金融市場翻訳に特化した LLM に対する LLM-as-a-Judge ベース蒸留学習

井本稔也 [1] 伊藤敬彦 [1] 川原一修 [1]
[1]*JapanDigitalDesign* 株式会社
{toshiya.imoto, takahiko.ito, takanobu.kawahara}@japan-d2.com

概要

金融市場レポートは, 為替・金利・株式など多様な資産クラスの動向や見通しを高頻度で報告する実務文書であり, 専門用語・略語や市場参加者特有の言い回し, 機関投資家向けの文体といったドメイン固有の特徴を持つ. したがって, 単なる文単位の意味対応だけでなく, レポート全体で一貫した用語選択と自然な市場コメントとしての表現が求められる. 近年の機械翻訳や大規模言語モデル（LLM）の性能向上にもかかわらず, 金融市場レポートでは用語の不自然さや不統一が残り, 人手による修正が依然として不可欠である. 本研究の貢献・新規性は以下の通りである.

- 高頻度更新かつ厳格な文体制約を持つ金融市場レポートを対象とした点.
- LLM-as-a-Judge と LLM ベース・ポストエディットを組み合わせ，高品質な合成対訳を自動構築し，金融特化 LLM を蒸留した点.
- 翻訳品質の平均値だけでなく分散（安定性）にも着目し，Judge ベース評価と統計的検定で効果を分析した点.

1 序論

1.1 背景：金融市場翻訳の課題

金融市場レポートは, 為替, 金利, 株式, コモディティなど多様な資産クラスについて, 市場動向やその要因, 今後の見通しを日次・週次・月次で解説する実務文書である. これらのレポートは, マクロ経済指標の公表, 金融政策決定会合, 地政学リスクの顕在化といったイベントに応じて高頻度で更新され, タイムリーな情報提供が求められる. このような金融市場レポートは,(i) ドメイン固有の専門用語・略語, (ii) 市場参加者の慣習に根ざした言い回しや比喩, (iii) 機関投資家やディーラーを主な読者と想定した独特の文体, といった特徴を持つ. そのため, 単に文単位で意味が対応しているだけでは不十分であり, レポート全体として一貫した用語選択と論理展開を保ちつつ, 市場コメントとして自然な表現を実現することが重要となる. 近年, 機械翻訳システムや大規模言語モデル（LLM）の性能向上により, 一般ニュースやウェブテキストに対しては高品質な自動翻訳が可能になりつつある. しかし, 金融市場レポートの翻訳に適用した場合, 金融ドメインの観点から見ると不自然・不統一な用語選択を行う訳文が生成されることがままあり, 人手による修正が不可欠である.

1.2 本研究の目的

本研究の狙いは, 対訳が少量にしか存在しない金融市場レポートを最大限に活用しつつ, 人手翻訳に近い自然さと一貫した用語選択を備えた翻訳モデルを構築することである. 具体的には,LLM-as-a-Judge を中核とする新しい翻訳データ蒸留パイプラインを提案する. パイプラインの詳細については 3 章に記載する.

2 関連研究

本節では，(i) ドメイン特化機械翻訳／ LLM のドメイン適応，(ii) 知識蒸留・データ蒸留に基づく翻訳性能向上，(iii) LLM-as-a-Judge による自動評価，の 3 つの観点から関連研究を整理し，本研究の位置づけを述べる.

2.1 ドメイン適応 MT / LLM

ニューラル機械翻訳（NMT）におけるドメイン適応は，追加対訳コーパスによる微調整，単言語データを用いた back-translation [1]，自然文書による継続事前学習（Continual Pre-Training; CPT）といった手法が広く利用されてきた.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

近年は，汎用 LLM を翻訳タスクに特化させる研究が注目を集めている．Zheng ら [2] は LLaMA 系モデルを IT ドメインの機械翻訳に適応させ，LoRA による効率的な微調整で専用 NMT に匹敵する性能を達成した．Moslem ら [3] は Mistral 7B を医療翻訳に適応させ，約 2 万セグメントという比較的少量の並列データでも高品質な翻訳を実現している．さらに Zhang ら [4] は，QLoRA 微調整が既存の専用 NMT を上回る翻訳性能を発揮することを示し，ドメイン適応における LLM の可能性をさらに拡張した．

