一般論文

Peer-Reviewed

オープンドメイン手順書のフローグラフ予測

白井　圭佑[†]・亀甲　博貴[††]・森　　信介[††]

機械による手順書理解は，文章中の手順に関する推論やこれらを元にした作業の自動化に必須である．先行研究では調理分野に焦点を当て，調理レシピの理解の表現としてレシピフローグラフ (recipe flow graph; r-FG) を提案し，そのアノテーションを作成した．r-FG は手順に関わる表現をノードとし，それらの関係をエッジとする有向非巡回グラフとして定義される．先行研究では，r-FG の自動予測のフレームワークとして，ノード予測とエッジ予測の 2 段階で行うものが提案されている．一方で，r-FG は調理分野に依存した表現となっており，調理以外の分野には適用されてこなかった．本論文では，一般的な手順書の理解の表現として wikiHow フローグラフ (wikiHow flow graph; w-FG) を提案する．w-FG は r-FG と互換性があり，既存の r-FG のアノテーションは w-FG に自動変換可能である．w-FG を用いて一般的な分野の手順書のフローグラフ予測精度を調査するために，wikiHow の記事を基に新たなコーパスである w-FG コーパスを構築する．実験では，調理分野から対象分野への分野適応を行うことで，ノード予測を 75.0% 以上，エッジ予測を 61.8% 以上の F 値で行えることを示す．

キーワード：手順書理解，フローグラフ表現，固有表現認識，依存構造解析

Towards Flow Graph Prediction of Open-Domain Procedural Texts

Keisuke Shirai[†], Hirotaka Kameko[††] and Shinsuke Mori[††]

Comprehension of procedural texts by machines is essential for reasoning about the steps in the texts and automating the procedures by robots. Previous work has focused on the cooking domain and proposed a recipe flow graph (r-FG) to represent an understanding of recipe texts with annotations. r-FG is defined as a directed acyclic graph with expressions related to procedures as nodes and the relationships between the nodes as edges. Previous work has proposed a framework that predicts r-FG representations in two steps: node prediction and edge prediction. While such advances have developed, the idea has only been applied to the cooking domain. This work proposes a wikiHow flow graph (w-FG) to represent an understanding of open-domain procedural texts. w-FG is compatible with r-FG, and the existing r-FG

[†] 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University
[††] 京都大学学術情報メディアセンター, Academic Center for Computing and Media Studies, Kyoto University
本論文は言語処理学会第 29 回年次大会 (白井 et al. 2023) および The 8th Workshop on Representation Learning for NLP (RepL4NLP 2023) (Shirai et al. 2023) で発表した 2 本の論文を基にしたものである．

annotations in the cooking domain can be automatically converted into those in w-FG. We introduce a novel dataset called the w-FG corpus from wikiHow articles to evaluate flow graph prediction accuracy in domains other than cooking. Experimental results show that domain adaptation from the cooking to the target domain enables predictions of nodes with more than 75.0% accuracy and edges with more than 61.8%.

Key Words: *Procedural Text Understanding, Flow Graph Representation, Named Entity Recognition, Dependency Parsing*

1 はじめに

手順書は調理や家具の組み立て等，幅広いタスクを実行するための手順のリストを提供する．手順は複数文から構成されることもあり，各文には通常一つ以上の動作と物体が含まれる．近年では，手順書の理解に向けた研究が盛んに行われている (Mori et al. 2014; Kiddon et al. 2015; Bosselut et al. 2018; Dalvi et al. 2018; Tandon et al. 2020)．この中でも，文章全体における手順の流れを理解すること (Mori et al. 2014; Kiddon et al. 2015) は，手順間の関係に関する推論 (Zhang et al. 2020b) や手順書を基にした作業の自動化 (Bollini et al. 2013) を目指す上で重要である．この方向において，先行研究では，調理レシピの理解の表現としてレシピフローグラフ (recipe flow graph; r-FG) がコーパスと共に提案されている (Mori et al. 2014; Yamakata et al. 2020)．図 1 の左図に示すように，r-FG は調理レシピ内の手順に関わる表現をノード，それらの関係をエッジとする有向非巡回グラフとして定義され，文章レベルでの手順の依存関係を捉えることが出来るという特徴を持つ．また，先行研究では，r-FG の自動予測を行うフレームワークとして，ノード予測とエッジ予測の 2 段階で行うものが提案されている (Maeta et al. 2015)．こうした発展がある一方で，r-FG は調理分野に依存した表現となっているため，その他の分野の手順書には未だ適用されていない．調理分野に依存しない一般化されたフローグラフ表現を開発することは，分野間で手順の知識の共有を可能にする上で意義があるといえる．また，フローグラフに共通する問題として，アノテーションが複雑であり，大規模なアノテーションが現実的ではない点が挙げられる．そのような際に，既存のアノテーションを活用し，新たな分野では少量アノテーションのみを用意して予測モデルを学習できれば，アノテーションコストの削減に繋がり有用である．

本論文では一般的な手順書の理解の表現として，wikiHow フローグラフ (wikiHow flow graph; w-FG) を提案する．これは r-FG における調理依存の表現（食材; Food）を手順書の最終生産物の材料 (Ingredient) として汎化して捉えることで得られる．w-FG は r-FG と互換性があり，既存の調理分野のデータは w-FG に変換可能である．ここでは，英語の手順書を対象とし，調理以外の分野における手順書のフローグラフ予測性能を調査する．この目的のため，新たに w-FG コーパスを構築する．これは，様々なタスクの手順を公開している wikiHow の記事を基に作成さ

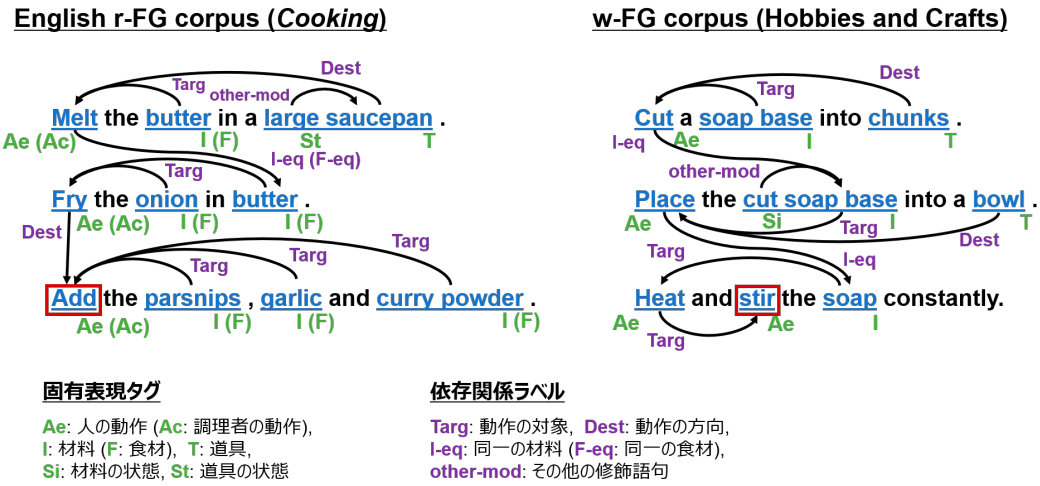

図 1　フローグラフアノテーションの例．左は English r-FG コーパスにおけるレシピのアノテーションを，右は w-FG コーパスの *Hobbies and Crafts* の手順書のアノテーションを，それぞれ示す．赤線で囲まれている表現は，フローグラフにおけるルートノードに対応している．

