摂動を加えた kNN 機械翻訳による多様な翻訳候補の生成

西田 悠人 [1]　森下 睦 [2]　上垣外 英剛 [1]　渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学,　[2]NTT コミュニケーション科学基礎研究所
nishida.yuto.nu8@is.naist.jp

概要

ニューラル機械翻訳システムの標準的な探索アルゴリズムであるビームサーチには，出力される複数の翻訳候補がほとんど同一になってしまう多様性低下の問題が存在する．本研究では，翻訳候補の生成に用例データからの近傍探索を用いることで通常は翻訳候補に入らないようなトークンを考慮できる kNN 機械翻訳を応用し，近傍探索時に摂動を与えて探索範囲を確率的に拡大する手法を提案する．実験の結果，提案法により，翻訳精度を維持しつつ翻訳候補の多様性を改善できること，摂動の大きさを調整することで多様性を制御できることを報告する．

1　はじめに

機械翻訳システムが多様かつ妥当な翻訳候補を出力することは，リランキング [1] や人手による後編集 [2] といった後段の処理によって翻訳精度の向上が見込める点で重要である．しかし，ニューラル機械翻訳 (NMT) システムの標準的な探索アルゴリズムであるビームサーチには，出力される複数の候補文がほとんど同一になってしまう多様性低下の問題が存在することが知られている [2, 3, 4]．これに対し，Vijayakumar ら [3] や Freitag ら [4] は翻訳候補の多様化を促進するビームサーチの変種 (多様化デコード手法) を提案している．

他方で，NMT システムの標準的な損失関数であるクロスエントロピー損失を用いて訓練したモデルには，過剰修正 (overcorrection [5]) の問題が存在することが知られている．過剰修正とは，訓練データの分布とは異なる分布の予測に対し，その予測が妥当であっても出力確率を過剰に低く見積もってしまう現象のことである．そのため，前述の多様化デコード手法を標準的な NMT システムが出力する確率分布に適用するだけでは，妥当であるが出力確率が低く翻訳候補に入りづらいトークンが存在し，多様性向上の効果が限定的になっている可能性がある．

したがって，多様化デコード手法の効果を最大化するためには過剰修正に対処する必要があると考えられる．本研究では，その手段として用例ベースの機械翻訳手法の一種である kNN 機械翻訳 [6] に着目する．kNN 機械翻訳は，デコード時に用例データからの k 近傍探索によって多くの妥当なターゲットトークンを検索・抽出できるため，過剰修正の問題に対処できることが Yang ら [7] によって示唆されている．しかし，kNN 機械翻訳の k 近傍探索では探索の対象が制限されており，多様性を損ねてしまう．そこで，我々は kNN 機械翻訳を応用し，kNN 機械翻訳における k 近傍探索時に摂動を加えることでトークンを探索する範囲を確率的に拡大する手法を提案する (図 1)．この手法と多様化デコード手法を併用することで，翻訳性能を維持しつつ翻訳候補の多様化が期待できる．

本稿では，ドメイン適応と一般ドメインで評価を行い，両設定において，提案法によって翻訳性能を維持しながら多様性が向上することを示した．また，本手法において，摂動の大きさを調整することで多様性が制御できることも判明した．

2　関連研究

2.1　多様な翻訳候補の生成

ビームサーチによる翻訳候補の多様性低下の問題に対処するため，出力される翻訳候補の多様化を促進するビームサーチの変種が提案されている．Vijayakumar ら [3] はビーム間でのトークンの重複にペナルティを設けたビームサーチである Diverse Beam Search (DBS) を提案した．また，Freitag ら [4] は同一の部分仮説を共有する候補の最大数を決める手法を提案した．これらの手法はいずれも翻訳候補の多様性向上を報告しているが，過剰修正に明示的には対処しておらず，妥当であるが出力確率が低く出力候補に入りづらいトークンが存在し，多様性向上の効果が限定的になっている可能性がある．

