文対モデリングのための言い換えに基づく対照学習

杉山 誠治[1]　近藤 里咲[2]　梶原 智之[2]　二宮 崇[2]
[1] 愛媛大学工学部　[2] 愛媛大学大学院理工学研究科
{sugiyama@ai.cs., kondo@ai.cs., kajiwara@cs., ninomiya.takashi.mk@}ehime-u.ac.jp

概要

本研究では，文対モデリングタスクの性能改善のために，事前学習済みマスク言語モデルに対する追加事前学習の手法を提案する．マスク言語モデリングによる事前学習は，意味的に近い文の埋め込み同士を埋め込み空間上で必ずしも近づけるようには設計されていない．そこで提案手法では，事前学習済みマスク言語モデルに対して，言い換え文対の文埋め込みを近づける対照学習を適用し，文対モデリングの性能改善を目指す．対照学習の先行研究で標準的に使用される自然言語推論コーパスは，英語以外の言語では大規模に利用できない課題があるが，本手法では生コーパスと言い換え辞書から低コストに対照学習のための学習データを構築できる．4 種類の文対モデリングタスクにおける実験の結果，英日の両言語において提案手法の有効性を確認できた．

1　はじめに

2 文間の関係を推定する文対モデリングは，類似度推定 [1] や含意関係認識 [2] などの基礎タスクから情報検索 [3] や質問応答 [4] などの応用タスクまで，様々な自然言語処理において重要な技術である．文対モデリングでは，伝統的には Bag-of-Words などの表層マッチングや word2vec [5] などの埋め込みが使用され，その後タスクごとに設計されたニューラルネットワーク [6, 7] の検討を経て，近年は BERT [8] などの事前学習済みマスク言語モデルのファインチューニングが標準的に採用されている．しかし，マスク言語モデリングの事前学習では，意味的に近い文の埋め込み同士を埋め込み空間上で必ずしも近づけるわけではない [9]．そこで，文対モデリングタスクにおけるファインチューニングの性能を充分に引き出すためには，マスク言語モデリングの事前学習に続いて，テキスト間の意味的な関係を推定する追加学習 (Transfer Fine-Tuning [10]) を挟むことが有効であると知られている．

そのような追加学習のひとつに対照学習がある．意味的に近い文の埋め込み同士を埋め込み空間上で近づけるために，SimCSE [11] などの対照学習 [11–13] が盛んに研究されている．これらの文埋め込みの対照学習では，自然言語推論 (NLI: Natural Language Inference) のアノテーションコーパスを用いて，含意関係にある文対の埋め込みを近づけ，矛盾関係にある文対の埋め込みを遠ざける．しかし，英語では数十万文対の規模の Stanford NLI (SNLI) [2] や Multi-Genre NLI (MNLI) [14] を使用できる一方，他の言語で使用できる NLI コーパスは小規模であるため，英語以外の言語において対照学習による高品質な文埋め込みを得ることは難しい．

本研究では，文対モデリングの性能改善のために，NLI コーパスに依存しない対照学習の手法を提案する．提案手法では，生コーパスと言い換え辞書を用いて，対照学習のための学習データを低コストかつ大規模に自動生成する．言い換え辞書は多くの言語[1)]において使用できるため，本手法は英語以外の言語においても適用できる．

英語および日本語における評価実験の結果，提案手法は 4 種類の文対モデリングタスク（商品検索・類似度推定・含意関係認識・言い換え認識）においてマスク言語モデルの性能を改善した．また，全タスクの平均性能では，言い換えを学習するが対照学習を用いない既存手法 [10]，生コーパスを用いる対照学習 [11]，NLI コーパスを用いる対照学習 [11] と比較して，英語および日本語の両言語において提案手法が最高性能を達成した．

2　提案手法

本研究では，NLI コーパスに依存しない対照学習によって，マスク言語モデルによる文対モデリングの性能を改善する．本手法は，以下の手順に示すように，マスク言語モデルの事前学習とファイン

