日本語 CommonGen における入力キーワード群のハブ単語の自動追加による生成文の改善

鈴木雅人 [1]， 新納浩幸 [2]
[1] 茨城大学工学部情報工学部, [2] 茨城大学大学院理工学研究科情報科学領域
{19T4042Y, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

概要

常識推論能力を試すタスクとして CommonGen がある．CommonGen は数個のキーワードを入力として，それらキーワードを含む妥当な文を生成するタスクである．論文 [1] では日本語 CommonGen のデータセットとベースラインシステムを作成している．またそこでは入力キーワード群のハブ単語を追加することで生成文が改善される可能性を論じている．ただしハブ単語の選出は人手で行っており，実装には至っていない．本論文ではハブ単語の選出法を提案，実装しその効果を検証する．

1 はじめに

常識推論は人工知能の重要な問題であり，自然言語処理の分野でも，常識推論を扱うタスクがいくつか考案されている [2][3][4][5][6][7]．そのなかで Lin らは CommonGen という新しい常識推論のタスクを提案した [8]．CommonGen は，概略，数個のキーワードを入力し，それらキーワードを用いて妥当な文を生成するという制約付き文生成のタスクである．例えば「彼，犬，餌」なら「彼が犬に餌をあげる」などといった文を生成するのが目標である．「犬が彼に餌をあげる」は文法的には正しくても常識的にはおかしな文であり，そのような文を生成しないために常識推論が必要とされる．

日本語 CommonGen の試作をおこなった研究 [1] では，データセットの作成や T5 を用いた CommonGen のモデルを作成し，このタスクが日本語においても常識推論能力が必要とされることを示している．また生成文を改善する手法として入力キーワード群のハブ単語を追加する手法が示された．しかしそこではハブ単語は人手により与え，ハブ単語の効果を論じているだけであり，ハブ単語をどのように選出し，どの程度の効果があるかは示されていない．ここではハブ単語の選出方法を提案し，生成文がどの程度妥当になるかを調査した．

2 関連研究

自然言語処理の分野では QA や対話などで，深い意味理解が必要となる場面で常識推論が使われ，様々なタスクが考案されている．CommonsenseQA [2] は元となる言葉に関係する 3 つの言葉を用意し，関係する元の言葉を含みながら，3 つの言葉の内それぞれひとつずつのみに当てはまる質問を用意する．またこの際に，追加で元となる言葉に関係する 2 つの誤答用の言葉が用意され，3 つの質問は 2 つの誤答用の言葉には当てはまらないようにする．そうして用意された 3 つの質問に対して 5 つの選択肢となる言葉からそれぞれ，どの言葉が当てはまるかを選択するタスクである．SocialIQA [3] は社会的に一般的な状況を提示し，その状況下で取る行動やその状況下での心情を尋ねる質問と 3 つの選択肢が与えられ，適切な回答を選ぶタスクである．WinoGrande [4] は 2 つの文章とその文章中にある代名詞に対して 2 つの単語の選択肢が与えられ，各文章に代名詞の指す正しい単語を選択するタスクであり，これは Winograd Schema Challenge からバイアスを除去し，クラウドソージングの手続きを改善したものである．KUCI [5] は日本語において中断されている文章とそれに続く蓋然的な関係を持つ文章を 4 つの選択肢から 1 つ選択するタスクである．SWAG [6] は「ある場面でのビデオキャプション」と 4 つの「次の場面でのキャプション」の選択肢となる文章を提示し，選択肢から本物の次の場面でのキャプションを選択するタスクである．HellaSwag [7] は SWAG を元としてさらに難しい不正解の選択肢を導入する Adversarial Filtering や元となるビデオキャプションの厳選などを行い，SWAG を改善したものである．

これらタスクは基本的には選択式の問題である．

CommonGen は制約付き文生成のタスクであり，選択式の問題では扱えない問題を扱っていると考えられる．

CommonGen と類似のシステムとしては株式会社 ELYZA がデモ版を公開している ELYZA Pencil[1] がある．ELYZA Pencil は数個のキーワードからそのキーワードを使ったニュース記事（タイトルとその本文）を生成する．生成される記事はかなり高品質であるが，商用システムであるため使われている技術の詳細は明らかにされていない．

