Multilingual CommonsenseQA

坂井 優介 上垣外 英剛 渡辺 太郎
奈良先端科学技術大学院大学
{sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

概要

言語モデルの自然言語理解能力を測る評価用データセットは多くの言語で不足している．また言語モデルのマルチリンガル性能に焦点を当てたとき，多言語間で対応の取れたデータセットは希少なため，その評価は限定的である．しかしデータセットを人手で構築するには限りがある．この問題を解決するため，本研究ではデータセットの作成過程を多段階に分割し，従来人手で作問していた工程を生成型マルチリンガル言語モデルに置換することで，効率的にマルチリンガルデータセットを作成する方法を提案する．本研究では CommonsenseQA に焦点を当て，提案手法を用いて 8 言語に拡張する．作成したデータセットは作成に使用した言語モデル自身にとっても十分難易度の高いデータセットとなった．

1 はじめに

言語モデルのマルチリンガル性能の評価に使用されるデータセットは主に，英語などで作成されたデータセットの翻訳 [16, 22, 9, 3, 32]，多言語間における類似タスクの組み合わせ [35, 12, 2, 23, 17]，多言語資源から同一手順に従って作成 [14, 13, 4, 6, 26] の 3 種類の方法で作成される（表 1）．しかし翻訳で作成されたデータセットは，翻訳由来の不自然さがあったり，言語固有の文化や知識・常識の違いが考慮されていない [16, 1, 6, 21, 15]．また多言語で類似タスクを組み合わせて作成されたマルチリンガルデータセットは，各言語ごとにデータ作成されているため，言語ごとの知識や常識の評価が可能である一方，言語数を増やすごとに完全に対応の取れたタスクは稀になり，作成過程やデータ取得元などの差異により，厳密な言語横断的評価はできなくなる．よって多言語資源から同一手順に従って作成されたマルチリンガルデータセットのみが，言語ごとの言語固有の知識や常識の違いを考慮した言語横断的評価を可能とする．しかし，そのようなデータセットを人手で作成することは作業者の確保や金銭的な観点から作成能力に限りがある．また，データセット作成時に言語資源が知識ベースなど構造化データの場合，それらを元に自然言語文を作成する必要がある．多言語環境において構造化された言語資源は，言語間で共通ルールに従って開発されており，言語間で対応が取れている．そのため言語によらず同一手順でデータセット作成が可能である．

表 1 マルチリンガルデータセットの作成方法の比較

作成方法	言語特有の知識	言語間の対応	作成コスト
翻訳	✗	✓	✓
類似タスク	✓	✗	✓
多言語資源	✓	✓	✗
提案手法	✓	✓	✓

本研究では，従来人手によって構造化された言語資源からデータセット作成していた工程を生成型マルチリンガル言語モデルに置換することで，低コストで効率的にマルチリンガルデータセットを作成する方法を提案する．本研究では，常識推論能力評価用データセットである CommonsenseQA（CSQA）[28] に焦点を当てる．CSQA は構造化された多言語知識ベースである ConceptNet [27] から人手によって作成された代表的な常識推論 QA データセットであるが，コスト的な問題から現状，英語と日本語版の JCommonsenseQA（JCSQA）[15] のみしか作成されていない．そのため提案手法を用いて CSQA を 8 言語[1)]に拡張した Multilingual CommonsenseQA（mCSQA）を作成する．mCSQA により言語モデルの言語横断的な常識理解能力を測ることが可能となる．提案手法により，十分に設計された作問手順に従って作成したデータセットは，作成に使用した言語モデル自身にとっても難易度の高いデータセットとなることがわかった．また mCSQA の 1 問あたりの作成金額は CSQA の 100 分の 1 まで削減できた．

1) 英語（en; English），日本語（ja; Japanese），中国語（zh; Chinese），独語（de; German），ポルトガル語（pt; Portuguese），オランダ語（nl; Dutch），仏語（fr; French），ロシア語（ru;Russian）