しかし，これらの研究は十分な量の並列データの存在を前提とすることが多い．対訳が存在しない実務ドメイン（金融市場レポートなど）をどのように活用するかについては，依然として未解決の課題が残っている．

2.2 知識蒸留とデータ蒸留

知識蒸留（knowledge distillation）は，教師モデルの出力分布を生徒モデルが模倣する枠組みであり，翻訳モデルの圧縮や性能向上に広く用いられている．NMT における代表的な手法として，Sequence-level knowledge distillation [5] や継続学習の文脈での continual knowledge distillation [6] がある．

また，単言語データを活用して疑似対訳を作成する self-training 型のデータ蒸留アプローチも広く用いられている [1]．一方，従来研究の多くは「教師モデルが生成した単一出力」を用いる構成が主流であり，複数候補のリランキングや，多次元的な品質評価（流暢性・忠実性・スタイル・情報被覆）を踏まえたデータ選別は十分に研究されていない．

本研究は，LLM-as-a-Judge による多数の候補訳のスコアリングと LLM ベースのポストエディットを組み合わせることで，高品質な合成対訳のみを抽出する「データ蒸留」パイプラインを構築する点で，既存手法と異なる．

2.3 本研究の位置づけ

本研究は，以上の文脈を踏まえ以下の点で新規性を有する．

- 高頻度更新・厳格な文体制約を持つ金融市場レポートを対象とする点．
- LLM-as-a-Judge による多軸評価とポストエディットを組み合わせ，高品質な合成対訳を自動生成する「データ蒸留」パイプラインを構築する点．
- 翻訳性能の平均値だけでなく，文書間の分散（品質の安定性）向上に焦点を当てた点．

これらにより，本研究は「LLM-as-a-Judge を評価器にとどめず，金融市場翻訳向けの学習データ構築そのものに利用する蒸留手法」として位置づけられる．

3 提案手法

3.1 対象データとベースモデル

本研究では，金融実務で利用されている以下の金融市場レポート群を対象とした．

- 金融機関による為替・金利・マクロ経済に関する市場レポート
- 日本銀行（BOJ）による金融経済レポート

これらのうち一部には人手による日英対訳が存在し，残りは日本語のみ（あるいは英語のみ）のモノリンガルデータである．既存の対訳データは，従来どおりの Supervised Fine-Tuning（SFT） 用教師データとして用いる一方で，モノリンガルデータは継続事前学習 (CPT) での利用や，提案パイプラインを通じて合成対訳へと変換して活用する．

ベースモデルには Llama 3.3 Swallow 70B v0.4 をモノリンガルデータで CPT し，金融向け対訳データで SFT したモデルを用いた．以降，本モデルを **初期翻訳モデル** と呼ぶ．初期翻訳モデルは金融用語の再現性に優れる一方で，日本語や英語としての自然さや訳文のクオリティのばらつきに課題があった．本研究の目的は，初期翻訳モデルの強みを保持しつつ，この品質のばらつきを小さくすることである．

3.2 多様な初期翻訳候補の生成

前項のデータセットをセグメント単位に分割し，各セグメント x に対して初期翻訳モデルから複数の翻訳候補をサンプリング生成する．具体的には，温度付きサンプリングを用い，温度 T，top-p p を適切に設定した上で $n = 50$ 本の候補 $\{y^{(1)}, \ldots, y^{(n)}\}$ を生成する．このステップにより，初期翻訳モデルが潜在的に生成可能な表現空間から，「良い訳」候補と「悪い訳」候補を含む多様なサンプル群を収集する．

3.3 LLM-as-a-Judge によるスコアリングとリランキング

次に，Judge モデルとして別の大規模モデル（DeepSeek-V3.1）を用い，各候補訳 $y^{(k)}$ に対して以

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

下の 4 軸でスコアリングを行う.

Accuracy　意味の正確さに関するスコア
Terminology　市場慣行・専門用語の適切さ
Fluency　読みやすさ・説得力
Overall　全体のスコア

Judge へのプロンプトは，原文セグメント x と候補訳 $\{y^{(1)}, \ldots, y^{(n)}\}$ を入力とし，各軸について 100 点満点でスコアを付与するよう設計した.