れている．w-FG コーパスは，wikiHow の上位カテゴリである *Food and Entertaining*，*Hobbies and Crafts*，*Home and Garden*，*Cars & Other Vehicles* を分野として選択し，各分野で 30 記事のアノテーションを提供する．w-FG コーパスは Web 上で公開済みである[1]．実験では，w-FG コーパスの各対象分野において，ノード予測とエッジ予測の性能を調査する．ここでは，フローグラフアノテーションのコストを考慮し，対象分野の学習に利用可能なデータが小規模であると想定する．この設定下でフローグラフ予測性能を向上するため，既存の調理分野のデータの利用を考える．具体的には，調理分野のデータである English r-FG コーパス (Yamakata et al. 2020) で事前学習を行い，w-FG コーパスの対象分野のデータでファインチューニングを行う分野適応によって，予測モデルを実現する．実験結果では，このような分野適応を行うことで，English r-FG コーパスか w-FG コーパスのいずれか一方を学習に用いる場合に比べ，大幅な性能向上が実現出来ることを示す．

2　レシピフローグラフ

本節では，レシピフローグラフ (recipe flow graph; r-FG) について概説する．まず，r-FG 表現について説明を行い，次に r-FG 予測のためのフレームワークを紹介する．

[1] https://github.com/kskshr/w-FG-Corpus

2.1　フローグラフ表現

図 1 の左図のように，r-FG はノードの集合 V とエッジの集合 E から成る有向非巡回グラフ $G(V, E)$ として定義される．ここで，V は食材や道具等の手順に関わる表現の集合であり，E はノード同士の依存関係を表すラベル付きエッジの集合である．グラフは連結であり，全手順の結果として得られる最終生産物に対応する特殊なルートノードが存在する．先行研究 (Yamakata et al. 2020) では，手順書の最後に登場した調理動作の表現をルートノードとしている．現在，r-FG のアノテーションとしては日本語のコーパス (Mori et al. 2014) と英語のコーパス（English Flow Graph コーパス; English r-FG コーパス） (Yamakata et al. 2020) が公開されている．ここで，r-FG は日本語のレシピを対象にデザインされており，英語 r-FG では英語特有の表現を扱うために少数のタグとラベルを新たに追加している．表 1 と表 2 に示す通り，英語 r-FG は 10 種類の固有表現タグと 13 種類の依存関係ラベルを用いる．本論文では英語の手順書を対象とするため，以降では英語 r-FG を主に扱う．

2.2　フローグラフ予測

先行研究では，ノード予測とエッジ予測の 2 段階による，r-FG 自動予測のためのフレームワークが提案されている (Maeta et al. 2015).

ノード予測では記事中のノードに対応する表現をタグ付きで検出する．これは系列ラベリング問題として定式化され，固有表現認識器を用いて予測を行う．固有表現認識では文レベルでタグ予測を行うのが一般的であるが (Lample et al. 2016)，先行研究では記事レベルで予測を行っ

固有表現タグ	意味
I (F)	材料（食材）
T	道具
D	継続時間
Q	分量
Ae (Ac)	人（調理者）による動作
Ae2 (Ac2)	不連続の動作
Ai (Af)	材料（食材）による動作
At	道具による動作
Si (Sf)	材料（食材）の状態
St	道具の状態

表 1　固有表現タグとその意味．括弧内は英語 r-FG におけるタグと意味を表す．

依存関係ラベル	意味
Agent	主語
Targ	動作の対象
Dest	動作の方向
T-comp	道具による補足
I-comp (F-comp)	材料（食材）による補足
I-eq (F-eq)	同一の材料（食材）
I-part-of (F-part-of)	材料（食材）の一部
I-set (F-set)	材料（食材）の集合
T-eq	同一の道具
T-part-of	道具の一部
A-eq	同一の動作
V-tm	動作のタイミング
other-mod	その他の修飾語句

表 2　依存関係ラベルとその意味．括弧内は英語 r-FG におけるラベルと意味を表す．

ており (Yamakata et al. 2020)，本研究もそれに従う[2].

エッジ予測ではノード間の依存関係をラベル付きで予測する．これは最大全域木問題として，以下のように定式化される．

$$\hat{G} = \underset{G \in \mathcal{G}}{\operatorname{argmax}} \sum_{(u,v,l)} s(u,v,l). \tag{1}$$

ここで，$s(u,v,l)$ はノード u からノード v へ，ラベル l が振られる場合のスコアを表しており，これは Chu-Liu-Edmonds アルゴリズムを用いて解かれる (Chu and Liu 1965; Edmonds 1967). ラベル付きエッジのスコアは依存構造解析器を用いて計算される (McDonald et al. 2005).

3　wikiHow フローグラフ

wikiHow フローグラフ (wikiHow flow graph; w-FG) は英語 r-FG を拡張したフローグラフ表現である．表 1 と表 2 に，w-FG の固有表現タグと依存関係ラベルのリストを示す．r-FG においては，原材料，手順の中間生産物はすべて食材 (Food) として扱われ，これらは手順の最後には最終生産物である料理に組み込まれる．例えば，サラダの調理レシピを考えたとき，最終生産物はサラダであり，レタスやドレッシングはサラダの材料として捉える．w-FG では，r-FG で扱う食材を，手順書一般で得られる最終生産物の材料 (Ingredient) として汎化して捉えることで，調理以外の分野の手順書を扱えるようにする．例えば，机の組み立てを考えたとき，最終生産物は机であり，机の脚やネジは机の材料として捉える．w-FG では固有表現タグの食材 (F) を材料 (I) に変更するほか，その他の食材に関わる固有表現タグや依存関係ラベルは，食材の部分を材料に置き換えて用いる（例えば，食材の状態 (State of foods; Sf) → 材料の状態 (State of ingredients; Si)，同一の食材 (Food equality; F-eq) → 同一の材料 (Ingredient equality; Ie)）. w-FG は r-FG との互換性があるため，既存の r-FG アノテーションは w-FG に変換して用いることが可能である．以降では，w-FG を用いた一般的な手順書からのフローグラフ表現の予測を考える．

3.1　フローグラフ予測

先行研究と同様に，フローグラフ予測はノード予測とエッジ予測の 2 段階で行う．予測モデルは w-FG アノテーションを用いて教師あり学習を行うことで得られる．ここで，フローグラフのアノテーションは高コストであるため，学習に利用可能な対象分野のデータは少量であると想定する．一方で，既存の調理分野のコーパスも w-FG 表現に変換可能であり，学習データとして用いることができる．従って，本論文では，低リソース問題を解消するために，モデル

[2] 予備実験では，記事レベルで予測を行うことで文レベルの場合と比較して予測精度が 10%向上することを確認した.