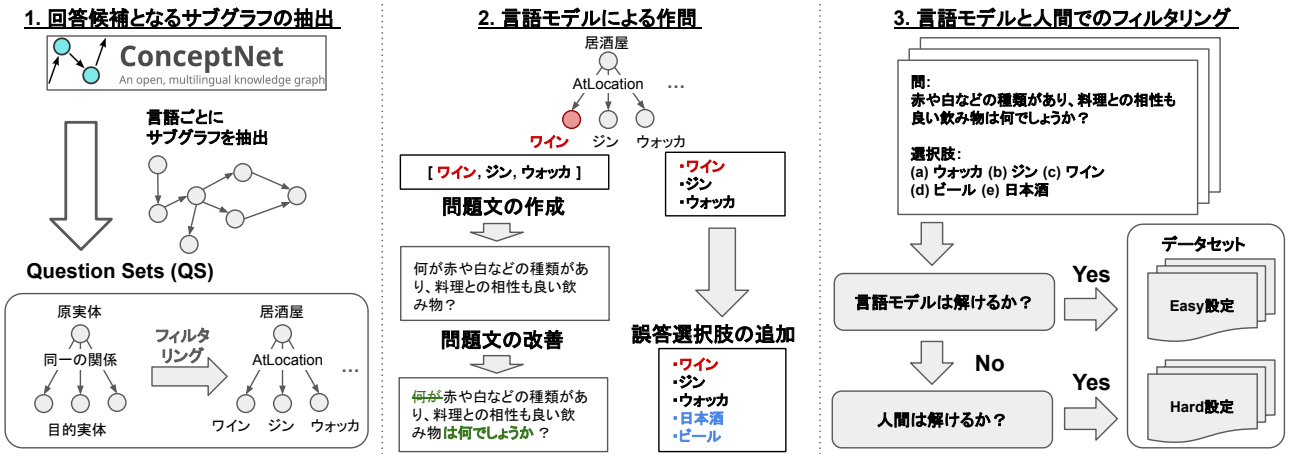

図 1　Multilingual CommonsenseQA の作成過程の概略図

2　関連研究

CSQA はある原実体と同じ関係名から派生する 3 つの異なる目的実体を持つようなサブグラフを ConceptNet から抽出する．3 つの目的実体を選択肢とみなし，そのうち 1 つのみが正解となるような問題文をクラウドワーカーによって作問する．その後，誤答選択肢を 2 つ追加することでデータセットを構築する．CSQA は知識ベースにおける関係の方向性を考慮していたが，JCSQA では多様な問題を作成するために逆方向の関係も用いる．XCSQA [16] は CSQA を機械翻訳で拡張した多言語データセットだが，機械翻訳の影響で言語固有の常識の評価は考慮されない．また，これらデータセットを言語モデルにとって，より難易度の高い問題にする研究も行われている [29, 36]．このように言語モデルの常識理解能力は基本的に質問応答形式で測られることが多い [33, 34, 18, 24, 25, 5]．また常識理解のためのデータセットを自動的に作成する試みも行われている．文献 [36, 33, 34] では与えられた問題文から言語モデルで回答選択肢を生成する．しかしマルチリンガルデータセットに焦点を当てたとき言語モデルによるデータセット自動作成はまだ発展段階である．

3　mCSQA の作成方法

図 1 に示すように mCSQA の作成過程は回答候補となるサブグラフの抽出，言語モデルによる作問，言語モデルと人間でのフィルタリングの 3 段階で行う．CSQA と JCSQA の作成手順を参考としつつ，多言語で共通処理を行えるよう修正を加えた．詳細な設定などは付録 A を参照．なおデータセット作成には GPT-3.5 [20]（gpt-3.5-turbo-1106）を用いた．

3.1　ConceptNet から回答候補の抽出

原実体と関係のクエリから導出可能な 3 つの異なる目的実体を持つようなサブグラフを，各言語ごとに ConceptNet から抽出する（図 1-1）．また JCSQA と同様に順方向の関係だけでなく逆方向の関係も使用する．このサブグラフを Question Set（QS）と名付ける．CSQA と同様に QS のノイズ除去のためにフィルタリングする．mCSQA では多言語で共通処理を行えるよう以下のフィルタリングを行う．

