んでいる様子が確認され

た．また，SR-dec を導入することで，transformer encoder 内で VSR との関係が強い物体領域に注目しつつ，同時に物体領域間の関係性も考慮している様子が確認された．

4.5 まとめ

本章では，文の意味構造的規則として格構造ラベルを採用し，格構造ラベルの文生成への有効性について検証した．タスクとして画像キャプショニングを設定し，画像を入力として，画像に対応する正しくかつ柔軟な単語や語句を選択しつつ，画像を説明する文として正しい意味構造をもつ文の生成の実現を目指した．従来の格構造ラベルを使用する画像キャプショニング手法は，3 つの推論について独立した複数のモデルが用意され，それらの独立したモデルが逐次的に処理されていた．具体的には，一つ目のモデルは，説明対象である各 VSR に相当する物体が存在する物体領域を推定し，二つ目は説明文の意味構造を示す SR の順序を推定する．三つ目は推定された物体領域と順序付けられた SR から説明文を生成する．各モデルにおける推論誤りが説明文の精度劣化の要因となっていた．上記課題に対し我々は，画像の説明文生成と共に，説明対象となる物体領域の推定と，SR の順序の推定を同時に解く End-to-End VSR-guided CIC モデルを提案した．さらに，SR の正しい順序を推定するために，デコーダの入力として直前まで推定した SR を使用する，SR-guided captioning decoder（SR-dec）を提案した．2 つの CIC データセットでの実験結果は提案手法が既存手法よりも高精度に正確な文を生成可能であることを示し，提案手法の End-to-End 構造により，既存手法における物体領域の推論誤りや，SR 系列の推論誤りを低減することを確認した．アブレーション評価では，SR-dec が SR の順序推定の精度向上に寄与することを示し，提案する SR-dec が，より正確な意味構造の説明文の生成に効果があることを確認した．また，提案手法では，説明文の意味構造を制御しながら，単語の多様性を許容していることが確認できた．今後の課題として，画像内の物体や動作の誤認識を削減する方法の検討のため，他の物体検出モデルの利用や，meshed-memory transformer 以外のネットワーク構造をベースに実験を行うことが挙げられる．また，動詞の使い方を指定するために VSR に加え主語を制御信号に追加するなど，制御信号の更なる拡張の検討も挙げられる．

第5章　むすび

5.1　まとめ

本節では，まず，3 章で述べた文法構造的規則を用いた文生成と，4 章で述べた意味構造規則を用いた文生成について内容をまとめる．最後に，二つの構造的規則を用いた文生成について概観したまとめを述べる．

5.1.1　文法構造的規則を用いた自然文生成

文の文法構造的規則として文脈自由文法を採用し，文脈自由文法を適用して構築される構文木を生成する手法を提案し，文脈自由文法の文生成への有効性について検証した．タスクの設定として，文の主要な要素となる重要単語のセットを入力として，それらの単語を使用しながら，正確かつ柔軟な単語や語句の選択と，正しい文法構造を備えた文の生成を目指した．文脈自由文法を使用して構文木を構築する際，文法種類数が増えるほど構築され得る構文木は増大し，膨大な種類の構文木から状況に応じた適切な構文木を探索することが必要となる．効率的に適切な構文木を探索するために，モンテカルロ木探索アルゴリズムを使用した文脈自由文法に基づく文生成手法を提案した．モンテカルロ木探索とは，2 章で説明した通り，UCB1 値に基づいて「知識の適用（exploitation）」と，「探査（exploration）」のバランスを保ちながら効率的に探索するアルゴリズムである．さらに，モンテカルロ木探索を使用して適切な構文木を探索する際に，生成文が適切か評価する必要がある．本章ではタスクの設定として，状況を説明する際に主要な要素となる複数の単語（以降，“Situational Input” とする．例えば，{dog,eat,bread} である．）を入力情報として与え，これらの単語に基づき適切な文が生成されているか評価した．具体的には，生成文の評価方法として，単語の繋がりの正しさと生成内容の適切さの観点から評価した．さらに，提案手法ではモンテカルロ木探索を効率的に行うために，探索方針の設定方法と探索範囲の絞り込み方法を