1）例えば Multilingual Paraphrase Database [15] は 23 言語において数百万から数億件の規模で言い換え対を収集している．

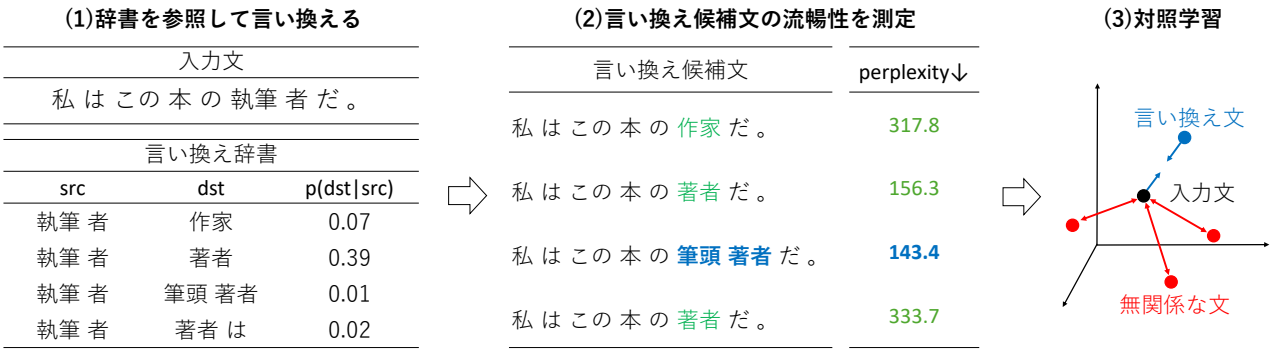

図 1　提案手法である言い換えベース対照学習の概要

チューニングの間で実施することにより，ファインチューニングの効果を高めるものである．

1. 事前学習：マスク言語モデリング
2. 追加学習：提案手法に基づく対照学習
3. ファインチューニング：文対モデリングの対象タスクにおける教師あり学習

図 1 に示すように，我々の対照学習は NLI コーパスにおける含意文対の代わりに言い換え文対を用いる．(1) 生コーパスから得た入力文を辞書を参照しつつ言い換え，(2) 言い換え候補の中から最も流暢な言い換え文を選択し，(3) 入力文と言い換え文の埋め込みを近づけ，入力文とバッチ内の他の文の埋め込みを遠ざける対照学習を実施する．

2.1　言い換え文対を用いる対照学習

提案手法では，SimCSE [11] と同様の対照学習を実施するが，対照学習の正例として含意文対の代わりに次節で説明する言い換え文対，負例として矛盾文対の代わりにバッチ内の他の文を用いる．本研究ではバッチ内の文間に意味的な関係を仮定しないため，バッチ内の他の文 x_j は入力文 x_i と意味的に無関係であると扱い，文埋め込みを遠ざけるべき負例として機能する．入力文 x_i の言い換え文を x_i^+，言い換え文対の埋め込みをそれぞれ $\mathbf{h}_i$ および $\mathbf{h}_i^+$ とし，式 (1) の損失関数を最小化するように学習する．

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h}_i,\mathbf{h}_j^+)/\tau}}, \tag{1}$$

ここで，N はバッチサイズ，τ は温度パラメータ，$\mathrm{sim}(\cdot)$ は文埋め込み間の余弦類似度である．

2.2　言い換え文対の自動生成

対照学習の正例として用いる言い換え文対は，生コーパスと言い換え辞書を用いて自動生成する．言い換え辞書は，言い換え元の語句 s，言い換え先の語句 d，言い換え確率 $p(d|s)$ の 3 つ組で構成され，本研究では閾値 θ 以上の言い換え確率を持つ言い換え対 $\{(s,d) \mid p(d|s) \geq \theta\}$ のみを扱う．生コーパスから得た入力文 $x_i \in \mathcal{D}$ に言い換え辞書を適用し，s から d へ語句を置換し，言い換え候補文を生成する．

ここで，図 1 の (2) に示すように，言い換え候補には文法的に不適格な表現が含まれ得る．これらの非文が対照学習に悪影響を与えることを防ぐために，最小の perplexity を持つ最も流暢な候補文を選択し，対照学習の正例として使用する．

このような言い換え文対は，機械翻訳モデルに基づく逆翻訳 [16] や折り返し翻訳 [17]，大規模言語モデルに基づく言い換え生成 [18] などの方法によっても獲得できる．これらの言い換え手法との比較は今後の課題であり，本研究では計算コストが低く解釈性が高い辞書ベースの言い換え手法を採用する．