1. CSQA や JCSQA で選択された 22 種類の関係が含まれる QS のみ保持．
2. 各実体が 5 単語以上，または 1 文字のみで構成される実体を含む QS を削除．
3. 同義語や近い表現を除去するため，目的実体同士が ConceptNet 内で Synonym の関係である QS と部分文字列である QS を削除．

フィルタリング後，各言語それぞれ 6,000 件の QS をランダムに選択した[2)]．

3.2　言語モデルによる作問

CSQA ではクラウドソーシングを用いて人手で作問するが，本研究では生成型マルチリンガル言語モデルである GPT-3.5 で代替する．CSQA では「問題文の作成」，「誤答選択肢の追加」の 2 つの工程に分割していたが，本研究では「問題文の改善」の工程を追加し，合計 3 工程で作問する（図 1-2）．各工程で使用したハイパーパラメタは付録の表 4 を参照．

問題文の作成　各 QS の 3 つの目的実体の内，1 つのみ正解となる問題文を生成する．言語モデルへ

2)　フランス語とロシア語の QS は 6,000 件に満たなかったため，全件採用し，それぞれ 4,125 件，3,901 件であった．

の指示内容は以下のとおりである．

1. 問題文に選択肢の単語を使用しない
2. 文字数などの表層的な情報の使用を避ける
3. 文末は疑問符（？）で終わる
4. 客観的な問題である
5. 1 文のみで構成

作問後，指示に従わなかったり，不適切な表現が含まれる問題文をパターンマッチングで除去する．

問題文の改善　言語モデルが作問した問題文には不自然な言い回しなどが含まれていることが判明した．この問題に対処するため，生成された問題文を言語モデルにより，意味的・文法的に正しい文へと改善を行う．改善後の結果に対して，再び問題文の作成時と同様のフィルタリングを行い，不適切な問題を除去する．修正した割合は付録の表 5 を参照．

誤答選択肢の追加　問題をより難しくするために，2 つの誤答選択肢の追加を行う．CSQA では問題文に関連しているか，もっともらしい選択肢を追加するようクラウドワーカーに依頼している．本研究では 3 つの選択肢のみ使用し，それらに関連するもっともらしい 2 つの誤答選択肢を言語モデルで作成した．問題文を提示しないことで，誤答選択肢の追加時に問題に対する回答能力と作問能力の分離が可能となる．しかしこの過程で，問題文を提示しないことで，誤答とはならず解答と成りうる選択肢が追加される可能性がある．そのため 3.3 節でフィルタリングを行い除去する．また追加した 2 つの誤答選択肢が既存の選択肢と重複している問題と，選択肢の言葉が問題文に含まれている問題を削除した．

3.3　言語モデルと人間でのフィルタリング

CSQA や JCSQA では複数の選択肢が回答と成り得たり，回答不可能な低品質な問題を除去するため，人手で全件確認しているが，mCSQA ではデータ数が多いため全件確認することは現実的ではない．そこで本研究では能動学習の枠組みを活用し，はじめに言語モデルが解答可能か検証し，言語モデルが解答できない問題のみ人手で確認する（図 1-3）．

言語モデルによるフィルタリング　作成したデータセットには，言語モデルが解ける問題，人間は解けるが言語モデルは解けない問題，問題に不備があるため解けない問題の 3 種類が含まれる．そこで問題作成に使用した言語モデル自身に問題を解答させることで，言語モデルが解ける問題のみ抽出し，残りを人手で精査することで，問題に不備があるため解けない問題のみデータセットから削除する．

人手によるフィルタリング　Amazon Mechanical Turk（MTurk）を使用し，各問題ごとに 2 人のクラウドワーカーを割り当てる．ワーカーには問題文と選択肢と解答候補を提示し，その解答候補が問題から解答可能であると全員返答した問題のみ採用した．

表 2　Multilingual CommonsenseQA の内訳

	Train	Dev			Test		
言語名	Total	Easy	Hard	Total	Easy	Hard	Total
English	10,910	1,071	292	1,363	1,071	292	1,363
Japanese	11,696	1,117	344	1,461	1,117	344	1,461
Chinese	12,159	972	546	1,518	972	546	1,518
German	12,504	1,279	283	1,562	1,279	283	1,562
Portuguese	12,659	1,234	348	1,582	1,234	348	1,582
Dutch	12,215	1,255	271	1,526	1,255	271	1,526
French	8,047	786	219	1,005	786	219	1,005
Russian	6,623	445	382	827	445	382	827