の学習に調理分野のデータと 4 節で述べる対象分野のデータの双方を用いる．具体的には，調理分野のデータでモデルを事前学習し，対象分野のデータでファインチューニングする分野適応を行う．本章の残りの部分では，3.2 節でタスクの定式化を行い，次に 3.3 節で低リソース下での性能向上のためのデータ拡張を考える．

3.2　タスク定義

調理分野における N 例のフローグラフ $(V_1^C, E_1^C), \cdots, (V_N^C, E_N^C)$ と対象分野における M 例のフローグラフ $(V_1^T, E_1^T), \cdots, (V_M^T, E_M^T)$ が与えられたとき，本タスクの目標は対象分野においてノード予測モデル F_{Node} とエッジ予測モデル F_{Edge} の予測性能を最大化することである．ここで，

$$F_{\text{Node}} : D \rightarrow V, \tag{2}$$

$$F_{\text{Edge}} : (D, V) \rightarrow E \tag{3}$$

であり，ここで D は手順書である．本設定では M は非常に小さい数であるため（実験では $M = 5$），本タスクは低リソース分野適応 (Xu et al. 2021) の側面を持つといえる．また，調理分野のデータ或いは対象分野のデータを用いない場合は，それぞれ few-shot, zero-shot の設定と見なすことが出来る．

3.3　データ拡張

低リソース設定における性能向上のためのアプローチとして，データ拡張は有望なアプローチである (Fadaee et al. 2017; Ding et al. 2020)．本研究では，手順の入れ替えと単語置換の 2 種類のデータ拡張を考える．

手順の入れ替え　では，図 2 のように，文章中の任意の 2 つの手順を入れ替えることでデータ拡張を行う．しかし，手順をランダムに入れ替えた場合，手順の順序関係に矛盾が生じる可能性がある．例えば，調理レシピにおいて，“1. Cut the potatoes.” と “2. Add the potatoes to the

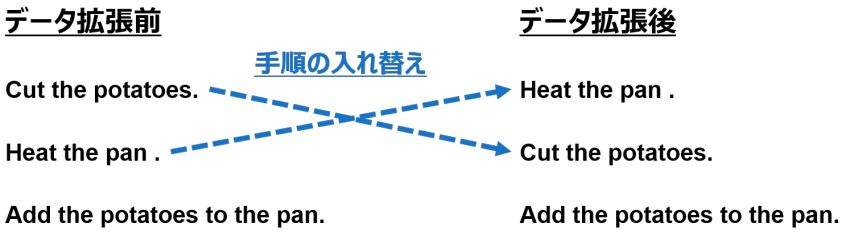

図 2　手順入れ替えの例．

pan.” という 2 つの手順は入れ替えることが出来ない．本論文では，フローグラフアノテーションを用いて手順の順序関係を考慮することで，この制約を守りつつデータ拡張を行う[3]．

単語置換　では，手順中のある単語を任意の単語に置換することでデータ拡張を行う．例えば，“Heat the pan .” の “pan” を “cooking_pan” に置換することで拡張する．しかし，無関係な単語に置換した場合，手順の意味が大きく変化する可能性がある．従って，本論文では WordNet (Dai and Adel 2020) を用いて，単語を類義語に置換する．常に単語の置換を行うわけではなく，$p(0 \leq p \leq 1)$ の確率で置換を行う．単語によっては複数の類義語が存在しうるが，その場合は候補の中からランダムに一つを選択する．単語の置換対象として，本論文では人の動作 (Ae)，材料 (I)，道具 (T) の固有表現タグに対応する表現を対象とする．

4　w-FG コーパス

w-FG コーパスは，wikiHow[4]上の記事から構築した新たなフローグラフコーパスである．wikiHow は 23 万を超える手順書を公開しており，近年では手順書の言語資源として先行研究で広く用いられている (Zhou et al. 2019; Zellers et al. 2019; Zhang et al. 2020b; Zhou et al. 2022; Lin et al. 2022)．以下では，データ収集，アノテーション手順，統計量，アノテーション一致率について順に説明する．

4.1　データ収集

対象分野として，wikiHow の上位カテゴリから *Food and Entertaining*, *Hobbies and Crafts*, *Home and Garden*, *Cars & Other Vehicles* の 4 つを選択した．ここで，*Food and Entertaining* は調理を主なタスクとする点で，English r-FG コーパスに非常に近い分野であるといえる．*Hobbies and Crafts* は工作を主に扱う分野であり，調理とは異なるが，材料を組み立てて最終生産物を得るという点で共通している．*Home and Garden* と *Cars & Other Vehicles* はそれぞれ園芸や乗り物の整備を主なタスクとしており，組み立て以外のタスクを多く含むため，他 2 分野と比較してより多様な手順を扱う分野であるといえる．

wikiHow コーパス (Zhang et al. 2020b) から wikiHow の記事を，分野ごとに 30 記事収集した．表 3 に分野ごとの記事タイトルの例を示す．このとき，低品質な記事を取り除くために，(i) 記事全体で 25 単語以上であり，(ii) ユーザからの評価が 50%以上である記事のみを収集した．また，タスクが曖昧なものや最終生産物が物体でないものに関しては，人手で除外した．ここで

[3] 実際には，“Heat the pan.” と “Add the potatoes to the pan.” の間の作業時間が変化し，温めたフライパンに影響があるため，このような手順間には暗黙的な順序の制約が存在するといえる．本論文では，このような暗黙的な制約は無視し，フローグラフによって与えられる明示的な制約のみに着目する．

[4] `https://www.wikihow.com`

は先行研究 (Zhang et al. 2020b; Zhou et al. 2022) に従い，段落の見出しを手順として利用し，フローグラフアノテーションを行った．

4.2 アノテーション手順

wikiHow 記事のアノテーションはアノテータに依頼して行った．アノテータのトレーニングとして，まず English r-FG コーパスからランダムに収集した 10 レシピを用いて，コーパス中のアノテーションとの一致率が 80%を超えるようにアノテーションの指示を行った．その後，wikiHow 記事を用いてより細かいアノテーション仕様を説明し，その後 120 記事のアノテーションを行った．アノテーションには，フローグラフアノテーションツールを用いた (Shirai et al. 2022)．また，アノテーションの前に，Stanza (Qi et al. 2020) を用いて手順を単語列に分割した．アノテーションには計 40 時間を要した．

4.3 統計量

まず，表 4 に w-FG コーパスの統計量を示す．各記事は平均 7.1 手順から成り，各手順は平均 10.5 単語から構成されていることがわかる．また，記事ごとに平均 30.6 個の固有表現タグと依存関係ラベルがアノテーションされていることがわかる．ここで，*Home and Garden* と *Cars & Other Vehicles* は他 2 分野に比べ記事中の単語数が少なく，それに伴いアノテーションされたタグ数とラベル数も少なくなっている．

次に，表 5 にタグごとのアノテーション数を示す．これより，Ae，I，T は分野を通して出現頻度が高いことがわかる．At は今回のアノテーションでは現れなかったが，これは English r-FG コーパスにおいても 15 件と出現頻度が非常に低い．English r-FG と同様に調理を扱う *Food and*

分野	記事タイトルの例
Food and Entertaining	*Cooking acorn squash*, *Making lavender tea*, *Baking a cherry pie*
Hobbies and Crafts	*Making a bar soap*, *Making a duct tape bow*, *Making a paper box*
Home and Garden	*Cleaning a mattress pad*, *Installing a microwave*, *Making a scented candle*
Cars & Other Vehicles	*Fixing a slipped bike chain*, *Cleaning car window*, *Cleaning tail lights*

表 3 分野ごとの記事タイトルの例．

分野	文字数	単語数	手順数	タグ数	ラベル数
Food and Entertaining	10,167	2,761	224	1,123	1,127
Hobbies and Crafts	9,407	2,556	247	1,048	1,059
Home and Garden	7,700	2,010	205	887	882
Cars & Other Vehicles	6,432	1,622	173	613	609

表 4 w-FG コーパスの統計量．

タグ	*Food and Entertaining*	*Hobbies and Crafts*	*Home and Garden*	*Cars & Other Vehicles*
I	380	419	250	218
T	136	56	186	91
D	41	9	17	4
Q	73	46	32	14
Ae	315	310	270	202
Ae2	15	38	19	23
Ai	28	22	17	11
At	0	0	0	0
Si	84	147	42	35
St	60	15	61	27
合計	1,132	1,062	894	625

表 5　固有表現タグごとのアノテーション数.