3　評価実験

提案手法の有効性を評価するために，商品検索・類似度推定・含意関係認識・言い換え認識の 4 種類の文対モデリングタスクを対象として，英語および日本語の評価実験を実施した．

3.1　タスク

商品検索　商品名とその検索クエリの間の関連度を 4 値分類するタスクであり，MicroF1 によって評価した．データセットは，英語および日本語の Shopping Queries データセット[2] [19] を使用した．

類似度推定　2 文間の意味的類似度を推定する回帰タスクであり，スピアマンの順位相関係数によって評価した．データセットは，英語では STS-B[3] [1]

2)　https://github.com/amazon-science/esci-data
3)　http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

表 1　英語データセットの文対数

	学習用	検証用	評価用
Shopping Queries	1,254,438	138,625	425,762
STS-B	5,749	1,500	1,379
SICK	4,439	495	4,906
SNLI	549,367	9,842	9,824
PAWS	49,401	8,000	8,000

表 2　日本語データセットの文対数

	学習用	検証用	評価用
Shopping Queries	294,874	32,272	118,907
JSTS	11,205	1,246	1,457
JSICK	4,500	500	4,927
JNLI	18,065	2,008	2,434
PAWS-X	49,401	2,000	2,000

および SICK[4] [20]，日本語では JSTS[5] [21] および JSICK[6] [22] を使用した．

含意関係認識　2 文間の意味的関係を 3 値分類するタスクであり，MacroF1 によって評価した．データセットは，英語では SNLI[7] [2] および SICK [20]，日本語では JNLI[5] [21] および JSICK [22] を使用した．

言い換え認識　2 文間の同義性を判定する 2 値分類タスクであり，MacroF1 によって評価した．データセットは，英語では PAWS[8] [23]，日本語では PAWS-X[8] [24] を使用した．

データセットの統計情報を表 1 および表 2 に示す．本実験では，英語の BERT[9] [8] および日本語の RoBERTa[10] [25] を，これらの文対モデリングタスクにおいてファインチューニングした．ファインチューニングの前に提案手法または比較手法の追加学習を適用することによって，各タスクの性能を改善できるかどうかを評価した．

3.2　提案手法の実装の詳細

前処理　追加学習には生コーパスとして Wiki-40B[11] の Wikipedia テキストを使用した．前処理として，英語では Moses [26] による文分割[12] および単語分割[13]，日本語では ja_sentence_segmenter[14] による文分割および MeCab (IPAdic)[15] [27] による単語分割を実施した．また，langdetect[16] の言語判定によって，英語および日本語の各コーパスにおいて対象言語の確率が 99% 以上と判定された文のみを使用した．そして，5 単語以下の短文および 50 単語以上の長文を除外した．

言い換え　言い換え辞書には，英語では PPDB 2.0[17] [28]，日本語では EhiMerPPDB[18] [29] を使用した．これらの辞書には，英語では最長 6 単語，日本語では最長 7 単語までの語句が含まれている．言い換え候補のフィルタリングのために，GPT-2 [30] の英語モデル[19] および日本語モデル[20] を用いて perplexity を計算した．

ハイパーパラメータ　追加学習の際には，学習率を 5×10^{-5}，温度パラメータ $\tau = 0.05$，バッチサイズを 64 文対とし，最適化手法には Adam を使用して，検証用データの損失が 3 エポック連続で改善されない時点で学習を早期終了した．その他のハイパーパラメータとして，言い換え確率の閾値 $\theta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ および追加学習の文数 $|\mathcal{D}| \in \{10k, 20k, 40k, 80k, 160k\}$ を調整した．ファインチューニングの際には，これらの中から文対モデリングタスクの検証用データの評価値を最大化する組み合わせを選択した．

3.3　比較手法

提案手法である言い換えに基づく対照学習の有効性を確認するために，言い換えを用いるが対照学習ではない既存手法 (Transfer Fine-Tuning) [10]，言い換えを用いない対照学習の既存手法 (SimCSE) [11]，追加学習なしでファインチューニングするベースラインと比較する．Transfer Fine-Tuning は約 3,000 万件の言い換え対を用いて句の言い換え認識を追加学習