3.4　データセットの分割方法

CSQA のデータ分割方法に従い，各言語ごとにランダムに 8 対 1 対 1 の割合で学習，検証，テストデータに分割した．また 3.3 節で言語モデルが解答できた問題を Easy 設定，人手によるフィルタリング後，採用された問題を Hard 設定とし，検証データとテストデータのみ明示的に区別する．最終的な問題数の内訳を表 2 に示す．各工程でフィルタリングされた割合は付録の図 3 を参照．1 問あたりの作成コストは CSQA が 0.33 ドルに対し，mCSQA は 0.002 ドルとなった．金額の内訳は付録の図 4 を参照．

4　作成したデータセットの評価

4.1　実験設定

言語モデルの実験設定　使用した言語モデルの詳細は付録表 6 を参照．Encoder 型言語モデルは付録表 7 の設定で訓練を行い，最良の結果を選択した．Decoder 型言語モデルは訓練を行わず，0-shot 設定と 3-shot 設定[3]で推論した．GPT-3.5 と GPT-4 は topp, temperature を 0 に設定した．Llama2-70B は貪欲法．

人手による評価　CSQA の設定に従い，各言語ごとに検証データとテストデータからそれぞれランダムに 100 問抽出した．MTurk 上で各言語ごとに 5 人のクラウドワーカーを募った．JCSQA の設定に従い，各問題の最終的な解答は収集した解答の多数決

3）　Easy 設定と Hard 設定の事例が最低 1 つずつ含まれる．

表 3　Multilingual CommonsenseQA の実験結果（正解率%）

	English		Japanese		Chinese		German		Portuguese		Dutch		French		Russian	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Human (Rand. 100 sent.)	87.0	93.0	89.0	95.0	91.0	87.0	96.0	96.0	93.0	93.0	98.0	97.0	96.0	92.0	87.0	94.0
mBERT-cased [10]	60.6	61.3	66.0	63.5	65.9	63.5	58.6	57.9	65.2	61.5	54.8	57.8	46.3	47.3	32.2	31.3
mBERT-uncased [10]	63.4	65.2	61.3	58.9	64.0	62.0	59.3	60.3	67.6	63.9	57.3	56.9	51.1	52.4	32.5	34.0
XLM-100 [8]	57.2	59.0	60.2	58.8	60.0	61.5	54.4	54.7	62.7	59.5	52.2	52.0	35.3	35.0	23.2	26.0
XLM-R_{BASE} [7]	68.0	69.1	68.5	66.2	69.8	68.3	63.9	62.8	69.5	67.3	62.0	64.0	47.6	45.5	36.9	37.0
XLM-R_{LARGE} [7]	77.2	77.5	75.7	72.6	**75.0**	**74.1**	76.2	75.4	79.0	76.4	73.0	74.7	62.0	62.3	48.9	48.6
mDeBERTa-v3 [7]	76.6	79.2	77.2	74.1	74.6	72.0	75.7	77.5	78.3	78.2	72.7	74.9	62.1	62.4	51.3	49.9
Llama2-70B (0-shot) [30]	48.1	47.7	25.6	24.8	26.5	25.9	32.5	32.7	38.7	37.6	40.9	39.4	42.3	44.1	23.5	22.9
Llama2-70B (3-shot) [30]	57.1	55.5	47.4	46.6	33.3	30.2	63.1	62.9	65.0	63.7	60.8	62.3	57.8	56.7	30.8	32.3
GPT-3.5 (0-shot) [20]	76.7	77.0	76.3	76.7	64.0	63.6	81.3	81.4	77.9	77.7	82.1	81.5	78.6	77.1	53.3	**53.0**
GPT-3.5 (3-shot) [20]	77.2	78.4	77.5	77.0	65.3	64.3	**83.2**	81.4	78.5	78.0	81.8	80.5	78.4	76.5	**54.1**	50.1
GPT-4 (0-shot) [19]	**80.9**	80.9	78.4	77.2	66.0	65.6	81.0	81.0	78.6	77.6	**83.4**	81.5	78.8	77.0	49.9	47.8
GPT-4 (3-shot) [19]	80.5	**81.0**	**78.5**	**77.5**	67.2	66.9	82.6	**81.6**	**80.5**	**78.8**	83.3	**81.6**	**79.0**	**77.4**	50.1	48.9