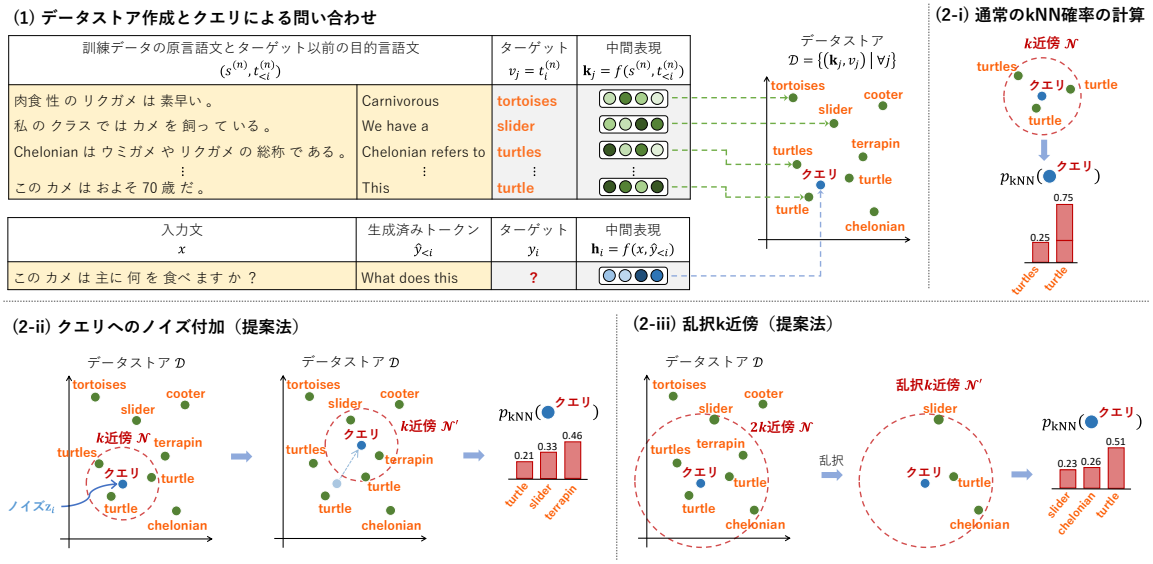

図 1　提案手法の概略図．得られた p_{kNN} によってデコードを行う段階については先行研究と同様であるため省略した．

2.2　kNN 機械翻訳

kNN 機械翻訳は以下に示す 2 ステップの方法で k 近傍探索のメカニズムを NMT モデルの推論に適用する手法であり，大規模な事例データに直接アクセスすることでよりよい推論を実現できる．

データストア構築　NMT モデルに訓練データを入力し，訓練データ中の目的言語側のトークンをキー，対応する隠れ層の表現をバリューとしてデータストアに保持する．訓練データ $(\mathcal{S}, \mathcal{T})$ のソース文 $s \in \mathcal{S}$，ターゲット文 $t \in \mathcal{T}$ に対する時刻 i の隠れ状態ベクトルを $f(s, t_{<i})$ とすると，データストア $\mathcal{D}$ は式 (1) で表される．

$$\mathcal{D} = \{(f(s, t_{<i}),\ t_i),\ \forall t_i \in t \mid (s, t) \in (\mathcal{S}, \mathcal{T})\} \quad (1)$$

デコーディング　入力文 x に対する時刻 i の隠れ状態ベクトル $\mathbf{h}_i$ を出力トークン y_i に対応するクエリとして，データストアから k 近傍 $\mathcal{N} \subset \mathcal{D}$ を抽出する．クエリ $\mathbf{h}_i$ と k 近傍点の距離から k 近傍確率 p_{kNN} を式 (2) のように計算する．なお，$d(\cdot)$ は距離関数，T は softmax 関数の温度パラメータである．

$$p_{\mathrm{kNN}}(y_i|x, y_{<i}) \propto \sum_{(\mathbf{k}_j, v_j) \in \mathcal{N}} \mathbb{1}_{y_i = v_j} \exp\left(\frac{-d(\mathbf{k}_j, \mathbf{h}_i)}{T}\right) \quad (2)$$

最後に，y_i の出力確率を k 近傍確率 p_{kNN} と NMT モデルの出力確率 p_{MT} の線形補間により式 (3) のように計算する．ここで，λ は k 近傍確率の重みを決定するハイパーパラメータである．

$$p(y_i|x, y_{<i}) = \lambda p_{\mathrm{kNN}}(y_i|x, y_{<i}) + (1 - \lambda) p_{\mathrm{MT}}(y_i|x, y_{<i}) \quad (3)$$

kNN 機械翻訳はモデルを追加で学習することなく翻訳精度を改善できることが報告されており，Yang ら [7] は k 近傍探索による過剰修正への対処が翻訳精度向上の理由であると示唆している．また，kNN 機械翻訳の変種として，翻訳性能を更に改善する手法 [8] や推論の遅さを克服する手法 [9, 10] が提案されている．その一方で，kNN 機械翻訳を多様性向上のために活用する研究は行われていない．