3 データセットの構築

本論文の実験は基本的に論文 [1] で作成されたデータセットを用いる．ここではそのデータセットがどのように構築されたかを記す．

データセットは Web 上で公開されている STAIR Captions[2] の画像キャプションのデータセットを元とした．キャプション 12,000 文から各キャプションを MeCab をもちいて形態素解析した結果を取得し，キーワード候補を品詞を用いて抽出を行った[3]．そこからランダムに 3 つの単語を選出し，選出した 3 単語を入力，対応するキャプションを出力とするデータセットを構築した（図 1 参照）．

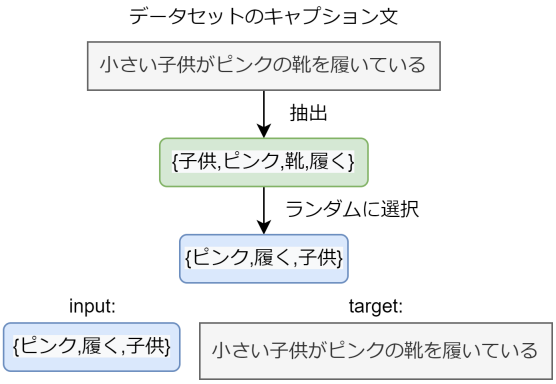

図 1 CommonGen 用データセット作成例

12,000 組のデータの内 9 割（10,800 組）が訓練データ，残りデータの半数（600 組）が検証データ，最後に残ったデータ（600 組）の中の 200 組をテストデータ（このテストデータを「自動作成したテストデータ」と呼ぶ）とした．作成したデータの例を表 1 に示す．

またキーワードによる生成の容易さを調べるために考案した 3 タイプ 60 組のテストデータを追加した．このテストデータを「独自に作成したテストデータ」と呼ぶ．独自に作成したテストデータには正解となる文は作成していないことを注記しておく．独自に作成したテストデータの一部を付録に示す．

表 1 自動作成したデータの例

キーワード	正解に対応する生成文
ビーチ, 人, 立つ	ビーチで数人の人が立っている
バット, 打つ, 球	子供がバットで球を打とうとしている
冷蔵庫, 置く, リフォーム	リフォーム中のキッチンに真新しい冷蔵庫が置いてある

4 実験

日本語 CommonGen のモデルは論文 [1] で構築されたモデルを用いる．モデルはウェブ上で公開されている日本語 T5 事前学習済みモデル[4] を前述したデータセットで fine-tuning したものである．学習時のパラメータは学習率を 3e-4，最大入力トークン数を 16，最大出力トークン数を 24，バッチサイズ 8，エポック数 10 としている．

4.1 自動作成したテストデータ

「自動作成したテストデータ」に対する上記モデルによる生成結果について述べる[5]．

「自動作成したテストデータ」は 200 組であり，各生成文に対して主観により 5 段階の評価基準のいずれか 1 つを付与した．付録に 5 段階の評価基準として提示されていた例を示す．分類結果を図 2 に示す．

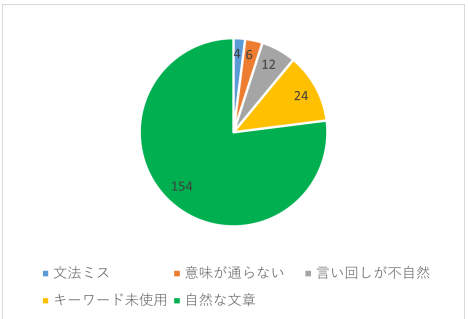

図 2 「自動作成したテストデータ」の評価結果

図 2 より約 8 割は評価基準の (v)（自然な生成文）と判定されている．このことから T5 モデルを利用することで，日本語においても CommonGen タスク

1) https://www.pencil.elyza.ai/
2) https://github.com/STAIR-Lab-CIT/STAIR-captions
3) ”する”, ”いる”, ”ある” など多機能な動詞は除かれている．
4) https://huggingface.co/sonoisa/t5-base-japanese/discussions
5) この結果は論文 [1] の結果を refine したものである．

は概ね解決できるといえる．また「妥当でない」と評価された生成文の内，前提条件である「キーワード未使用」を除くと最も割合が高いのは「言い回しの不自然さを含む文」であり，日本語においても常識推論が求められることを示している．

4.2 独自に作成したテストデータ

「独自に作成したテストデータ」に対する上記モデルによる生成結果について述べる[6]．この実験は入力キーワード群による生成の容易さを考察するために行っている．

独自に作成したキーワードと生成文の例として，「キーワード同士の関連度を段階的に変えたもの」の生成例を付録の表 5 に，「3 つのキーワードがすべて抽象名詞」の生成例を付録の表 6 に，「抽象名詞 2 つと具体的な名詞 1 つ」の生成例を付録の表 7 に示す．