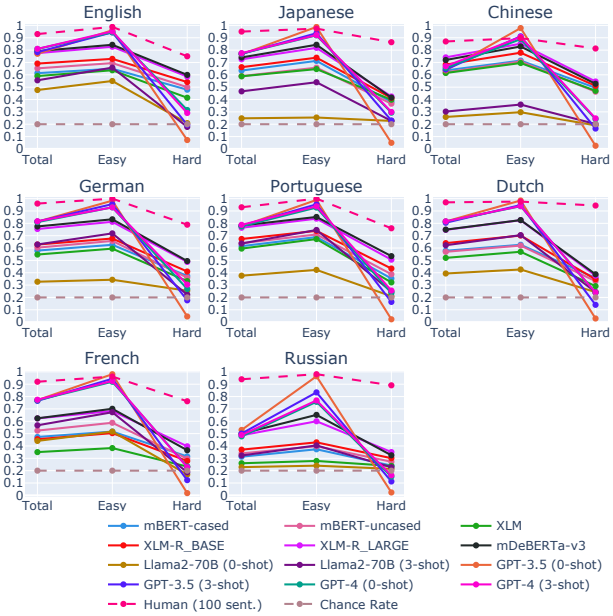

図 2　Easy 設定と Hard 設定における評価結果の比較

とし，同数の場合はランダムで選択する．

4.2　実験結果

表 3 に示す実験結果では，各言語モデルと人間の評価結果に隔たりがある．さらに Decoder モデルに着目すると，ほとんどの結果で 0-shot 時より 3-shot 時のほうが良い評価結果となる．また本手法で使用した GPT-3.5 は，ドイツ語とロシア語以外の言語では Encoder モデルの最も良い結果と比較して，同等以下の評価結果である．また GPT-4 の結果と比較した場合，ほとんどの言語に対して，GPT-3.5 は評価結果が劣っている．このことから GPT-3.5 が解答できない問題は，自身が有している知識だけでは解けない問題だと言える．以上より，作問に使用した言語モデル自身にとっても十分難易度の高いデータセットが作成できたため，本手法は有効である．さらなる分析については付録 C を参照いただきたい．

4.3　Easy 設定と Hard 設定の比較

言語モデルと人間それぞれのフィルタリングによる問題の難易度比較を行う．テストデータにおける Easy 設定と Hard 設定の評価結果の比較を図 2 に示す．Easy 設定の場合，特に GPT-3.5 と GPT-4 はほぼ正答できるが，Hard 設定の場合，人間の結果と大きく隔たりがある．また他の言語モデルでも，Hard 設定では Easy 設定より評価結果が低下している．このことから言語モデルが問題を作成できたとしても，それが言語モデル自身が解けるとは限らないと言えるため，作問能力と解答能力は切り分けて考えられる．よって構造化データからデータセットを作成する際，一部の工程において言語モデルは人間の代替として機能することが可能だとわかった．

5　おわりに

本研究では CSQA の作成手順を参考に，構造化された多言語資源 ConceptNet から，生成型マルチリンガル言語モデルを用いて，マルチリンガルデータセット mCSQA を作成した．また作問過程で作問能力と解答能力を切り分けたことで，作成したデータセットは作成に使用した言語モデル自身にとっても十分難易度の高いデータセットとなった．今後の展望として，多言語資源の言語間での質や量の不均衡への解決や，依然として人件費が高額（付録図 4）なため，さらに効率的な手法の開発の余地がある．

参考文献

[1] Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. Towards an atlas of cultural commonsense for machine reasoning, 2020.

[2] David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 4488–4508, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[3] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics.