3　提案法：kNN-Diversify Decoding

本研究では，kNN 機械翻訳と多様化デコード手法の組み合わせ，および近傍探索時の探索範囲を確率的に拡げる手法を提案する．手法の概略図を図 1 に示す．この手法では，k 近傍探索によって，通常では出力確率の上位には含まれないが妥当なトークンの出力確率の向上が期待できる．さらに，その確率分布を多様化デコード手法を用いて幅広く探索することで妥当かつ多様な翻訳候補の生成を図る．また，従来の kNN 機械翻訳では k 近傍探索の範囲が制限されているという問題を解決し，さらに多様性を向上させるために，k 近傍探索時に摂動を加えて探索の範囲を確率的に拡げる手法を 2 種類提案する．

3.1　クエリへのノイズ付加

k 近傍分布に摂動を与える最も単純な方法として k 近傍探索のクエリにノイズベクトルを加える手法 (図 1 の (2-ii)) を導入する．この手法では，出力トークン y_i に対する中間表現 $\mathbf{h}_i$ にノイズベクトル $\mathbf{z}_i$ を加えた $\mathbf{h}_i + \mathbf{z}_i$ をクエリとして k 近傍探索を行い k 近

傍 $\mathcal{N}'$ を得る．その後，式 2 のように，得た $\mathcal{N}'$ から k 近傍確率を計算する．ノイズベクトル $\mathbf{z}_i$ はノルム 1 のホワイトガウスノイズ $\mathbf{n}_i$ と平均 m，分散 s^2 の正規分布に従う確率変数 a_i の積，すなわち $\mathbf{z}_i = a_i \times \mathbf{n}_i$ として各時刻・各ビームで独立に生成する．パラメータ m, s は以下の 2 つの方法を用いて決定する．

静的ノイズ　最も単純な方法として，ハイパーパラメータ h_m, h_s を用いて $m = h_m, s = h_s$ とする静的ノイズを導入する．この方法では，追加の推論コストをほぼ必要とせずに k 近傍探索の範囲を確率的に拡げられ，翻訳候補の多様化が期待できる．しかし，ノイズの大きさを適切に決定するためには，データストアの分布の事前調査が必要である[1)]．

適応的ノイズ　データストアの分布の事前調査を必要としない方法として，各時刻で予め k 近傍探索を行って得た分布を基にノイズの大きさを決定する適応的ノイズを導入する．具体的には，通常の k 近傍探索を行い，k 近傍点までの距離の最大値 $d_{\max}$ と k 近傍点までの距離の標準偏差 d_{std} を得る．ハイパーパラメータ h_m, h_s を用いて $m = h_m \times d_{\max}, s = h_s \times d_{\mathrm{std}}$ とする．この方法では，各時刻における k 近傍分布に基づいてノイズの大きさを決定できるため，データストアの分布の事前調査は不要である．しかし，各時刻で 2 回 k 近傍探索を行うため，追加の推論コストが必要である．

3.2　乱択 k 近傍

前節の手法の欠点であったデータストアの分布の事前調査または追加の推論コストを必要とせずに k 近傍探索の範囲を確率的に拡げる手法として，近傍点の一部をサンプリングする乱択 k 近傍 (図 1 の (2-iii)) を導入する．具体的には，$h > 1$ を満たすハイパーパラメータ h を用いて $\lfloor h \times k \rfloor$ 近傍を抽出し，$\lfloor h \times k \rfloor$ 個の近傍点から k 個を一様ランダムにサンプリングすることで乱択 k 近傍 $\mathcal{N}'$ を得る．この手法では，より多くの近傍点を探索範囲に含めることができるため，翻訳候補の多様化が期待できる．また，データストアの分布を事前に調査する必要はなく，各時刻ごとに 1 回のみ k 近傍探索を行うため，追加の推論コストも必要ない．