提案した．探索方針の設定方法として，文の主要な要素から優先的に探索する方法を提案した．具体的には，文の述語や主語が文の主要な要素となるという考えのもと，述語を優先的に探索した後，述語よりも左の文法規則を適用することで優先的に主語を探索し，最後に述語よりも右の文法規則を適用した．探索範囲の絞り込み方法として，文の主要な要素となる重要単語のセット “Situational Input” への関連強さに基づいて文法規則のサンプリング確率を設定し，設定したサンプリング確率に基づいて語彙の絞り込みを行った．

実験では，与えられた文の主要な要素となる重要単語のセットに基づいた様々な文が生成されることを確認した．文法構造については，文頭から “主語”→“述語”→“目的語” の順番で正しく単語選択されていることが確認された．一方で，“主語”，“述語”，“目的語” の周辺で不自然な意味を持つ単語が選択され，文全体として意味が通らない例も確認された．また，動詞や名詞の変形が不自然な例も確認された．今後の課題として，重要単語のみならず，文全体としての意味を考慮した手法を検討し，文全体として意味的に自然な文を生成することが挙げられる．

5.1.2 意味構造的規則を用いた自然文生成

文の意味構造的規則として格構造ラベルを採用し，格構造ラベルの文生成へ活用する方法を検討し，格構造ラベルの文生成への有効性について検証した．タスクとして画像キャプショニングを設定し，画像を入力として，画像に対応する正しくかつ柔軟な単語や語句を選択しつつ，画像を説明する文として正しい意味構造をもつ文の生成の実現を目指した．従来の格構造ラベルを使用する画像キャプショニング手法は，3 つの推論について独立した複数のモデルが用意され，それらの独立したモデルが逐次的に処理されていた．具体的には，一つ目のモデルは，説明対象である各 VSR に相当する物体が存在する物体領域を推定し，二つ目は説明文の意味構造を示す SR の順序を推定する．三つ目は推定された物体領域と順序付けられた SR から説明文を生成する．各モデルにおける推論誤りが説明文の精度劣化の要因となっていた．上記課題に対し我々は，画像の説明文生成と共に，説明対象となる物体領域の推定と，SR の順序の推定を同時に解く End-to-End VSR-guided CIC モデルを提案した．さらに，SR の正しい順序を推定するために，デコーダの入力として直前まで推定した SR を使用する，SR-guided captioning decoder（SR-dec）を提案した．

2 つのデータセットでの実験結果は提案手法が既存手法よりも正確に文を生成可能であることを示し，提案手法の End-to-End 構造により，既存手法における物体領域の推論誤りや，SR 系列の推論誤りを低減することを確認した．アブレーション評価では，SR-dec が SR の順序推定の精度向上に寄与することを示し，提案する SR-dec が，より正確な意味構造の説明文の生成に効果があることを確認した．また，提案手法では，説明文の意味構造を制御しながら，単語の多様性を許容していることが確認できた．今後の課題として，画像内の物体や動作の誤認識を削減する方法の検討のため，他の物体検出モデルの利用や，meshed-memory transformer 以外のネットワーク構造をベースに実験を行うことが挙げられる．また，動詞の使い方を指定するために VSR に加え主語を制御信号に追加するなど，制御信号の更なる拡張の検討も挙げられる．

5.1.3 文の構造的規則に基づく自然文生成

本節では，文法構造的規則と意味構造的規則の文生成への有効性について検証した結果をまとめる．文法構造的規則は，3 章の検証結果より，文の意味の制御が不十分になる傾向がある．例えば，生成文例の中に，“another dog ate connecticut breads” や “another boy play overweight soccer” が含まれ，意味的に不自然な文が生成されることが確認された．また，動詞の時制や名詞の単数形・複数形などの単語の語形変化について制御が不十分になる傾向がある．例えば，“another dog eaten smallest breads” や “these woman wrote letter washington” が確認された．意味構造的規則は，4 章の検証結果より，正しい意味構造の生成文は，同時に正しい文法構造を持つ傾向が確認された．例えば，“a man sitting on a bench next to a white dog” は，正しい意味構造 “patient-verb(sit)-instrument-prediction” を持つと同時に，正しい文法構造 “主語-述語-副詞句” を持つ．意味構造を制御することができれば同時に正しい文法構造を持つことが可能である．

以上の検証結果より，意味構造的規則に基づく文の生成を行うことで，同時に文の文法構造を考慮することも可能であり，文法構造的規則に基づく文生成手法よりも比較的正確な文の生成を実現できると考えられる．