[4] Sabine Buchholz and Erwin Marsi. CoNLL-X shared task on multilingual dependency parsing. In **Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)**, pp. 149–164, New York City, June 2006. Association for Computational Linguistics.

[5] Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. CODAH: An adversarially-authored question answering dataset for common sense. In **Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP**, pp. 63–69, Minneapolis, USA, June 2019. Association for Computational Linguistics.

[6] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 454–470, 2020.

[7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[8] Alexis CONNEAU and Guillaume Lample. Cross-lingual language model pretraining. In **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[9] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In **The Eleventh International Conference on Learning Representations**, 2023.

[12] Hai Hu, Ziyin Zhang, Weifang Huang, Jackie Yan-Ki Lai, Aini Li, Yina Patterson, Jiahui Huang, Peng Zhang, Chien-Jer Charles Lin, and Rui Wang. Revisiting acceptability judgements, 2023.

[13] Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 1440–1448, Marseille, France, May 2020. European Language Resources Association.

[14] Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual Amazon reviews corpus. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4563–4568, Online, November 2020. Association for Computational Linguistics.

[15] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[16] Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1274–1287, Online, August 2021. Association for Computational Linguistics.

[17] Shervin Malmasi and Mark Dras. Large-scale native language identification with cross-corpus evaluation. In **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1403–1409, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[18] Kazumasa Omura, Daisuke Kawahara, and Sadao Kurohashi. 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, Online, November 2020. Association for Computational Linguistics.

[19] OpenAI. GPT-4 Technical Report, 2023.

[20] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744. Curran Associates, Inc., 2022.

[21] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. KLUE: Korean language understanding evaluation. In **Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)**, 2021.

[22] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2362–2376, Online, November 2020. Association for Computational Linguistics.

[23] Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5919–5930, Online, November 2020. Association for Computational Linguistics.

[24] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. **Commun. ACM**, Vol. 64, No. 9, p. 99–106, aug 2021.

[25] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 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, Hong Kong, China, November 2019. Association for Computational Linguistics.

[26] Holger Schwenk and Xian Li. A corpus for multilingual document classification in eight languages. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[27] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. In **Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence**, AAAI'17, p. 4444–4451. AAAI Press, 2017.

[28] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 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, Volume 1 (Long and Short Papers)**, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[29] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In **Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)**, 2021.

[30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[31] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[32] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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. 3687–3692, Hong Kong, China, November 2019. Association for Computational Linguistics.

[33] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 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, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[34] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.

[35] Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. Mela: Multilingual evaluation of linguistic acceptability, 2023.

[36] 栗原健太郎, 河原大輔, 柴田知秀. JCommonsenseQA 2.0: 計算機と人の協働による常識推論データセットの改良. 言語処理学会第 29 回年次大会発表論文集, pp. 2908–2913, March 2023.

表 4　作問時の言語モデルのハイパーパラメタ

工程	temperature	top_p	seed
問題文の作成	0.0	0.0	0
問題文の改善	0.7	0.5	0
誤答選択肢の追加	1.2	0.7	0

表 5　修正された問題文の割合

	en	ja	zh	de	pt	nl	fr	ru
全体	14,722	15,695	17,254	16,542	16,679	15,992	10,770	10,215
修正文	3,654	12,007	6,534	765	585	7,927	3,109	6,734
割合 (%)	24.82	76.50	37.87	4.63	3.51	49.57	28.87	65.92

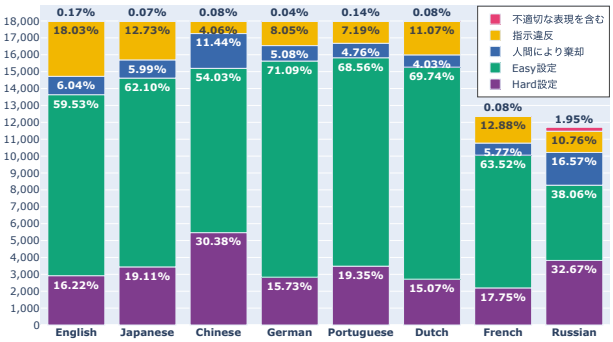

図 3　各工程で処理された文数の割合．Easy 設定と Hard 設定が最終的にデータセットとして採用され，その他は生成過程で除去された．

A　mCSQA の作成方法の詳細

作成したデータセットは https://huggingface.co/datasets/yusuke1997/mCSQA にて公開予定．各工程で使用したプロンプトは紙面の都合上，https://github.com/yusuke1997/mCSQA/blob/main/base_template.py に全て掲載している．3.2 節の各工程における作問時のハイパーパラメタを表 4 に示す．また，「問題文の改善」パートによって，修正された問題文の数と割合を表 5 に示す．