1)　本研究では，事前に開発データで kNN 機械翻訳を行い，その際に k 近傍点までの距離の平均・分散を取得した．
2)　他に Neucleus Sampling [11] の実験も行ったが，DBS と概ね同様の結果であったため本稿では割愛する．
3)　N が偶数のときは中央順位の 2 文のうち文レベルの BLEU が高い文を選ぶ．
4)　https://huggingface.co/bert-base-multilingual-cased

4　実験設定

4.1　データ・モデル・ハイパーパラメータ

実験は大別してドメイン適応の設定と一般ドメインの設定の 2 つを行う．ドメイン適応の実験では独英の Koran, IT, Medical, Law, Subtitles の 5 つのドメインデータ [12, 13] を用いる．一般ドメインの実験では WMT'19 独英翻訳のデータを用いる．両方の実験において，データストアの構築およびテストデータの翻訳には，`fairseq` ライブラリ [14] で利用可能な WMT'19 独英訓練済みモデル [15] を用いた．ビーム幅は 20 とし，提案法のハイパーパラメータは開発データで最も良いパラメータを選んだ．kNN 機械翻訳の詳細な設定については付録 A に記す．多様化デコード手法には DBS を用いた[2)]．DBS のグループ数は 20 とし，多様性強度は 0.5 とした．

4.2　評価指標

翻訳性能　翻訳性能は，オラクル BLEU，Median BLEU，平均文長比によって評価する．オラクル BLEU（BLEU@N）は N-best 候補文で文レベルの BLEU [16] が最大の文をそれぞれ選定したときの BLEU であり，リランキングによる性能の上界に相当する．実験では BLEU@1 および BLEU@20 を報告する．なお，BLEU@1 (1-best 訳に対するオラクル BLEU) は通常の BLEU である．Median BLEU (MedBLEU) は N-best 候補文で文レベルの BLEU が中央値である文[3)]をそれぞれ選定したときの BLEU であり，全体の平均的な翻訳品質を示す．平均文長比 (AveLen) は参照訳の文長に対する翻訳候補の文長の比の平均である．

多様性　翻訳候補の多様性の指標には BLEU-based discrepancy metric (DP) [17] を用いる．DP は，出力された候補文に含まれるユニークな n-gram の多さを示す指標であり，DP が高いと多様性が高いことを示す．指標の詳細は付録 B に記す．

流暢性　翻訳候補の流暢性の指標には pseudo-log-likelihood score (PLL) [18] を用いる．実験では，翻訳候補全体の PLL の平均 (AvePLL) および，各 N-best 候補文の PLL の最大値/最小値の平均（MaxPLL/MinPLL）を報告する．また，比較のために参照訳の AvePLL も報告する．指標の詳細は付録 B に記す．PLL を計算するための MLM モデルには多言語 BERT[4)] [19] を使用した．

表 1　ドメイン適応の評価結果．値は平均 ± 母標準偏差の形式で記載した．

Method	DP	BLEU@1	BLEU@20	MedBLEU	AveLen	MinPLL	MaxPLL	AvePLL
Reference	-	-	-	-	-	-	-	$-3.35_{\pm 0.82}$
Baseline	$31.4_{\pm 10.5}$	$34.1_{\pm 10.2}$	$42.6_{\pm 10.6}$	$30.8_{\pm 10.2}$	$\mathbf{0.956}_{\pm 0.028}$	$-2.26_{\pm 0.37}$	$\mathbf{-4.55}_{\pm 1.43}$	$\mathbf{-3.28}_{\pm 0.83}$
DBS	$35.9_{\pm 6.4}$	$33.6_{\pm 9.8}$	$40.0_{\pm 9.8}$	$31.4_{\pm 9.5}$	$0.946_{\pm 0.026}$	$-2.23_{\pm 0.38}$	$-4.63_{\pm 1.30}$	$\mathbf{-3.28}_{\pm 0.80}$
kNN	$32.3_{\pm 11.5}$	$\mathbf{43.2}_{\pm 15.0}$	$\mathbf{51.8}_{\pm 14.6}$	$38.3_{\pm 14.7}$	$0.947_{\pm 0.026}$	$-2.23_{\pm 0.38}$	$-4.74_{\pm 1.58}$	$-3.32_{\pm 0.90}$
kNN+DBS	$42.0_{\pm 9.5}$	$42.0_{\pm 14.8}$	$48.6_{\pm 13.9}$	$\mathbf{38.8}_{\pm 14.6}$	$0.950_{\pm 0.017}$	$-2.18_{\pm 0.51}$	$-4.90_{\pm 1.45}$	$-3.35_{\pm 0.88}$
適応的ノイズ	$53.7_{\pm 12.2}$	$41.0_{\pm 15.1}$	$49.0_{\pm 14.2}$	$36.2_{\pm 14.9}$	$0.951_{\pm 0.017}$	$-2.04_{\pm 0.49}$	$-5.21_{\pm 1.56}$	$-3.38_{\pm 0.90}$
静的ノイズ	$\mathbf{55.2}_{\pm 11.7}$	$40.4_{\pm 14.6}$	$49.0_{\pm 14.0}$	$35.4_{\pm 14.5}$	$0.949_{\pm 0.016}$	$\mathbf{-2.02}_{\pm 0.51}$	$-5.23_{\pm 1.56}$	$-3.37_{\pm 0.89}$
乱択 k 近傍	$54.4_{\pm 7.3}$	$39.5_{\pm 13.3}$	$48.4_{\pm 13.6}$	$34.3_{\pm 12.6}$	$0.950_{\pm 0.017}$	$-2.08_{\pm 0.51}$	$-5.16_{\pm 1.51}$	$-3.38_{\pm 0.90}$