キーワード同士の関連度を段階的に変えたものからは，現在のモデルではキーワード自体は含めることは出来ても 3 つのキーワード間に関連性がなければ妥当な文の生成は難しいことを示している．また学習データであるキャプションには出現しない抽象名詞をキーワードとしたとしても，意味を捉えた生成ができており，学習データの種類による特定の語彙を用いた生成能力の差は小さいのではないかと考えられる．抽象名詞 2 つと具体的な名詞 1 つの実験では，キーワード未使用が減少しており，キーワードの具体性がキーワード未使用に関連していることが予想できる．

4.3 ハブ単語の追加

論文 [1] では，独自に作成したテストデータの中の「3 つのキーワード全てが抽象名詞」に対する生成結果が妥当でないと判断した 10 組に関して，3 つのキーワードの中で関連性が低いキーワードを一つ任意の単語に変更する実験とハブ単語を追加する実験を行い生成結果の妥当性の変化を検証している．それぞれの実験の結果を表 8，表 9 を付録に示す．任意のキーワードを変更する実験では全体的にキーワードが未使用であるケースが減少し，生成文に関しても一部を除き妥当性が増したと述べられている．ハブ単語を追加する実験では，一部の生成結果でキーワード未使用や生成文の妥当性の改善が見られ，効果的な手法である可能性が示された．

6） この結果は論文 [1] の結果を refine したものである．

4.4 ハブ単語の自動選出

前述したハブ単語を追加する実験では，追加するハブ単語は人手により与えられた．本論文では入力キーワード群からハブ単語を自動選出する手法を提案し，その効果を調査する．

ハブ単語の選出は以下の手順で行う．まず chiVe[7] を用いて各入力キーワードの分散表現を得て，それらの平均ベクトルを求める．その平均ベクトルと類似度が高い上位 3 単語を chiVe の語彙から選択する．この例を表 2 に示す．表 2，2 行目左の列が入力キーワード群「表情，ミス，ショット」であり，このキーワードの分散表現の平均ベクトルと類似度が高い上位 3 単語が表 2 の 3，4，5 行目左の列の最後のキーワードである「ナイスショット，ミスショット，凡ミス」である．そして表 2 の 3，4，5 行目，右の列が追加後の生成文となっている．

表 2 ハブ単語の追加例

キーワード群	生成文
表情, ショット, ミス	テニスコートでミスをしてしまい悔しそうな表情をしている男性
表情, ショット, ミス, ナイスショット	ナイスショットをしようとしている黒猫の表情が強張っている
表情, ショット, ミス, ミスショット	ミスショットをしてしまい悔しそうな表情をしている女性
表情, ショット, ミス, 凡ミス	ミスをしてしまった瞬間の黒猫の表情

この操作を 3 節で説明した「自動作成したテストデータ」と「独自に作成したテストデータ」の内，生成文に全てのキーワードを含めていないキーワード群と関連研究で [1] で自然でないと評価したキーワード群に対して実行し，追加前と追加後の生成文の比較を行った．

全てのキーワードを含めていないキーワード群に対して比較結果は，「自動作成したテストデータ」では平均して 0.450 個，「独自に作成したテストデータ」では平均して 0.148 個の使用キーワード数の増加となった．またその例を表 3 に示す．

次に自動作成したテストデータ内，自然でないと評価したテストデータにおいてキーワードを追加した実験結果の例を表 4 に示す．主観評価となるが，3 割程度の生成結果で改善が見られた．

5 考察

ハブ単語の自動選出の実験結果では自動作成したテストデータと独自に作成したテストデータの間の

7） https://github.com/WorksApplications/chiVe

表 3 キーワード未使用と評価されたテストデータに対するタブ単語の追加例

キーワード群	生成文
こちら, 向こう, 走る	道路の向こうでシマウマが走っている
こちら, 向こう, 走る, 走り抜ける	柵の向こうを走り抜けてきた馬とこちらを見ている男性
こちら, 向こう, 走る, 向こう側	道路の向こう側にシマウマが走っている
こちら, 向こう, 走る, 歩く	道路の向こうで馬が二頭こちらを向いて走っている
周り, 両手, 模様	壁にカラフルな凧を描いてある
周り, 両手, 模様, 両腕	顔の周りに模様のついた両手を挙げている男性がいる
周り, 両手, 模様, 周囲	象の周りを両手で持つ模様のあるもの
周り, 両手, 模様, 背中	模様のついたリュックを背負った男性が背中を向けている
愛情, 最高, 感動	家族に最高の愛情でいっぱいのプレゼントをもらった
愛情, 最高, 感動, 感激	犬に愛情をこめて作ったピザに感激している
愛情, 最高, 感動, 素晴らしい	家族に最高の愛情でケーキを頂いて感動している
愛情, 最高, 感動, 感動的	家族が最高に感動した瞬間だった