3.4　LLM ベース・ポストエディットによる高品質訳の構築

LLM-as-a-Judge により選択された Top1 候補 $y^\star$ は，依然として細かな不自然さや金融ドメイン特有の表現揺れを含むことがある. そこで本研究では，再度 DeepSeek-V3.1 を用いた **ポストエディット** により，$y^\star$ を高品質な翻訳文 $\tilde{y}$ へと洗練させる.

原文 x とポストエディット後の訳文 $\tilde{y}$ のペア $(x, \tilde{y})$ を**合成対訳** として得ることができる.

3.5　蒸留モデルの Supervised Fine-Tuning

構築したデータセットを用いて，再び Llama 3.3 Swallow 70B v0.4 に対して SFT を行う. ここで学習されるモデルを **蒸留モデル** と呼ぶ.

これにより，蒸留モデルでは, 運用時の計算コストを抑えつつ，初期翻訳モデルに比べて平均品質が高いだけでなく品質分散が小さい，より安定した金融市場翻訳モデルを実現することを狙う.

4　実験設定

本章では初期モデルと, 提案パイプラインで作成した蒸留モデルを比較するための実験設定を述べる.

4.1　評価データセット

評価には，人手で作成された金融市場レポート対訳コーパスの一部から評価用のサブセットを学習データと重複しないように抽出した. 評価は，日英・英日の両方向について実施した.

モデルについては, 初期モデルと蒸留モデルの二つを比較する.

4.2　多様サンプリングによる翻訳候補生成

評価セット中の各ソース文 q に対して，初期モデルおよび蒸留モデルそれぞれから $N = 50$ 回の翻訳をサンプリング生成した．両モデルとも，同一の評価セット，同一のデコード設定（温度や top-p など）で評価を行い，モデル以外の条件差が結果に影響しないように統制した．得られたスコア系列を，問題 q ごとの繰り返し測定とみなし，平均と分散の両方を統計的に比較する.

4.3　評価指標

各翻訳候補に対して，以下の 2 系統の指標を算出した.

4.3.1　自動評価指標

参照訳との一致度に基づく標準的な自動評価指標として，BLEU および METEOR を用いた.

4.3.2　LLM-as-a-Judge スコア

翻訳の自然さや金融ドメイン特有のスタイルをより直接的に評価するため，Judge モデルとして LLM（ChatGPT4.1）を用いた LLM-as-a-Judge ベースの自動評価も併用した．評価軸は以下の 5 つとした.

- **Fluency**：訳文の流暢さ，文法的自然さ
- **Adequacy**：原文の意味内容がどれだけ忠実に保持されているか
- **Market Style**：金融市場レポートとしての用語選択・文体の適切さ
- **Coverage**：情報の抜け漏れの少なさ
- **Overall**：総合的な翻訳品質

Judge モデルには，原文と候補訳を入力し，各軸について 0–4 の整数スコアを付与させた.

4.4　サンプル単位の統計量

上記の自動指標および LLM-as-a-Judge スコアについて，各評価サンプル q ごとに，モデル m の平均スコアと分散を算出した．具体的には，モデル $m \in$ 初期, 蒸留，問題 q，サンプリングインデックス $i = 1, \ldots, N$ に対してスコア $s_{m,q,i}$ が与えられているとき，

$$\mu_{m,q} = \frac{1}{N}\sum_{i=1}^{N} s_{m,q,i},\ \sigma^2_{m,q} \quad = \mathrm{Var}\big(s_{m,q,1:N}\big) \tag{1}$$

と定義する（ここで Var は不偏分散）．その上で，問題 q ごとにモデル間差分を

$$d_q = \mu_{\text{蒸留},q} - \mu_{\text{初期},q},\ v_q \quad = \log \sigma^2_{\text{蒸留},q} - \log \sigma^2_{\text{初期},q} \tag{2}$$