表 2　一般ドメイン (WMT'19 独英翻訳) の評価結果

Method	DP	BLEU@1	BLEU@20	MedBLEU	AveLen	MinPLL	MaxPLL	AvePLL
Reference	-	-	-	-	-	-	-	-2.93
Baseline	32.1	39.6	51.4	36.8	0.981	-2.18	-3.53	-2.80
DBS	41.3	38.9	47.6	34.9	0.969	-2.09	-3.77	-2.84
kNN	31.7	**40.6**	**52.2**	**37.5**	**0.986**	-2.17	**-3.51**	**-2.79**
kNN+DBS	41.7	39.7	48.7	35.5	0.976	-2.03	-3.77	-2.82
適応的ノイズ	41.9	39.6	48.7	35.3	0.977	-2.02	-3.78	-2.82
静的ノイズ	**42.7**	39.7	48.6	35.4	0.979	-2.02	-3.78	-2.82
乱択 k 近傍	42.2	39.6	48.6	35.2	0.981	**-1.99**	-3.77	-2.81

5　実験結果

5.1　ドメイン適応

ドメイン間の平均をとった結果を表 1 に示す．各ドメインの結果は付録の表 5 に示す．提案法の kNN+DBS は DBS に対して DP が向上した．kNN に比べて BLEU@20 が低下したが，その減少幅は Baseline と DBS の差と同程度である．また，摂動を加えることで kNN+DBS と比較して DP および BLEU@20 が改善した．なお，摂動の種類によって BLEU および DP に大きな差はなかった．kNN は Baseline と比較して若干 PLL が低いが，提案法では kNN と比較して PLL はほぼ毀損されず，参照訳の PLL と比べてもその差は小さい．以上より，ドメイン適応の設定で提案法は翻訳精度と流暢性を低下させることなく翻訳候補の多様性を向上させられることが示された．

5.2　一般ドメイン

評価結果を表 2 に示す．提案法である kNN+DBS を DBS と比較すると，DP と各種 BLEU が向上しており，摂動を加えることで BLEU を維持したままさらに DP が向上している．また，提案法の PLL は参照訳の PLL と比較して高く，Baseline と比較しても低下していない．よって，一般ドメインにおいても提案法によって翻訳精度と流暢性を維持しながら多様化の促進がみられた．

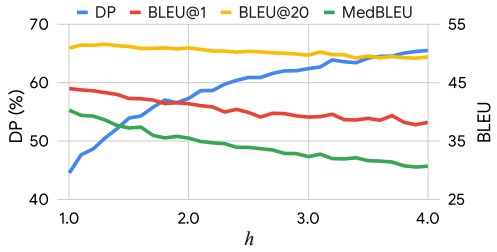

図 2　IT ドメインの乱択 k 近傍における，摂動の大きさ h(横軸) と翻訳精度や多様性 (縦軸) への影響．$h = 1$ のスコアは摂動無し，すなわち DBS+kNN の値である．