5.2 今後の課題

1 章で述べた通り，我々人間の生活・仕事の効率化や人間の能力拡張のために，機械に人間同様の文生成の能力を実装することを目的としたとき，本論文で検証した手法の到達点との差について考察し，今後解決されるべき課題について述べる．

まず，本論文で検証した文生成手法の到達点について述べる．文法構造的規則に基づく手法においては，主要な単語セットを入力として，それらの単語を使った文長が 3 8 の単一の文を生成した．意味構造的規則に基づく手法においては，画像を入力として，画像を説明する単一の文を生成した．文長制約は設定していないが，平均約 10 程度の長さの文が生成された．文法構造的規則と意味構造的規則に基づく両文生成手法とも，文の構造規則に基づく文長 10 程度の短い単一文の生成が実現されたと捉えることができる．

一方で，人間の文生成の能力を代替するために，求められる文生成能力の到達点について考える．例えば複雑な作業について他人に説明するための文を生成するとき，単一文ではなく複数文になる可能性がある．カレーの作り方を説明するためには，“にんじんを一口大に切る”，“玉ねぎをくし切りに切る”，“鍋を強火で温め，鍋に油をしく” などの複数文で説明する必要がある．このとき，“鍋を強火で温め，鍋に油をしく” の文よりも前に，“にんじんを一口大に切る” や “玉ねぎをくし切りに切る” などの手順を説明する必要があり，文同士の関係性を考慮しながら順番に並べ，一つのレシピが作成される．人間同様の文生成の能力を代替するためには，文と文の関係性を考慮しながら文を適切な順番に並べた文書を作成する能力が必要だと考えられる．

また，単一文で説明する内容は，10 以上の単語が必要になる可能性がある．例えば，自分の経験を他人に説明するための文を生成するとき，“When I was a student, I studied natural language processing, but after joining the company, I started researching image processing as a specialty.” など，文長 10 以上の長文となる可能性がある．このとき，文中の単語間の関係性について，より長期的な単語間の関係性を考慮する必要がある．4 章で使用した Transformer は長期的な文脈考慮に適したネットワーク構造であるが，本論文における検証では文長 10 程度の文の生成の検証にとどまっている．より長い文の文脈を考慮する必要があり，単一文中の文脈の考慮について，既存手法の到達点の検証と課題の整理をする必要がある．

上述より今後の課題として，より文長の長い文の生成のために，より長期的な単語間の関係性の考慮の検討と，文と文の関係性を考慮する手法の検討が挙げられる．

謝辞

本論文は，お茶の水女子大学，人間文化創成科学研究科，理学専攻，情報科学コース博士前期課程及び博士後期課程において筆者が行なった研究をまとめたものです．この研究は，多くの先生方からの御指導や，小林研究室の先輩・同輩・後輩，その他の友人達による数多くの御支援，御協力なくしては，本論文の完成には決して至りませんでした．この場を借りて，皆様への謝意を表します．まず最初に，本論文の主査であり，日頃より多大なる御指導，御鞭撻をいただいたお茶の水女子大学院人間文化創成科学研究科理学専攻情報科学領域・小林一郎教授に深く感謝致します．御多忙な中，研究の方向性について相談する時間を多く割いていただきました．私が学部 4 年の頃，研究室に配属した時から約 8 年間，研究のご指導を頂き，課題設定，解決策検討，実験，そして論文の執筆及び発表といった研究に関するすべての局面において，常に的確なご助言をいただけたことで，研究者にとって不可欠なことを多く学ばせていただくことができました．また，過去から現在に至るまで，常に人工知能・言語処理分野の第一人者として活躍し続けるその御姿は，私が目標とすべきものだと常に感じており，日々研究に励む原動力の 1 つとなっております．長きにわたりご指導いただき，誠にありがとうございました．心より深謝致します．また，本論文の審査の過程において，お茶の水女子大学の吉田教授、小口教授、伊藤教授、浅井教授には数々の御助言と御指導を賜りました．大変お忙しい中，本論文の副査をお引き受け下さいましたこと深く感謝致します．日々のディスカッション，ミーティング等を通して数多くの御助言をいただいた小林研究室の諸氏に深く感謝いたします．また，日頃から研究について議論させていただいているだけでなく，社会人ドクターとして大学に通うことを快く受け入れ，支えていただいた日本電信電話株式会社 NTT 人間情報研究所の諸氏に深く感謝いたします．最後に，今日に至るまでさまざまな面で支えていただいた，家族，友人に深く感謝致します．

参考文献

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

[2] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[3] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[4] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

[5] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020.