Entertaining が計 30 記事のみであり，r-FG コーパス（300 記事）の 10 分の 1 の規模であることを考慮すると，これは妥当な数値であるといえる.

各分野における Ae，I，T がアノテーションされた上位 10 件の表現の頻度分布を図 3 に示す. これより，Ae に関しては，Food and Entertaining と Hobbies and Crafts の両分野では “add” や “cut” が，Home and Garden と Cars & Other Vehicles では “remove” や “use” が多用されていることがわかる．また，I においては，全分野で代名詞 “it” の頻度が高いこともわかる．さらに，I や T の高頻度の表現には，各分野で材料や道具として用いられやすい物体の特色が現れているといえる.

表 6 にラベルごとのアノテーション数を示す．Targ，Dest，I-eq，other-mod は全分野を通して高頻度であることがわかる．また，*Home and Garden* と *Cars & Other Vehicles* では T-comp の頻度が高く，これらの分野では特に道具を用いた動作が多いことがわかる．また，*Hobbies and Crafts* では I-part-of の頻度が他 3 分野と比較して高いが，これはこの分野において，材料の一部分に対する動作が多いことを示唆している.

4.4　アノテーション一致率

アノテーションの質を評価するため，別のアノテータに依頼し，分野ごとに 3 記事を再アノテーションした後，それらの一致率を F 値で計算した．表 7 にその結果を示す．ノードのアノテーションに関しては，89.68%という非常に高い一致率が得られた．これは手順に関わる表現とタグの検出は，比較的容易であることを表している．また，エッジのアノテーションに関しては，68.79%という一致率が得られた．これは，ノードアノテーション時のミスが影響することを考慮すれば，十分高い一致率であるといえる.

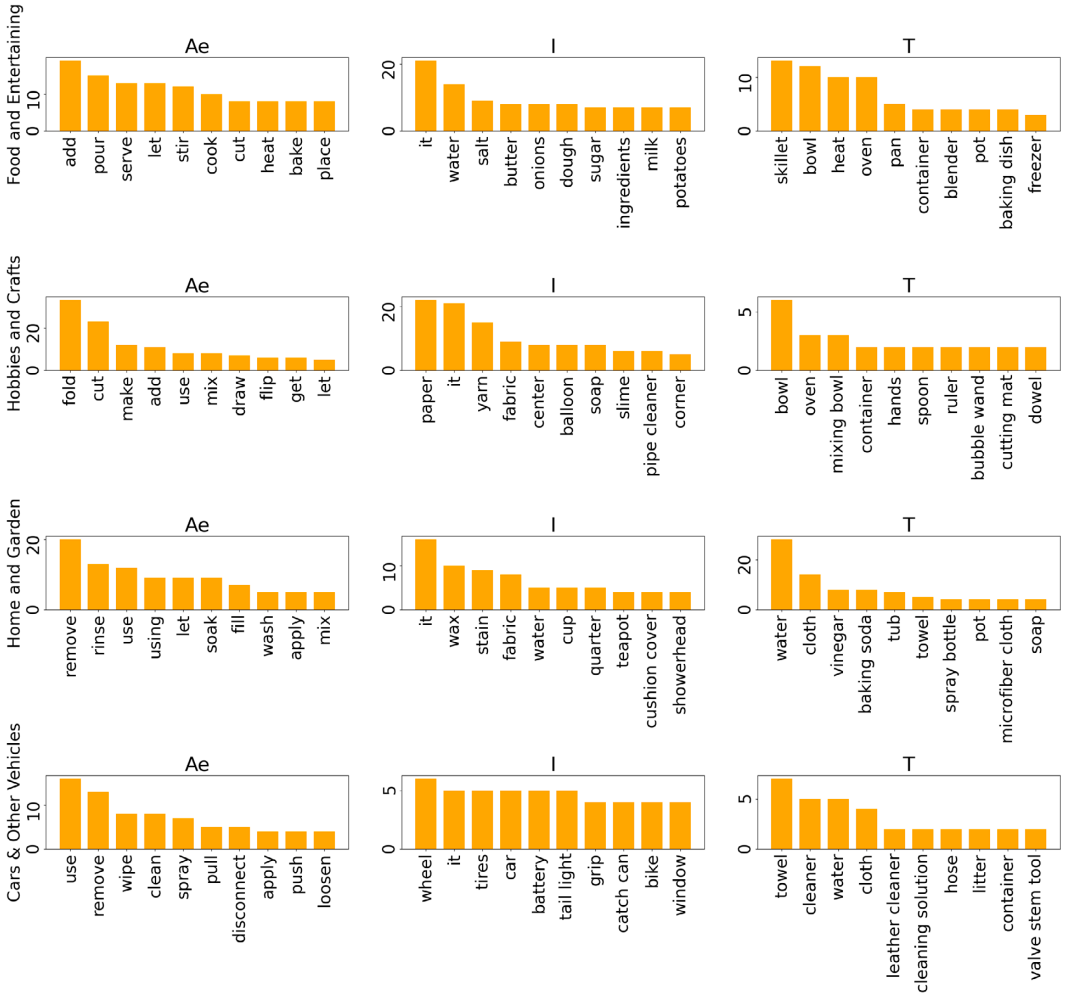

図 3　Ae，I，T における上位 10 件の表現とその頻度.

5　ノード予測

フローグラフ予測においては，手順に関する表現をタグ付きで検出するノード予測を最初に行う．ここでは，先行研究と同様に固有表現認識器をノード予測モデルとして用いる．以降では，実験設定について説明した後，実験結果について述べる．最後に，タグレベルでの予測性能について議論する.

5.1　実験設定

モデル.　固有表現認識器として BiLSTM-CRF を採用した (Lample et al. 2016)．エンコーダ

ラベル	*Food and Entertaining*	*Hobbies and Crafts*	*Home and Garden*	*Cars & Other Vehicles*
Agent	46	49	25	20
Targ	396	341	301	231
Dest	151	145	100	47
T-comp	25	19	64	42
I-comp	7	4	3	3
I-eq	146	149	110	68
I-part-of	29	124	63	76
Set	8	7	0	2
T-eq	21	8	18	10
T-part-of	5	3	16	6
A-eq	4	6	2	6
V-tm	32	9	2	2
other-mod	254	212	182	109
合計	1,124	1,076	886	622

表 6　依存関係ラベルごとのアノテーション数.

アノテーションの種類	一致率
ノード	89.68%
エッジ	68.79%

表 7　アノテータ間のアノテーション一致率.

としては，BiLSTM ではなく事前学習済みの DeBERTa (He et al. 2021) を用いた[5]．このモデルのパラメータ数は合計 1.40 億個であった.

学習.　分野適応を行うモデルについては，まず English r-FG コーパスで事前学習を行い，その後 w-FG コーパスの対象分野のデータでファインチューニングを行った．分野適応を行わないモデルについては，English r-FG コーパスあるいは w-FG コーパスの一方のみを用いて学習を行った．学習時には事前学習済みのエンコーダを含む全てのパラメータを調節した.

パラメータの最適化には AdamW (Loshchilov and Hutter 2019) を用いた．初期学習率は 5.0×10^{-5}，weight decay は 1.0×10^{-5} に設定した．学習率の調節には warmup は S_{w} ステップ，cosine-annealing (Loshchilov and Hutter 2019) を S_{d} ステップ行った．ミニバッチは B 記事から構築した．ハイパーパラメータの選択に関して，ミニバッチサイズ B は {3,5} から，warmup と cosine-annealing のステップ数の組み合わせ $(S_{\mathrm{w}}, S_{\mathrm{d}})$ は {(100, 900), (500, 4500), (1000, 9000)} から，開発データを用いて最適なパラメータを選択した．その結果，English r-FG コーパス上での

[5] 予備実験では，BERT ではなく DeBERTa を用いることで，English r-FG コーパスにおいて 0.47%精度が向上することを確認した.