3 章の各工程で処理された問題数を図 3 に示す．各言語・各工程ごとに使用した金額の割合を図 4 に示す．

B　実験設定の詳細

実験に使用した言語モデルを表 6 に，設定したパラメータを表 7 に示す．

C　言語モデルの言語転移性能

言語転移性能の評価は言語資源的な理由で英語から多言語への評価が行われることが多い．そのため，本節ではマルチリンガル言語モデルにおいて，英語以外の言語で学習されたモデルの転移性能を測定することで，各言語に特化した学習が必要か議論する．XLM-R_{LARGE} を用いて，表 7 の設定で各言語ごとにそれぞれ学習を行い，8 言語全てのテストデータで評価を行った．図 5 に示す実験結果は，言語転移性能評価において，各言語ごとに学習したモデルの評価結果からどの程度下落するかを示している．図 5 の結果より，概ね各言語ごとにそれぞれ学習させた場合に最も良い結果となりつつも，どの言語で学習しても一定の言語転移能力の獲得が確認できる．しかし Hard 設定では Easy 設定と比較して言語転移性能は低い．この結果より，人間の判断が深い背景知識が必要な問題に対して，言語固有の言語モデルの開発や学習が必要であることがわかった．

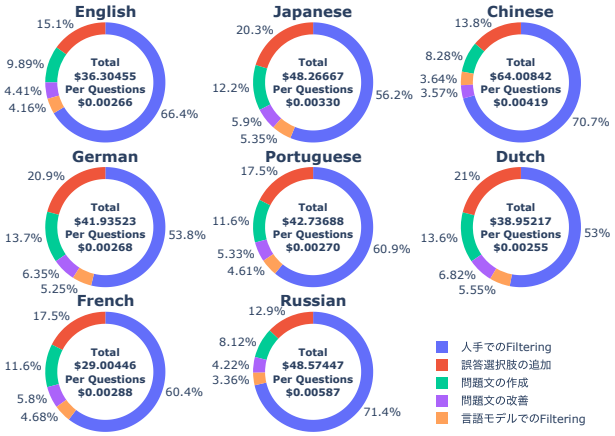

図 4　各言語・各工程ごとに使用した金額の割合と合計値

表 6　検証に使用したマルチリンガル言語モデル

種類	本稿でのモデル名	HuggingFace / OpenAI API
Encoder 型	mBERT-cased [10]	bert-base-multilingual-cased
	mBERT-uncased [10]	bert-base-multilingual-uncased
	XLM-100 [8]	xlm-mlm-100-1280
	XLM-R_{BASE} [7]	xlm-roberta-base
	XLM-R_{LARGE} [7]	xlm-roberta-large
	mDeBERTa-v3 [11]	microsoft/mdeberta-v3-base
Decoder 型	Llama2-70B [30]	meta-llama/Llama-2-70b-chat-hf
	GPT-3.5 [20]	gpt-3.5-turbo-1106
	GPT-4 [19]	gpt-4-1106-preview

表 7　実験に使用したハイパーパラメタ．その他のパラメタは標準設定を用いた．Transformers [31] を用いて実験．

ハイパーパラメタ名	値
Batch Size	64
Learning Rate	2e-5, 3e-5, 5e-5
Seed	42
Early Stopping	3
Warmup Ratio	0.1
Max Sequence Length	128

図 5　言語モデルの転移性能の評価．y 軸に示す各言語で Fine-tuning 後，x 軸方向の言語のテストデータで評価を行う．同一言語で学習と評価を行った場合と比較して，何割程度の評価性能になるか示している．