[6] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*, 2019.

[7] Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. Human-like controllable image captioning with verb-specific semantic roles. In *CVPR*, 2021.

[8] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[10] 自然言語処理: 基礎と応用. 電子情報通信学会 (コロナ社), 1999.

[11] 確率的言語モデル. 言語と計算. 東京大学出版会, 1999.

[12] J.E. Hopcroft and J.D. Ullman. *Formal Languages and Their Relation to Automata*. Addison-Wesley Series in Computer Science. Addison-Wesley Publishing Company, 1969.

[13] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, Vol. 19, No. 2, pp. 313–330, 1993.

[14] 野村浩郷, 内藤昭三ほか. 計算言語学: 自然言語理解における意味表現. 情報処理, Vol. 27, No. 8, 1986.

[15] 村木一至, 亀井真一郎, 野村直之ほか. 機械翻訳システムの中間言語. 情報処理学会研究報告自然言語処理 (NL), Vol. 1989, No. 54 (1989-NL-073), pp. 99–106, 1989.

[16] G. Chaslot, S.C.J. Bakkes, I. Szita, and P.H.M. Spronck. Monte-Carlo tree search: A new framework for game AI. *Proceedings of the BNAIC 2008, the twentieth Belgian-Dutch Artificial Intelligence Conference*, pp. 389–390, 2008.

[17] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, Vol. 47, pp. 235–256, 2002.

[18] R. Kneser and H. Ney. Improved backing-off for n-gram language modeling. In *ICASSP*, Vol. 1, pp. 181–184, 1995.

[19] Stuart M. Shieber. A uniform architecture for parsing and generation. In *Proceedings of the 12th Conference on Computational Linguistics - Volume 2*, COLING '88, pp. 614–619, Stroudsburg, PA, USA, 1988. Association for Computational Linguistics.

[20] Martin Kay. Chart generation. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 200–204, Santa Cruz, California, USA, June 1996. Association for Computational Linguistics.

[21] Irene Langkilde-Geary. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the International Natural Language Generation Conference*, 2002.

[22] Wei Lu, Hwee Tou Ng, and Wee Sun Lee. Natural language generation with tree conditional random fields. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pp. 400–409, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[23] Tsung-Hsien Wen, Milica Gašić, Dongho Kim, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Stochastic Language Generation in Dialogue using Recurrent Neural Networks with Convolutional Sentence Reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, September 2015.

[24] Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, September 2015.

[25] Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. Multi-domain neural network language generation for spoken dialogue systems. In *Proceedings of the 2016 Conference on North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics, June 2016.

[26] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Nat. Lang. Eng.*, Vol. 3, No. 1, pp. 57–87, 1997.

[27] A. Koller and M. Stone. Sentence generation as a planning problem. In *International Natural Language Generation Workshop*, Vol. 12, pp. 17–24, 2007.

[28] D.Bauer and A. Koller. Sentence generation as planning with probabilistic ltag. In *Proceedings of the 10th International Workshop on Tree Adjoining Grammar and Related Formalisms*, 2010.

[29] Avrim L. Blum and Merrick L. Furst. Fast planning through planning graph analysis. *ARTIFICIAL INTELLIGENCE*, Vol. 90, No. 1, pp. 1636–1642, 1995.

[30] Nathan McKinley and Soumya Ray. A decision-theoretic approach to natural language generation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 552–561, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[31] Jonathan Pfeil and Soumya Ray. Scaling a natural language generation system. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1148–1157, Berlin, Germany, August 2016. Association for Computational Linguistics.

[32] J. H. Lau, A. Clark, and S. Lappin. Unsupervised prediction of acceptability judgements. In *ACL 2015*, Vol. 53, pp. 15–1000, 2015.

[33] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, January 2003.

[34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[35] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE TPAMI*, 2016.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[39] Sainbayar Sukhbaatar, Edouard Grave, Guillaume Lample, Herve Jegou, and Armand Joulin. Augmenting self-attention with persistent memory. *arXiv*, 2019.

[40] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *NeurIPS*, 2019.

[41] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *ICCV*, 2019.

[42] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019.

[43] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

[44] Annika Lindh, Robert J Ross, and John D Kelleher. Language-driven region pointer advancement for controllable image captioning. In *COLING*, 2020.

[45] Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. Length-controllable image captioning. In *ECCV*, 2020.

[46] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2015.