学習では $(B, S_{\mathrm{w}}, S_{\mathrm{d}}) = (5, 500, 4500)$，w-FG コーパス上での学習では $(B, S_{\mathrm{w}}, S_{\mathrm{d}}) = (3, 100, 900)$ とした．データ拡張は分野適応モデルのファインチューニング時にのみ行い，手順の入れ替えに関しては 1 記事あたり 5 記事を，単語の置換に関しては 1 記事あたり 10 記事を，$p = 0.5$ として行った．

評価. English r-FG コーパスは全体で 300 記事あり，そのうちの 80%を学習データとして，10%を開発データとして，残りの 10%をテストデータとして用いた．ここで，テストデータは付録 A で述べる追加実験の評価時に用いた．w-FG コーパスは分野ごとに 30 記事あり，それらを 6 分割したときの 1 分割を学習データとして，別の 1 分割を開発データとして，残りの 4 分割をテストデータとして用いた．より信頼性の高い結果を得るために，本論文ではテストデータの分割を変えながら 6 分割交差検証を行った．先行研究 (Maeta et al. 2015; Yamakata et al. 2020) に従い，評価指標には精度，再現率，F 値を用いた．

モデル設定. 分野適応を行うモデルを **domain-adaptation**，English r-FG コーパスのみで学習したモデルを **cooking-only**，w-FG コーパスのみで学習したモデルを **target-only** とそれぞれ参照する．

5.2 実験結果

表 8 に実験結果を示す．**target-only** モデルのスコアから，少量の学習データを用いて 66.9% 以上の F 値でノード予測を行えることがわかる．また，**cooking-only** モデルのスコアから，調理分野のデータを用いた場合にも **target-only** と競合するスコアを実現可能であることがわかる．*Food and Entertaining* では，**cooking-only** モデルが **target-only** を F 値で 10.3%上回っているが，これはこの分野の手順書が調理レシピを主に扱っているためである．次に，**domain-adaptation** モデルのスコアから，全ての分野において分野適応を行った場合に，最高のスコアを実現することがわかる．特に *Food and Entertaining* 以外の 3 分野においては，**target-only** から **domain-adaptation** で F 値において 9.6%以上の向上を実現しており，English r-FG コーパスでの事前学習が効果的であることを示唆している．

データ拡張を行った場合の結果に関しては，手順の入れ替えを行う場合には *Food and Entertaining* と *Home and Garden* の 2 分野においてそれぞれ 0.4%, 0.3%の微量の改善が確認出来たのみであった．また，単語の置換を行った場合には，全ての分野で性能向上が得られなかった．これらの結果は，ノード予測において，これらのデータ拡張の効果が薄いことを表している．

5.3 タグごとの予測性能

各固有表現タグに対応する表現は分野間で大きく異なる．このため，**domain-adaptation** モデルを考えたとき，English r-FG コーパスと w-FG コーパスの対象分野のデータとの表現の重複率が低いほど，ファインチューニング時に分野依存の表現を多く学習するため，**cooking-only**

モデルからの性能改善の幅が大きくなると予想される．これを調査するため，コーパス間におけるタグごとに表現の重複率を計算し，**cooking-only** モデルと **domain-adaptation** モデルのタグごとの予測性能を F 値で評価した．ここでは，タグの中でも特に出現頻度の高い Ae, C, T を対象とした．

表 9 に結果を示す．Ae では予想とは異なり，コーパス間の重複率に関わらず **cooking-only** から **domain-adaptation** への性能改善の度合いは小さいことが確認出来る．これは，学習デー

対象分野	モデル	データ拡張		精度	再現率	F 値
		手順の入れ替え	単語置換			
Food and Entertaining	target-only			0.770	0.784	0.777
	cooking-only			0.884	0.877	0.880
	domain-adaptation			0.890	0.892	0.891
	domain-adaptation	✓		**0.894**	**0.895**	**0.895**
	domain-adaptation		✓	0.885	0.891	0.888
Hobbies and Crafts	target-only			0.698	0.707	0.702
	cooking-only			0.703	0.684	0.693
	domain-adaptation			**0.794**	**0.805**	**0.799**
	domain-adaptation	✓		0.784	0.795	0.789
	domain-adaptation		✓	0.781	0.790	0.785
Home and Garden	target-only			0.663	0.676	0.669
	cooking-only			0.734	0.742	0.738
	domain-adaptation			0.780	0.786	0.783
	domain-adaptation	✓		**0.787**	**0.791**	**0.786**
	domain-adaptation		✓	0.765	0.773	0.769
Cars & Other Vehicles	target-only			0.650	0.690	0.669
	cooking-only			0.646	0.695	0.670
	domain-adaptation			**0.748**	**0.784**	**0.765**
	domain-adaptation	✓		0.734	**0.784**	0.761
	domain-adaptation		✓	0.729	0.772	0.750

表 8　ノード予測の実験結果．表中のチェックマーク (✓) は用いたデータ拡張手法を指す．

分野	Ae			I			T		
	重複率	F 値		重複率	F 値		重複率	F 値	
		Cook.	Adapt.		Cook.	Adapt.		Cook.	Adapt.
Food and Entertaining	92.06%	0.941	0.952	72.11%	0.932	0.933	77.94%	0.896	0.882
Hobbies and Crafts	69.03%	0.943	0.951	10.33%	0.717	0.833	51.79%	0.398	0.588
Home and Garden	65.19%	0.954	0.961	18.40%	0.716	0.795	43.55%	0.567	0.678
Cars & Other Vehicles	46.04%	0.905	0.919	6.88%	0.666	0.805	27.47%	0.459	0.557

表 9　Ae，I，T のタグごとの予測性能と English r-FG コーパス中の表現との重複率．Cook.，Adapt. はそれぞれ **cooking-only**，**domain-adaptation** を指す．

タの分野に依らず，人の動作表現は高い精度で検出可能であることを意味している．C と T においては，*Food and Entertaining* を除く全ての分野で，分野適応による大幅の性能改善が確認できる．これらの分野における重複率は *Food and Entertaining* よりも低く，これは C と T に対応するノードの予測性能に関しては，当初の予想通りファインチューニングによる影響が大きいことを示している．

6 エッジ予測

ノード予測後は，ノード間の依存関係をラベル付きで予測するエッジ予測を行う．ここでは，先行研究と同様に依存構造解析器をエッジ予測モデルとして用いる．先行研究 (Yamakata et al. 2020) に従い，手順書の最後に登場した動作表現 (Ae) をルートノードとする．以降では，実験設定，ノードが既知である場合の実験結果，5.2 節で予測したノードを与えた場合の実験結果について順に述べる．

6.1 実験設定

モデル. 依存構造解析器として，Biaffine dependency parser (Dozat and Manning 2018) を採用した[6]．このモデルはエッジとラベルの予測に異なるモジュールを用いており，最終的な誤差はそれぞれのモジュールの誤差の重み付け和として以下のように与えられる．

$$l = \lambda l^{\mathrm{Edge}} + (1-\lambda) l^{\mathrm{Label}}, \tag{4}$$

ここで，λ は各誤差の強さを制御し，ここでは 0.5 に設定した．また，言語エンコーダとしては事前学習済みの DeBERTa (He et al. 2021) を用いた[7]．このモデルのパラメータ数は合計 1.49 億個であった．