4) https://zenodo.org/records/2787612
5) https://github.com/yahoojapan/JGLUE
6) https://github.com/verypluming/JSICK
7) https://nlp.stanford.edu/projects/snli/
8) https://github.com/google-research-datasets/paws
9) https://huggingface.co/google-bert/bert-base-uncased
10) https://huggingface.co/rinna/japanese-roberta-base
11) https://www.tensorflow.org/datasets/catalog/wiki40b
12) https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/mosestokenizer/sentsplitter.py
13) https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl
14) https://github.com/wwwcojp/ja_sentence_segmenter
15) https://taku910.github.io/mecab/
16) https://pypi.org/project/langdetect/
17) http://paraphrase.org/#/download
18) https://github.com/EhimeNLP/EhiMerPPDB
19) https://huggingface.co/openai-community/gpt2
20) https://huggingface.co/rinna/japanese-gpt2-medium

表 3　4 種類の文対モデリングタスクの実験結果

英語	商品検索	類似度推定		含意関係認識		言い換え認識	平均
	Shopping Queries	STS-B	SICK	SNLI	SICK	PAWS	
追加学習なし	0.654	0.824	0.815	**0.904**	0.858	0.913	0.828
Transfer Fine-Tuning	0.652	0.854	0.821	0.902	0.860	0.901	0.832
教師なし SimCSE	**0.655**	0.830	0.806	**0.904**	**0.868**	**0.918**	0.830
教師あり SimCSE	**0.655**	**0.857**	0.824	0.901	0.865	0.913	0.836
提案手法	**0.655**	0.841	**0.842**	**0.904**	0.866	**0.918**	**0.838**

日本語	商品検索	類似度推定		含意関係認識		言い換え認識	平均
	Shopping Queries	JSTS	JSICK	JNLI	JSICK	PAWS-X	
追加学習なし	0.576	0.859	0.890	0.785	0.839	0.793	0.790
教師なし SimCSE	**0.587**	**0.861**	0.886	0.781	0.837	0.790	0.790
教師あり SimCSE	0.576	0.825	0.886	**0.843**	0.843	**0.800**	0.796
提案手法	**0.587**	**0.861**	**0.896**	0.828	**0.856**	0.791	**0.803**

する手法であり，著者らによって公開されている英語の学習済みモデル[21]を用いるため，英語の実験でのみ比較する．教師なし SimCSE は生コーパスのみを用いる Dropout ベースの対照学習であり，本実験では前節と同じ設定で Wikipedia を用いて再現する．教師あり SimCSE は NLI コーパスを用いる対照学習であり，英語では SNLI [2] および MNLI [14]，日本語では SNLI を日本語訳した JSNLI[22]を用いて，それぞれ再現する．

SimCSE のハイパーパラメータは，前節における提案手法の設定と同じである．ただし，JSNLI コーパスは 160k 文対に満たないため，追加学習の文数の最大値を 140k 文対とした．

3.4　実験結果

表 3 に実験結果を示す．提案手法によって，英語では全タスクにおいて追加学習なしベースライン以上の性能を達成し，日本語では言い換え認識以外の 3 タスクにおいて追加学習なしベースライン以上の性能を達成した．なお，日本語の言い換え認識タスクで性能が悪化した原因として，PAWS-X の言い換え文対は単語の並べ替えを中心に構築されているため，並べ替えを行わず語句の置換のみを行う本研究の言い換えとは異なる言語現象を扱っていることが影響している可能性がある．

Transfer Fine-Tuning [10] および SimCSE [11] の既存手法と比較しても，英日の両言語において，全タスクの平均性能において提案手法が最高性能を達成した．提案手法は，教師あり SimCSE における NLI コーパスのような高コストなアノテーションが不要であるため，マスク言語モデルの性能を低コストに改善できる強みを持つと言える．

4　おわりに

本研究では，マスク言語モデルによる文対モデリングの性能改善のために，追加学習としての言い換えベースの対照学習を提案した．提案手法は，自動生成された言い換えデータに基づいて高い性能を達成できるため，既存の対照学習が頼ってきた NLI コーパスのような高コストな人手アノテーションを必要としないという強みを持つ．商品検索・類似度推定・含意関係認識・言い換え認識の 4 種類の文対モデリングタスクを対象とする評価実験の結果，提案手法は英語と日本語の両言語において，マスク言語モデルの性能を既存手法以上に改善できた．