5.3　翻訳性能と多様性のトレードオフ

摂動の大きさによる翻訳精度や多様性への影響を観察するため，IT ドメインの乱択 k 近傍における摂動の大きさと DP および BLEU の関係を図 2 に示す．図より，DP と各種 BLEU はトレードオフ関係にあり，提案法は摂動の大きさを変化させることで多様性と翻訳精度を調整可能であることが示された．

6　おわりに

本研究では，kNN 機械翻訳の k 近傍探索時に摂動を加えることで探索範囲を確率的に拡げる手法を提案し，この手法によって翻訳精度と流暢性を維持しつつ翻訳候補の多様化ができることを示した．また，提案法において摂動の大きさを変化させることで多様性と翻訳精度を調整できることを報告した．今後は，kNN 機械翻訳の推論速度を向上させる手法の適用や，より多様性を向上させられる手法の検討を行いたい．

謝辞

本研究は JSPS 科研費 JP21H05054，JP21K17801 の助成を受けたものです．

参考文献

[1] Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine translation. **arXiv preprint arXiv:1601.00372**, 2016.

[2] Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. A systematic exploration of diversity in machine translation. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1100–1111, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[3] Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 32, No. 1, Apr. 2018.

[4] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In **Proceedings of the First Workshop on Neural Machine Translation**, pp. 56–60, Vancouver, August 2017. Association for Computational Linguistics.

[5] Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. Bridging the gap between training and inference for neural machine translation. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4334–4343, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In **International Conference on Learning Representations**, 2021.

[7] Zhixian Yang, Renliang Sun, and Xiaojun Wan. Nearest neighbor knowledge distillation for neural machine translation. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5546–5556, Seattle, United States, July 2022. Association for Computational Linguistics.

[8] Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. Adaptive nearest neighbor machine translation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 368–374, Online, August 2021. Association for Computational Linguistics.

[9] Dexin Wang, Kai Fan, Boxing Chen, and Deyi Xiong. Efficient cluster-based k-nearest-neighbor machine translation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2175–2187, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li. Fast nearest neighbor machine translation. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 555–565, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[11] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In **International Conference on Learning Representations**, 2020.

[12] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In **Proceedings of the First Workshop on Neural Machine Translation**, pp. 28–39, Vancouver, August 2017. Association for Computational Linguistics.

[13] Roee Aharoni and Yoav Goldberg. Unsupervised domain clusters in pretrained language models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7747–7763, Online, July 2020. Association for Computational Linguistics.

[14] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook FAIR's WMT19 news translation task submission. In **Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)**, pp. 314–319, Florence, Italy, August 2019. Association for Computational Linguistics.

[16] Boxing Chen and Colin Cherry. A systematic comparison of smoothing techniques for sentence-level BLEU. In **Proceedings of the Ninth Workshop on Statistical Machine Translation**, pp. 362–367, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[17] Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. Generating diverse translations with sentence codes. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 1823–1827, Florence, Italy, July 2019. Association for Computational Linguistics.

[18] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2699–2712, Online, July 2020. Association for Computational Linguistics.

[19] 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.

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. **IEEE Transactions on Big Data**, Vol. 7, No. 3, pp. 535–547, 2019.

A　実験設定の詳細

データストアの構築および近傍探索には FAISS [20] を用いた．近傍数 k および k 近傍確率の温度 T，補間係数 λ は，ドメイン適応では Khandelwal ら [6] のパラメータを用い，一般ドメインでは開発データで $k = \{16, 32, 64, 128\}$, $T = \{10, 100, 1000\}$, $\lambda = \{0.1, 0.2, \ldots, 0.9\}$ でグリッドサーチを行い，最良のパラメータを選んだ．各手法で使用したハイパーパラメータを表 3 に示す．

表 3　ハイパーパラメータ

	Koran	IT	Medical	Law	Subtitles	WMT
kNN 機械翻訳						
k	64	64	64	64	64	32
λ	0.8	0.7	0.8	0.8	0.7	0.2
T	100	10	10	10	10	100
静的ノイズ						
h_m	49.4	23.8	38.2	16.9	36.0	18.9
h_s	1.15	3.20	0.35	2.00	0.20	1.40
適応的ノイズ						
h_m	0.2	0.1	0.1	0.05	0.1	0.025
h_s	0.1	0.05	0.1	0.05	0.05	0.8
乱択 k 近傍						
h	2.9	2.0	2.7	3.2	3.1	3.7