学習. 5.1 節と同様に，English r-FG コーパスと w-FG コーパスを用いて分野適応を行うモデル，いずれかのコーパスのみを用いて学習するモデルをそれぞれ学習した．パラメータの最適化にも同様に AdamW (Loshchilov and Hutter 2019) を用い，学習率の調節は warmup と cosine-annealing を用いて行った．これらのハイパーパラメータについては，5.1 節と同様に開発データ上で調節し，English r-FG コーパス上での学習では $(B, S_{\mathrm{w}}, S_{\mathrm{d}}) = (5, 500, 4500)$，w-FG コーパス上での学習では $(B, S_{\mathrm{w}}, S_{\mathrm{d}}) = (3, 100, 900)$ とした．

評価. Englsih-FG コーパス，w-FG コーパス共に 5.1 節と同様の分割を用い，6 分割の交差検証を行った．評価指標には，ラベル付きエッジ (u, v, l) を基に，精度，再現率，F 値を計算した．

[6] 先行研究 (Yamakata et al. 2020) では線形モデルを用いて実装されているが，biaffine dependency parser を用いることで English r-FG コーパスにおいてより高い性能を実現出来ることを確認した．この予備実験の結果については付録 A で示す．

[7] 言語エンコーダとしてその他の言語モデルを用いた場合の比較については，付録 A で説明する．

モデル設定.　5 節と同様に，**cooking-only**，**target-only**，**domain-adaptation** で各モデルを参照する．

6.2　実験結果

正解のノードが既知であり，エッジのみを予測した結果を表 10 に示す．ノード予測の実験とは異なり，**target-only** モデルのスコアは全分野において 33.8%以下と低いことが確認できる．一方で，**cooking-only** モデルのスコアは全分野で 58.7%以上であり，各分野の **cooking-only** モデルのスコアを倍以上上回っている．これらの結果はエッジ予測モデルの学習にはノード予測モデル以上に多くのデータが必要であり，結果として学習サンプル数の多い **cooking-only** モデルの方が高い性能を示したと考えられる．次に，**domain-adaptation** モデルは全ての分野で **target-only** と **cooking-only** を上回り，最高のスコアを実現している．これは，English r-FG コーパスからの分野適応がノード予測と同様に効果的であることを示している．また，データ拡張を行った際には，手順の入れ替えは効果が無い一方で，単語の置換は *Food and Entertaining* 以外の 3 分野で最大 0.15%の改善が得られている．しかし，Bootstrap resampling (Koehn 2004)

分野	モデル	データ拡張		精度	再現率	F 値
		手順の入れ替え	単語置換			
Food and Entertaining	target-only			0.335	0.338	0.337
	cooking-only			0.725	0.731	0.728
	domain-adaptation			0.750	**0.756**	**0.753**
	domain-adaptation	✓		0.747	0.752	0.750
	domain-adaptation		✓	**0.761**	0.752	0.749
Hobbies and Crafts	target-only			0.285	0.281	0.283
	cooking-only			0.613	0.605	0.609
	domain-adaptation			0.649	0.640	0.644
	domain-adaptation	✓		0.646	0.638	0.642
	domain-adaptation		✓	**0.653**	**0.644**	**0.648**
Home and Garden	target-only			0.229	0.232	0.231
	cooking-only			0.644	0.649	0.646
	domain-adaptation			0.659	0.665	0.662
	domain-adaptation	✓		0.656	0.662	0.659
	domain-adaptation		✓	**0.674**	**0.680**	**0.677**
Cars & Other Vehicles	target-only			0.154	0.155	0.154
	cooking-only			0.587	0.590	0.587
	domain-adaptation			0.607	0.610	0.609
	domain-adaptation	✓		0.607	0.610	0.608
	domain-adaptation		✓	**0.617**	**0.620**	**0.618**

表 10　エッジ予測の実験結果．表中のチェックマーク (✓) は用いたデータ拡張手法を表す．

分野	F 値
Food and Entertaining	0.679　(−9.8%)
Hobbies and Crafts	0.501 (−22.2%)
Home and Garden	0.494 (−25.4%)
Cars & Other Vehicles	0.449 (−26.3%)

表 11　パイプラインモデルの実験結果．括弧内の数値は表 10 の数値との差分を表す．

に基づく有意差検定を行ったところ，有意差は認められなかったため，このデータ拡張手法の有効性に関しては引き続き調査する必要がある．

6.3　パイプラインモデルの予測性能

ここまではノードが既知である設定で，エッジ予測のみを行った．しかし，実際にフローグラフを予測する際には，予測したノードを基にエッジを予測する必要がある．この場合，ノード予測時の誤りが影響を与えるため，エッジの予測性能は 6.2 節の時より低下すると考えられる．このときの予測性能を調べるために，ここでは表 8 の予測結果を基にエッジ予測を行う．評価は，(u, v, l) にノード u，v の固有表現タグ n_u，n_v を加えた (u, v, l, n_u, n_v) の一致率を F 値で計算することで行った．

表 11 に結果を示す．これらのスコアは，この設定下において，ラベル付きエッジを 44.9%から 67.9%の F 値で予測可能であることを示している．また，表 10 からの性能低下については，*Food and Entertaining* では 9.8%であるのに対し，他 3 分野では平均 24.6%の大きな低下が見られた．これには，ノード予測の性能差が影響している可能性が高い．表 8 を見ると，**domain-adaptation** モデルの *Food and Entertaining* での F 値は 89.1%であるが，その他 3 分野では 76.5%から 79.9%と，9.2%以上の開きがあり，この性能差がこれら 3 分野における性能低下に繋がっていると考えられる．

7　w-FG の限界

w-FG は調理分野以外の手順書をフローグラフに変換するために，r-FG を拡張して得られた表現である．しかし，w-FG のアノテーションで対応可能な手順書には，以下の 3 つの限界が存在する．

まず，w-FG は English r-FG (Yamakata et al. 2020) を基にした表現であるため，対象の手順書は英語で記述されている必要がある．2.1 節で述べたとおり，日本語の r-FG (Mori et al. 2014) を英語に拡張する際には，英語特有の表現を扱うために，一部の固有表現タグの追加が行われ

ている．従って，w-FG を英語以外の言語に適用する際には，同様に対象の言語特有の表現を扱えるように，タグやラベルの追加が必要となる可能性がある．

また，w-FG のアノテーションは，手順が物理的な物体を対象とする手順書に限られている．これは，w-FG が r-FG を基に設計されており，r-FG では物理的な材料を扱う手順を対象としているためである．例えば，wikiHow においては，“how to be popular” や “how to not get a nervous” といった抽象的な目標や手順を含む手順書が存在するが，これらに対する w-FG のアノテーションは想定していない．

最後に，w-FG のアノテーションでは，対象とする手順書が一つの最終成果物を持つと仮定している．これは，r-FG (Mori et al. 2014; Yamakata et al. 2020) の定義に従い，フローグラフを根付き有向非巡回グラフとして表現するためである．手順書によっては，複数の最終成果物が存在しうるが，w-FG のアノテーションではこのような手順書は想定していない．

8　関連研究

r-FG コーパスにはテキストのみのアノテーション (Mori et al. 2014; Yamakata et al. 2020) に加え，視覚的なアノテーションを施したものが提案されている (Nishimura et al. 2020; Shirai et al. 2022)．Nishimura et al. (2020) は手順ごとに 1 枚の画像を付与し，それが動作中か動作後かという情報に加え，画像中の食材と道具のバウンディングボックスをアノテーションした．Shirai et al. (2022) はレシピに付随する調理動画を用いて，各調理動作の前後に対応するフレームのアノテーションを行った．本論文で提案した w-FG コーパスではテキストのみのアノテーションであるが，各手順を視覚的に説明する画像が紐づいており，これを用いてクロスモーダルなアノテーションを行うことが可能である．