今後の課題として，対照学習における正例と負例の作成方法の改善が挙げられる．特に，正例の作成について機械翻訳 [16, 17] や大規模言語モデル [18] に基づく言い換え生成に取り組みたい．

謝辞

本研究は，株式会社メルカリ R4D の支援を受けて実施した．

21）https://github.com/yukiar/TransferFT

22）https://nlp.ist.i.kyoto-u.ac.jp/?%E6%97%A5%E6%9C%AC%E8%AA%9ESNLI%28JSNLI%29%E3%83%87%E3%83%BC%E3%82%BF%E3%82%BB%E3%83%83%E3%83%88

参考文献

[1] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In **SemEval**, pp. 1–14, 2017.

[2] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A Large Annotated Corpus for Learning Natural Language Inference. In **EMNLP**, pp. 632–642, 2015.

[3] Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar, and Amran Bhuiyan. Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges. **ACM Computing Surveys**, Vol. 56, No. 7, pp. 1–33, 2024.

[4] Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. A Survey for Efficient Open Domain Question Answering. In **ACL**, pp. 14447–14465, 2023.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. **arXiv:1301.3781**, 2013.

[6] Hua He and Jimmy Lin. Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement. In **NAACL**, pp. 937–948, 2016.

[7] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for Natural Language Inference. In **ACL**, pp. 1657–1668, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **NAACL**, pp. 4171–4186, 2019.

[9] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the Sentence Embeddings from Pre-trained Language Models. In **EMNLP**, pp. 9119–9130, 2020.

[10] Yuki Arase and Jun'ichi Tsujii. Transfer Fine-Tuning: A BERT Case Study. In **EMNLP**, pp. 5393–5404, 2019.

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In **EMNLP**, pp. 6894–6910, 2021.

[12] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings. In **NAACL**, pp. 4207–4218, 2022.

[13] Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. RankCSE: Unsupervised Sentence Representations Learning via Learning to Rank. In **ACL**, pp. 13785–13802, 2023.

[14] Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In **NAACL**, pp. 1112–1122, 2018.

[15] Juri Ganitkevitch and Chris Callison-Burch. The Multilingual Paraphrase Database. In **LREC**, pp. 4276–4283, 2014.

[16] J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. PARABANK: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation. In **AAAI**, pp. 6521–6528, 2019.

[17] Tomoyuki Kajiwara, Biwa Miura, and Yuki Arase. Monolingual Transfer Learning via Bilingual Translators for Style-sensitive Paraphrase Generation. In **AAAI**, pp. 8042–8049, 2020.

[18] Sam Witteveen and Martin Andrews. Paraphrasing with Large Language Models. In **NGT**, pp. 215–220, 2019.

[19] Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. **arXiv:2206.06588**, 2022.

[20] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK Cure for the Evaluation of Compositional Distributional Semantic Models. In **LREC**, pp. 216–223, 2014.

[21] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese General Language Understanding Evaluation. In **LREC**, pp. 2957–2966, 2022.

[22] Hitomi Yanaka and Koji Mineshima. Compositional Evaluation on Japanese Textual Entailment and Similarity. **TACL**, Vol. 10, pp. 1266–1284, 2022.

[23] Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase Adversaries from Word Scrambling. In **NAACL**, pp. 1298–1308, 2019.

[24] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In **EMNLP**, pp. 3687–3692, 2019.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **arXiv:1907.11692**, 2019.

[26] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In **ACL**, pp. 177–180, 2007.

[27] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In **EMNLP**, pp. 230–237, 2004.

[28] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB 2.0: Better Paraphrase Ranking, Fine-grained Entailment Relations, Word Embeddings, and Style Classification. In **ACL-IJCNLP**, pp. 425–430, 2015.

[29] 近藤里咲, 梶原智之, 二宮崇. JParaCrawl からの大規模日本語言い換え辞書の構築. 言語処理学会第 30 回年次大会, pp. 1736–1740, 2024.

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models Are Unsupervised Multitask Learners. 2019.