[47] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[48] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[49] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[51] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005.

[52] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL workshop*, 2004.

[53] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[54] Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv*, 2019.

[55] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*.

[56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*.

[57] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.

研究業績

本論文に関する研究業績

学術論文誌

- Kaori Kumagai, Ichiro Kobayashi, Daichi Mochihashi, Hideki Asoh, Tomoaki Nakamura, Takayuki Nagai: “Natural Language Generation Using Monte Carlo Tree Search,”Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.22, No.5, pp. 777-785, 2018.
- 熊谷 香織, 高木 基宏, 近藤 重邦, 青野 裕司, 小林 一郎”動詞固有意味役割ラベルを使用した制御可能な End-to-End 画像キャプショニング” 情報処理学会論文誌，Vol.63, No.12, Dec. 2022.

国際会議

- Kaori Kumagai, Ichiro Kobayashi, Daichi Mochihashi, Hideki Asoh, Tomoaki Nakamura, Takayuki Nagai: ”Human-like Natural Language Generation Using Monte Carlo Tree Search,” Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation.

研究会・シンポジウム等

- 熊谷香織，持橋大地，小林一郎，麻生英樹，Muhammad Attamimi，中村友昭，長井隆行”モンテカルロ木探索を用いた確率文脈自由文法に基づくテキスト生成，” 第 21 回言語処理学会年次大会，2015.
- 熊谷 香織 持橋 大地 小林 一郎 麻生 英樹 Attamimi Muhammad 中村 友昭 長井 隆行”モンテカルロ木探索を用いた確率文脈自由文法に基づくテキスト生成，” 2015 年度人工知能学会全国大会（第 29 回），2015.

- 熊谷 香織 持橋 大地 小林 一郎 麻生 英樹 Attamimi Muhammad 中村 友昭 長井 隆行”モンテカルロ木探索を用いた統語情報を考慮した文生成，” 第 22 回言語処理学会年次大会，2016.
- 熊谷 香織 持橋 大地 小林 一郎 麻生 英樹 Attamimi Muhammad 中村 友昭 長井 隆行”モンテカルロ木探索を用いた構造的正しさと言語モデルを考慮した文生成，” 2016 年度人工知能学会全国大会（第 30 回），2016.
- 熊谷香織, 小林一郎, 持橋大地, 麻生英樹, 中村友昭, 長井隆行”モンテカルロ木探索を用いた構文木構築に基づく頑健な文生成，” 第 23 回言語処理学会年次大会，2017.
- 熊谷 香織, 高木 基宏, 近藤 重邦, 青野 裕司, 小林 一郎”動詞固有意味役割ラベルを使用した制御可能なエンドツーエンド画像キャプショニング” 第 25 回画像の認識・理解シンポジウム（MIRU2022），2022.

その他の研究業績

学術論文誌

- 熊谷 香織, 渡邉 之人, 細野 峻司, 早瀬 和也, 島村 潤”属性の一貫性を考慮したカテゴリ非依存な物体画像変換” 電子情報通信学会論文誌, Vol.J104-D, No.02, pp.130-142, Feb. 2021.

国際会議

- Kaori Kumagai, Yukito Watanabe, Takashi Hosono, Jun Shimamura, Atsushi Sagata ”Category Independent Object Transfiguration with Domain Aware GAN” Proceedings of Asian Conference on Pattern Recognition (ACPR), Nov. 2019.

研究会・シンポジウム等

- 熊谷香織, 渡邉之人, 田良島周平, 島村　潤, 杵渕哲也, CNN の中間層における高次特徴表現の統合による可視化手法の検討 2018 年電子情報通信学会総合大会, 2018.
- 熊谷 香織, 渡邉 之人, 細野 峻司, 島村 潤, 嵯峨田 淳”ドメイン領域と強度の明示的推定に基づく未知物体画像変換” 第 22 回画像の認識・理解シンポジウム（MIRU2019）, 2019.
- 熊谷香織, 梅田崇之, 入江豪, 北原正樹, 島村潤, 小林一郎”人と物体との関係性に基づくマルチモーダル物体候補領域推定手法の検討” 第 24 回画像の認識・理解シンポジウム（MIRU2021）, 2021.
- 熊谷香織, 佐藤禎哉, 高木基宏, 近藤重邦, 青野裕司”画像の状況を示すキャプションを用いた物体検出手法の検討” 映像情報メディア学会 2021 年冬季大会, 2021