表 4 不自然な生成文と評価されたテストデータに対するタブ単語の追加例

キーワード群	生成文
しまうま, 並ぶ, 鼻	しまうまと象が鼻が並んでいる
しまうま, 並ぶ, 鼻, 鼻先	しまうまが二頭鼻先が並んでいる
しまうま, 並ぶ, 鼻, 鼻筋	鼻筋が弱いしまうまが二頭並んでいる
しまうま, 並ぶ, 鼻, 顔	鼻が高いしまうまが 2 匹並んでいる
こちら, 路線バス, 連結	2 階建ての路線バスがこちらを連結している
こちら, 路線バス, 連結, 下方	路線バスが下方に連結されている
こちら, 路線バス, 連結, 端部	路線バスが端部のみに連結されている
こちら, 路線バス, 連結, 部材	路線バスに使われている部材がこちらに連結されている
ぶら下がる, 信号機, 手前	手前の信号機にオレンジ色の光がぶら下がっている
ぶら下がる, 信号機, 手前, 踏切	踏切の手前に赤い信号機がぶら下がっている
ぶら下がる, 信号機, 手前, 歩道	歩道の手前に黄色い信号機がぶら下がっている
ぶら下がる, 信号機, 手前, 踏み切り	踏み切りの信号機が手前にぶら下がっている

使用キーワード増加数に 0.3 キーワード程の差があるが，この原因として独自に作成したテストデータの生成難易度が高いことが一つの要因だと考える．根拠として，この実験の結果を検証した際，自動作成したテストデータでは多くのキーワード群で上位 3 単語の追加キーワードに対してどれか一つは使用キーワードの増加が見られていたが，独自に作成したテストデータでは，上位 3 単語の追加キーワード全てでキーワード使用数に変化がない，もしくはキーワード使用数の減少があるケースが多く見られたことに加え，Elyza Pencil に対して，3 つの追加キーワード全てで改善が見られなかったキーワード群を用いて記事を出力する実験を行った際，同様にキーワードを含む出力が見られないケースが多く見られたことが挙げられる．このことから，独自に作成したテストデータの一部に人間にとっても生成が難しいキーワード群が存在し，たとえ適切なアプローチを行っても改善を確認できないテストデータが存在する可能性を検証する必要がある．

また自動追加の手法に関してこの論文では，分散表現を用いてハブとなるキーワードの探索を行うという手法を採用したが，その他の手法としてキーワードを追加する前の生成文からキーワード以外の単語をキーワードとして追加する実験を行った．この手法は，キーワードが生成文に含まれないという問題に有効ではないかと考え実験を行ったが，追加前の生成文と追加後の生成文が殆ど同じになってしまう場合や追加キーワードが元となる 3 つのキーワードよりも優先されて含まれてしまうことが多く，生成文を変化させるには，追加前の生成文から関係が低いキーワードが望ましいと考える．

6 おわりに

論文 [1] で示された日本語 CommonGen のデータセットとベースラインシステムを利用し，論文内で示された入力キーワード群のハブ単語を追加することで生成文を改善する手法について機械によるハブとなる単語の自動追加システムを実装し，その効果を検証した．

キャプションを元としたテストデータでは，キーワードの未使用に対して一定の効果があることを示した．生成文の妥当性に対しては，主観となるが改善が見られることを確認した．

今後の課題としては，不自然な生成文の自動判定がある．生成文が自然かどうかを自動判定できなければ，本手法は利用できないからである．現在，不自然な生成文を収集し，生成文が自然かどうかを判定する分類器の構築を試みている．

謝辞

本研究は 2022 年度国立情報学研究所公募型共同研究（22FC04）の助成を受けています.

参考文献

[1] 鈴木 雅人, 新納 浩幸. 日本語 CommonGen の試作と入力単語間の関連性からの考察. 自然言語処理研究会 (第 253 回), 2022.

[2] Talmor Alon, Herzig Jonathan, Lourie Nicholas, and Berant Jonathan. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4149–4158. Association for Computational Linguistics, 2019.