手順書を手順の依存関係を表すグラフ構造として表現する研究は，r-FG や w-FG の他にも存在する．調理分野という枠組みでは，手順書のグラフ表現を教師なし学習によって獲得するアプローチ (Kiddon et al. 2015) や，r-FG と同様に人手アノテーションで構築したコーパスを用いてグラフ予測のためのモデルを学習する研究 (Pan et al. 2020; Papadopoulos et al. 2022) がある．生化学分野においては，実験の自動化のためにプロトコルをグラフ表現に変換するアプローチが提案されている (Kulkarni et al. 2018; Tamari et al. 2021)．材料科学の分野では，科学文献に含まれる合成プロセスの解析のため，合成手順を有向非巡回グラフとして表現したコーパスが提案されている (Mysore et al. 2019; Kuniyoshi et al. 2020)．生化学と材料科学分野におけるこれらの手順書は，手順に関わる表現をノード，ノード間の依存関係をエッジとして表現する点において r-FG や w-FG と共通している．

wikiHow は多様な手順の知識を含む言語資源として，先行研究で広く用いられている．Zhou et al. (2019) や Zhou et al. (2022) では手順の知識ベースとして，Zellers et al. (2019) では常識

推論のためのデータセットとして利用している．Zhang et al. (2020b) や Zhang et al. (2020a) では，wikiHow 記事中の見出しを手順として，記事タイトルを目標として捉えることで，手順から目標を推論するタスクを提案している．また，Lin et al. (2022) や Zhou et al. (2023) は，作業動画中の手順の理解のために wikiHow の手順知識を活用している．本研究で提案した w-FG は，手順全体の流れをグラフ構造として表現するため，手順単体に加え，それらの依存関係も含めた知識として利用可能である．

9 おわりに

本論文では，調理分野に留まらない一般的な手順書の理解の表現として，wikiHow フローグラフを提案した．これを基に，wikiHow の記事をフローグラフアノテーションを施した w-FG コーパスを新たに構築し，フローグラフ予測の実験を行った．実験では，既存の調理分野のデータを用いて事前学習し，対象分野のデータでファインチューニングを行うことで，片方のデータのみを学習に用いた場合に比べて大幅な性能改善が見込めることを確認した．今後の方向性としては，手順書フローグラフを材料科学 (Kuniyoshi et al. 2020) や生化学 (Kulkarni et al. 2018) の分野の手順書に適用することや，大規模言語モデルを用いて学習なしでフローグラフ表現を予測すること等が挙げられる．

謝 辞

本研究は JSPS 科研費 JP20H04210, JP21H04910 の助成を受けたものです．

本論文の一部は言語処理学会第 29 回年次大会 (白井 他 2023) および The 8th Workshop on Representation Learning for NLP (RepL4NLP 2023) (Shirai et al. 2023) で発表したものです．

参考文献

Beltagy, I., Peters, M. E., and Cohan, A. (2020). “Longformer: The Long-document Transformer.” *arXiv preprint arXiv:2004.05150*.

Bollini, M., Tellex, S., Thompson, T., Roy, N., and Rus, D. (2013). “Interpreting and Executing Recipes with a Cooking Robot.” In *Experimental Robotics*, pp. 481–495. Springer.

Bosselut, A., Levy, O., Holtzman, A., Ennis, C., Fox, D., and Choi, Y. (2018). “Simulating Action Dynamics with Neural Process Networks.” In *Proceedings of the 6th International Conference on Learning Representations*.

Chu, Y.-J. and Liu, T.-H. (1965). “On the Shortest Arborescence of a Directed Graph.” *Science Sinica*, **14**, pp. 1396–1400.

Dai, X. and Adel, H. (2020). “An Analysis of Simple Data Augmentation for Named Entity Recognition.” In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3861–3867. International Committee on Computational Linguistics.

Dalvi, B., Huang, L., Tandon, N., Yih, W.-t., and Clark, P. (2018). “Tracking State Changes in Procedural Text: a Challenge Dataset and Models for Process Paragraph Comprehension.” In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1595–1604.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics.

Ding, B., Liu, L., Bing, L., Kruengkrai, C., Nguyen, T. H., Joty, S., Si, L., and Miao, C. (2020). “DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6045–6057. Association for Computational Linguistics.

Dozat, T. and Manning, C. D. (2018). “Simpler but More Accurate Semantic Dependency Parsing.” In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 484–490. Association for Computational Linguistics.

Edmonds, J. (1967). “Optimum Branchings.” *Journal of Research of the National Bureau of Standards: Mathematics and mathematical physics. B*, **71**, pp. 233–240.

Fadaee, M., Bisazza, A., and Monz, C. (2017). “Data Augmentation for Low-Resource Neural Machine Translation.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 567–573. Association for Computational Linguistics.

He, P., Liu, X., Gao, J., and Chen, W. (2021). “Deberta: Decoding-Enhanced Bert with Disentangled Attention.” In *9th International Conference on Learning Representations*.

Kiddon, C., Ponnuraj, G. T., Zettlemoyer, L., and Choi, Y. (2015). “Mise en Place: Unsupervised Interpretation of Instructional Recipes.” In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 982–992. Association for Computational

Linguistics.

Koehn, P. (2004). "Statistical Significance Tests for Machine Translation Evaluation." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395. Association for Computational Linguistics.

Kulkarni, C., Xu, W., Ritter, A., and Machiraju, R. (2018). "An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 97–106. Association for Computational Linguistics.

Kuniyoshi, F., Makino, K., Ozawa, J., and Miwa, M. (2020). "Annotating and Extracting Synthesis Process of All-Solid-State Batteries from Scientific Literature." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1941–1950. European Language Resources Association.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). "Neural Architectures for Named Entity Recognition." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270. Association for Computational Linguistics.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." In *Proceedings of the 8th International Conference on Learning Representations*.

Lin, X., Petroni, F., Bertasius, G., Rohrbach, M., Chang, S.-F., and Torresani, L. (2022). "Learning To Recognize Procedural Activities With Distant Supervision." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13853–13863.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). "Roberta: A Robustly Optimized BERT Pretraining Approach." *arXiv preprint arXiv:1907.11692*.

Loshchilov, I. and Hutter, F. (2019). "Decoupled Weight Decay Regularization." In *Proceedings of the 7th International Conference on Learning Representations*.

Maeta, H., Sasada, T., and Mori, S. (2015). "A Framework for Procedural Text Understanding." In *Proceedings of the 14th International Conference on Parsing Technologies*, pp. 50–60. Association for Computational Linguistics.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). "Non-Projective Dependency Parsing using Spanning Tree Algorithms." In *Proceedings of Human Language Technology Con-*

ference and Conference on Empirical Methods in Natural Language Processing, pp. 523–530. Association for Computational Linguistics.

Mori, S., Maeta, H., Yamakata, Y., and Sasada, T. (2014). "Flow Graph Corpus from Recipe Texts." In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pp. 2370–2377.

Mysore, S., Jensen, Z., Kim, E., Huang, K., Chang, H.-S., Strubell, E., Flanigan, J., McCallum, A., and Olivetti, E. (2019). "The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures." In *Proceedings of the 13th Linguistic Annotation Workshop*, pp. 56–64. Association for Computational Linguistics.