B　評価指標の詳細

ソース文集合を $\mathfrak{X} = \{x_1, \ldots, x_M\}$，ソース文 x_k に対する N-best 候補文集合を $\mathbf{B}_k = \{\hat{y}_k^1, \ldots, \hat{y}_k^N\}$ とする．

多様性の指標 DP はソース文集合 $\mathfrak{X}$ に対する N-best 候補文集合を $\mathbb{H} = \{\mathbf{H}_1, \ldots, \mathbf{H}_N\}; \mathbf{H}_n = \{\hat{y}_1^n, \ldots, \hat{y}_M^n\}$ としたとき，式 (4) のように計算される．なお，BLEU$(\mathbf{H}, \mathbf{H}')$ は参照 $\mathbf{H}'$ に対する仮説 $\mathbf{H}$ のコーパス単位の BLEU である．

$$\mathrm{DP}(\mathbb{H}) = \frac{1}{N(N-1)} \sum_{\mathbf{H} \in \mathbb{H}} \sum_{\mathbf{H}' \in \mathbb{H}, \mathbf{H}' \neq \mathbf{H}} 1 - \mathrm{BLEU}(\mathbf{H}, \mathbf{H}') \tag{4}$$

流暢性の指標 PLL は文 $y = (w_1, \ldots, w_{|y|})$ に対して式 (5) のように計算される．ここで，式中の $y_{\setminus t}$ は時刻 t のトークン w_t がマスクされた文であり，$P_{\mathrm{MLM}}(w_t | y_{\setminus t})$ は MLM モデルがマスクされた文 $y_{\setminus t}$ から元のトークン w_t を予測する確率である．また，MaxPLL，MinPLL，AvePLL はシステム出力を $\mathbb{W} = \{\mathbf{B}_1, \ldots, \mathbf{B}_M\}$ とするとき，それぞれ式 (6)，式 (7)，式 (8) のように定義される．

$$\mathrm{PLL}(y) = \sum_{t=1}^{|y|} \log P_{\mathrm{MLM}}(w_t | y_{\setminus t}) \tag{5}$$

$$\mathrm{MaxPLL}(\mathbb{W}) = \frac{1}{|\mathbb{W}|} \sum_{\mathbf{B} \in \mathbb{W}} \max_{\hat{y} \in \mathbf{B}} \left(\frac{1}{|\hat{y}|} \mathrm{PLL}(\hat{y}) \right) \tag{6}$$

$$\mathrm{MinPLL}(\mathbb{W}) = \frac{1}{|\mathbb{W}|} \sum_{\mathbf{B} \in \mathbb{W}} \min_{\hat{y} \in \mathbf{B}} \left(\frac{1}{|\hat{y}|} \mathrm{PLL}(\hat{y}) \right) \tag{7}$$

$$\mathrm{AvePLL}(\mathbb{W}) = \frac{1}{|\mathbb{W}|} \sum_{\mathbf{B} \in \mathbb{W}} \left(\frac{1}{|\mathbf{B}|} \sum_{\hat{y} \in \mathbf{B}} \left(\frac{1}{|\hat{y}|} \mathrm{PLL}(\hat{y}) \right) \right) \tag{8}$$

C　推論コスト

ドメイン適応実験における各ドメインおよび一般ドメイン (WMT) の推論の速さ (tokens/s) を表 4 に示す．表中の報告値は `fairseq` の出力ログに記載されている値である．表より，提案法のうち DBS+kNN，静的ノイズ，乱択 k 近傍では推論の速さが kNN と比較して若干低下する傾向があり，適応的ノイズでは推論速度の低下が顕著であることがわかる．

表 4　推論の速さ．単位は tokens/s．行ラベルは (1) Baseline，(2) DBS，(3) kNN，(4) DBS+kNN，(5) 適応的ノイズ，(6) 静的ノイズ，(7) 乱択 k 近傍である．

	Koran	IT	Medical	Law	Subtitles	WMT
(1)	897.2	874.8	706.0	822.4	1095.9	969.6
(2)	565.2	597.7	401.2	428.1	857.7	717.9
(3)	86.8	59.5	17.9	27.2	4.5	8.6
(4)	75.8	57.3	15.7	26.1	4.4	9.0
(5)	44.4	32.9	8.0	13.7	2.2	4.6
(6)	67.2	50.6	14.0	25.8	4.4	9.1
(7)	65.2	51.0	14.6	25.5	4.4	8.9