[3] Sap Maarten, Rashkin Hannah, Chen Derek, Le Bras Ronan, and Choi Yejin. Social iqa: Commonsense reasoning about social interactions. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 4463–4473. Association for Computational Linguistics, 2019.

[4] Sakaguchi Keisuke, Le Bras Ronan, Bhagavatula Chandra, and Choi Yejin. Winogrande: An adversarial winograd schema challenge at scale, 2019.

[5] Omura Kazumasa, Kawahara Daisuke, and Kurohashi Sadao. A method for building a commonsense inference dataset based on basic events. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2450–2460. Association for Computational Linguistics, 2020.

[6] Zellers Rowan, Bisk Yonatan, Schwartz Roy, and Choi Yejin. Swag: A large-scale adversarial dataset for grounded commonsense inference. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 93–104. Association for Computational Linguistics, 2018.

[7] Zellers Rowan, Holtzman Ari, Bisk Yonatan, Farhadi Ali, and Choi Yejin. 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, 2019.

[8] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 1823–1840. Association for Computational Linguistics, 2020.

A　生成例

表 5　「キーワード同士の関連度を段階的に変えたもの」の生成例

キーワード	生成文
先生, 生徒, 教科書	生徒と先生と生徒が教科書を見ている
音楽, 配送, パン	配送中のパンに音楽がかかっている
テーブル, デート, 動画	テーブルの上でデートの動画が放送されている

表 6　「3 つのキーワードがすべて抽象名詞」の生成例

キーワード	生成文
愛情, 最高, 感動	家族に最高の愛情でいっぱいのプレゼントをもらった
哀愁, 故郷, 風景	故郷の風景と重なるところが多く哀愁を感じる
好み, 配慮, 選択	家族の意向を配慮したチョイスである

表 7　「抽象名詞 2 つと具体的な名詞 1 つ」の生成例

キーワード	生成文
愛情, 最高, テレビ	家族が最高に愛情たっぷりのテレビを見ている
哀愁, 故郷, 置き物	故郷を哀愁漂うさせる置き物だ
好み, 配慮, パン	ご主人の好みに配慮したパンが置いてある

表 8　表 6 で生成が妥当でないと評価した組み合わせについて
キーワードを一つ関連度の高そうなものに変更した結果

キーワード	生成文
愛情, 映画, 感動	家族に愛情たっぷりの映画で感動した瞬間
好み, 配慮, 選ぶ	家族に配慮して好みの焼き菓子を選んでいる
家族, 別れ, 心配	家族が別れの電話をして別れを心配している

表 9　表 6 で生成が妥当でないと評価されたキーワード群にハブ単語を追加した実験結果

キーワード	生成文
愛情, 最高, 感動, 映画	家族が最高に感動した映画がある
好み, 配慮, 選択, 食事	食事の好みを配慮して女性が選択している
悲しみ, 別れ, 心配, 電話	悲しみに暮れている男性が電話で別れの心配をしている

B　独自に作成したテストデータ例

タイプ 1　キーワード同士の関連度を段階的に変えたもの（20 組）
（先生, 生徒, 教科書），（音楽, 配送, パン），（テーブル, デート, 動画）

タイプ 2　3 つのキーワードがすべて抽象名詞（20 組）
（愛情, 最高, 感動），（哀愁, 故郷, 風景），（好み, 配慮, 選択）

タイプ 3　（2）の内 1 つのキーワードを具体的な名詞に変更したもの（抽象名詞 2 つと具体的な名詞 1 つ）（20 組）
（愛情, 最高, テレビ），（哀愁, 故郷, 置き物），（好み, 配慮, パン）

C　キャプションから作成のテストデータに対する 5 段階評価の分類例

(i)生成文に対して未使用なキーワードを含む
(入力): 携帯，写真，見える
(出力): 携帯を触っている男性の顔がぼやけて見える

(ii)意味の通らない生成文
(入力): くちばし，目，鷲
(出力): 鷲の目の近くにあるくちばしの大きな鳥がいる

(iii)文法的な誤りを含む生成文
(入力): しまうま，並ぶ，鼻
(出力): しまうまと象が鼻が並んでいる

(iv)言い回しの不自然さを含む生成文
(入力): 座る，道端，植木
(出力): 道端に小さな植木が座っている

(v)自然な生成文
(入力): 目，動物，茶色
(出力): 茶色の目をした動物が草を食べている