Nishimura, T., Tomori, S., Hashimoto, H., Hashimoto, A., Yamakata, Y., Harashima, J., Ushiku, Y., and Mori, S. (2020). "Visual Grounding Annotation of Recipe Flow Graph." In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4275–4284. European Language Resources Association.

Pan, L.-M., Chen, J., Wu, J., Liu, S., Ngo, C.-W., Kan, M.-Y., Jiang, Y., and Chua, T.-S. (2020). "Multi-Modal Cooking Workflow Construction for Food Recipes." In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1132–1141. Association for Computing Machinery.

Papadopoulos, D. P., Mora, E., Chepurko, N., Huang, K. W., Ofli, F., and Torralba, A. (2022). "Learning Program Representations for Food Images and Cooking Recipes." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16559–16569.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 101–108. Association for Computational Linguistics.

Shirai, K., Hashimoto, A., Nishimura, T., Kameko, H., Kurita, S., Ushiku, Y., and Mori, S. (2022). "Visual Recipe Flow: A Dataset for Learning Visual State Changes of Objects with Recipe Flows." In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3570–3577. International Committee on Computational Linguistics.

Shirai, K., Kameko, H., and Mori, S. (2023). "Towards Flow Graph Prediction of Open-Domain Procedural Texts." In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pp. 87–96. Association for Computational Linguistics.

白井圭佑，亀甲博貴，森信介 (2023)．オープンドメインの手順書のフローグラフ予測とデータセットの構築．言語処理学会第 29 回年次大会発表論文集, pp. 2869–2873．[K. Shirai et

al. (2023). Flow Graph Prediction of Open-Domain Procedural Texts and Dataset Creation. Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing, pp. 2869–2873.].

Tamari, R., Bai, F., Ritter, A., and Stanovsky, G. (2021). "Process-Level Representation of Scientific Protocols with Interactive Annotation." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2190–2202. Association for Computational Linguistics.

Tandon, N., Sakaguchi, K., Dalvi, B., Rajagopal, D., Clark, P., Guerquin, M., Richardson, K., and Hovy, E. (2020). "A Dataset for Tracking Entities in Open Domain Procedural Text." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6408–6417.

Xu, H., Ebner, S., Yarmohammadi, M., White, A. S., Van Durme, B., and Murray, K. (2021). "Gradual Fine-Tuning for Low-Resource Domain Adaptation." In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pp. 214–221. Association for Computational Linguistics.

Yamakata, Y., Mori, S., and Carroll, J. A. (2020). "English Recipe Flow Graph Corpus." In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5187–5194. European Language Resources Association.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). "HellaSwag: Can a Machine Really Finish Your Sentence?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics.

Zhang, L., Lyu, Q., and Callison-Burch, C. (2020a). "Intent Detection with WikiHow." In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 328–333. Association for Computational Linguistics.

Zhang, L., Lyu, Q., and Callison-Burch, C. (2020b). "Reasoning about Goals, Steps, and Temporal Ordering with WikiHow." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4630–4639. Association for Computational Linguistics.

Zhou, H., Martín-Martín, R., Kapadia, M., Savarese, S., and Niebles, J. C. (2023). "Procedure-Aware Pretraining for Instructional Video Understanding." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10727–10738.

Zhou, S., Zhang, L., Yang, Y., Lyu, Q., Yin, P., Callison-Burch, C., and Neubig, G. (2022). "Show Me More Details: Discovering Hierarchies of Procedures from Semi-structured Web Data." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pp. 2998–3012. Association for Computational Linguistics.

Zhou, Y., Shah, J., and Schockaert, S. (2019). "Learning Household Task Knowledge from WikiHow Descriptions." In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pp. 50–56. Association for Computational Linguistics.

付録

A　English r-FG コーパス上でのエッジ予測

ここでは，English r-FG コーパス上でのエッジ予測の性能について，先行研究 (Yamakata et al. 2020) との比較を行う．先行研究では Maeta et al. (2015) と同様の線形モデルが用いられている．先行研究と実験設定を合わせるため，コーパス全体の 80%を学習データとして，10%を開発データとして，残りの 10%をテストデータとして，10 分割交差検証を行った．また，エンコーダとして事前学習済みの BERT (Devlin et al. 2019)，RoBERTa (Liu et al. 2019)，DeBERTa (He et al. 2021)，Longformer (Beltagy et al. 2020)，ALBERT (Lan et al. 2020) をそれぞれ用い，実験を行った．

表 12 に結果を示す．これより，事前学習済みの言語モデルを用いることで，先行研究を上回る性能を実現出来ることがわかる．また，言語モデルによっても性能差があり，今回の実験では DeBERTa を用いることで，全指標において最もよいスコアが得られることがわかる．

B　w-FG コーパス上でのエッジ予測と各言語モデルの比較

w-FG コーパスにおけるエッジ予測について，異なる言語モデルを用いた場合の性能を比較する．ここでは，BERT，RoBERTa，Longformer，ALBERT，DeBERTa を用いる．ここでは **domain-adaptation** モデルの結果を報告する．

言語モデル	精度	再現率	F 値
Yamakata et al. (2020)	0.737	0.686	0.711
BERT	0.737	0.703	0.720
RoBERTa	0.754	0.719	0.736
Longformer	0.751	0.716	0.733
ALBERT	0.744	0.710	0.727
DeBERTa	**0.756**	**0.721**	**0.738**

表 12　English r-FG コーパス上におけるエッジ予測について，言語モデルを変えた場合の実験結果．

分野	言語モデル	F 値
Food and Entertaining	BERT	0.729
	RoBERTa	0.736
	Longformer	0.735
	ALBERT	0.725
	DeBERTa	**0.753**
Hobbies and Crafts	BERT	0.637
	RoBERTa	**0.644**
	Longformer	0.631
	ALBERT	0.603
	DeBERTa	**0.644**
Home and Garden	BERT	**0.672**
	RoBERTa	0.665
	Longformer	0.662
	ALBERT	0.627
	DeBERTa	0.662
Cars & Other Vehicles	BERT	0.611
	RoBERTa	0.610
	Longformer	0.600
	ALBERT	**0.623**
	DeBERTa	0.609

表 13　エッジ予測において，言語モデルを変えた場合の実験結果.

表 13 に結果を示す．これより，BERT や ALBERT を用いることで *Home and Garden* と *Cars & Other Vehicles* において最もよい F 値が実現出来ることがわかる．また，その他の分野においては，DeBERTa が最高のスコアを実現することが確認できる．

略歴

白井　圭佑：2017 年愛媛大学工学部卒業．2020 年京都大学大学院情報学研究科修士課程修了．2020 年より京都大学大学院情報学研究科博士後期課程在籍中．2024 年同大学院博士取得見込み．自然言語処理，マルチメディアに関する研究に従事．言語処理学会会員.

亀甲　博貴：2018 年東京大学大学院工学系研究科博士課程修了．博士（工学）．同年より京都大学学術情報メディアセンター助教．自然言語処理，ゲーム AI 等に関する研究に従事．言語処理学会，情報処理学会各会員.

森　　信介：1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了．同年日本アイ・ビー・エム株式会社入社．2007 年より京都大学学術情報メディアセンター准教授．2016 年同教授．現在に至る．計算言語学ならび

に自然言語処理の研究に従事．博士（工学）．1997 年情報処理学会山下記念研究賞受賞．2010 年，2013 年情報処理学会論文賞受賞．2010 年第 58 回電気科学技術奨励賞．2023 年言語処理学会論文賞受賞．言語処理学会，情報処理学会，日本データベース学会各会員．

（2023 年 9 月 30 日　受付）
（2024 年 1 月 11 日　再受付）
（2024 年 2 月 20 日　採録）