D　各ドメインの結果

ドメイン適応設定の評価結果を表 5 に示す．

表 5　各ドメインの評価結果．行ラベルの (1)~(7) は表 4 と同一，列ラベルは① DP，② BLEU@1，③ BLEU@20，④ MedBLEU，⑤ MaxPLL，⑥ MinPLL，⑦ MaxPLL である．

	①	②	③	④	⑤	⑥	⑦	⑧
Koran								
ref	-	-	-	-	-	-	-	-2.95
(1)	27.4	16.9	22.5	16.2	0.950	-2.08	-3.48	-2.73
(2)	39.4	17.0	22.4	15.9	0.937	-1.97	-3.74	-2.77
(3)	26.2	21.0	27.4	20.0	0.946	-2.01	-3.48	-2.69
(4)	47.2	20.5	27.0	18.7	0.945	-1.72	-3.84	-2.73
(5)	63.6	18.6	26.6	16.3	0.948	-1.55	-4.14	-2.76
(6)	60.1	19.3	27.1	17.1	0.941	-1.54	-3.92	-2.69
(7)	55.9	19.5	26.7	17.1	0.953	-1.64	-3.86	-2.69
IT								
ref	-	-	-	-	-	-	-	-4.93
(1)	31.5	37.7	47.3	33.9	1.004	-2.96	-7.21	-4.87
(2)	35.2	37.1	44.2	34.9	0.998	-2.97	-7.05	-4.82
(3)	32.7	45.9	55.0	39.7	0.974	-2.94	-7.69	-5.04
(4)	44.6	43.9	50.9	40.2	0.975	-3.09	-7.60	-5.05
(5)	59.5	42.5	51.2	36.2	0.975	-2.89	-8.08	-5.10
(6)	57.0	42.8	51.6	37.0	0.973	-2.92	-7.97	-5.07
(7)	57.1	41.5	50.8	35.7	0.972	-2.97	-7.92	-5.10
Medical								
ref	-	-	-	-	-	-	-	-3.24
(1)	27.7	40.4	49.1	37.3	0.946	-2.33	-4.23	-3.17
(2)	31.4	39.9	46.0	37.7	0.938	-2.26	-4.29	-3.15
(3)	29.5	55.4	63.0	49.2	0.928	-2.31	-4.59	-3.28
(4)	36.7	54.0	59.6	50.6	0.937	-2.31	-4.77	-3.31
(5)	49.2	52.9	60.3	47.3	0.934	-2.17	-5.20	-3.38
(6)	55.8	50.6	59.3	44.3	0.934	-2.11	-5.50	-3.41
(7)	52.1	50.0	59.0	44.0	0.931	-2.21	-5.26	-3.40
Law								
ref	-	-	-	-	-	-	-	-2.57
(1)	19.5	46.1	52.4	44.2	0.963	-1.97	-3.17	-2.50
(2)	27.4	45.0	50.2	42.6	0.936	-2.00	-3.35	-2.58
(3)	19.6	61.9	68.8	58.9	0.977	-2.02	-3.27	-2.57
(4)	26.9	60.8	65.8	58.0	0.965	-2.06	-3.44	-2.64
(5)	31.7	60.6	66.5	57.2	0.964	-2.02	-3.53	-2.64
(6)	33.8	60.0	66.6	56.6	0.964	-2.01	-3.55	-2.65
(7)	42.2	56.2	64.8	51.1	0.963	-1.99	-3.71	-2.69
Subtitles								
ref	-	-	-	-	-	-	-	-3.07
(1)	51.0	29.3	41.7	22.3	0.917	-1.98	-4.65	-3.12
(2)	45.9	29.0	36.9	25.7	0.923	-1.96	-4.74	-3.10
(3)	53.6	31.7	45.1	23.9	0.911	-1.87	-4.65	-3.04
(4)	54.6	30.6	39.5	26.3	0.930	-1.70	-4.86	-3.02
(5)	64.4	30.4	40.5	24.0	0.932	-1.58	-5.09	-3.03
(6)	69.5	29.5	40.4	21.8	0.933	-1.51	-5.19	-3.03
(7)	64.7	30.2	40.6	23.4	0.931	-1.56	-5.03	-3.02