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

と定義し，d_q および v_q を評価セット全体にわたる「性能差」「安定性差」の系列として扱う．d_q は各問題における平均スコアの改善量，v_q は分散比の対数（$v_q < 0$ で蒸留モデルの方がばらつきが小さい）を意味する．

平均・分散スコアの差の検定　平均スコア差系列 d_q , 分散差系列 v_q の両者について，以下の 2 種類の検定を行った．

- **対応のある t 検定**：d_q,v_q が正規分布に従うと仮定したときの平均差の検定として，両側一標本 t 検定を実施した．
- **Wilcoxon の符号付順位検定**：d_q,v_q を順序尺度として扱い，分布仮定に依存しないノンパラメトリック検定として Wilcoxon 符号付順位検定を適用した．

5　結果（Results）

5.1　自動評価指標（BLEU/METEOR）

参照ベースの自動評価指標 BLEU および METEOR では，いずれも初期モデルが平均値で優位だった．METEOR は蒸留モデル 0.519，初期モデル 0.580（平均差 −0.062，95% CI [−0.069, −0.055]，Cohen's $d = -0.509$，$p < 0.01$），BLEU は蒸留モデル 29.69，初期モデル 36.50（平均差 −6.82, 95% CI [−7.67, −6.02]，Cohen's $d = -0.496$，$p < 0.01$）であった．

一方，スコア分散の比較では METEOR の分散比が 0.65（95% CI [0.57, 0.76]），BLEU の分散比が 0.61（95% CI [0.50, 0.75]）となり，蒸留モデルの方が出力のばらつきはやや小さいことが分かった．

5.2　LLM-as-a-Judge による評価結果

ChatGPT-4.1 を Judge モデルとし，Fluency，Adequacy，Style，Coverage，Overall の 5 指標で評価したところ，すべての指標で蒸留モデルが初期モデルを平均スコア・分散の両面で上回った．代表的な数値を以下に示す．

- **Fluency:** 平均差 +0.080（95% CI [0.066, 0.093]），Cohen's $d = 0.366$，蒸留モデルの勝率 64.7%.
- **Adequacy:** 平均差 +0.211（95% CI [0.194, 0.228]），Cohen's $d = 0.768$，蒸留モデルの勝率 80.9%.
- **Style:** 平均差 +0.066（95% CI [0.053, 0.080]），Cohen's $d = 0.297$.
- **Coverage:** 平均差 +0.088（95% CI [0.076, 0.099]），Cohen's $d = 0.471$.
- **Overall:** 平均差 +0.178（95% CI [0.161, 0.195]），Cohen's $d = 0.632$，蒸留モデルの勝率 75.7%.

全ての指標で対応あり t 検定・Wilcoxon 検定ともに $p < 0.01$ となり，統計的に有意な差が確認された．

5.3　分散（安定性）の比較

各指標について，50 サンプルから得られるスコア分散の log 比 $v_q = \log \sigma^2_{\text{蒸留},q} - \log \sigma^2_{\text{初期},q}$ を比較した．$v_q < 0$ は蒸留モデルの方が安定であることを意味する．

結果として，すべての Judge 指標で蒸留モデルの分散は大きく減少した．例として，

- **Fluency:** 分散比 $\exp(-2.10) \approx 0.12$（初期モデルの約 1/8）
- **Adequacy:** 分散比 $\exp(-4.64) \approx 0.01$（初期モデルの約 1/100）
- **Style:** 分散比はほぼ $\exp(-3.80) \approx 0.02$（初期モデルの約 1/50）
- **Overall:** 分散比 $\exp(-2.89) \approx 0.056$（初期モデルの約 1/18）

以上より，提案する蒸留モデルは，参照ベース指標では平均スコアで劣るものの，Judge ベース評価では平均スコアと安定性の両面で初期モデルを明確に上回り，「品質の下振れを大幅に抑制しつつ安定した翻訳品質を出力する」ことが示された．実験結果の詳細は付録の表 1，2 に示す．

6　おわりに

本研究では，金融市場レポートという高頻度更新かつ文体制約の厳しいドメインを対象に，新規のデータ蒸留パイプラインを構築し，金融市場翻訳特化 LLM を蒸留した．参照ベース指標では初期モデルに劣る一方で，Judge ベース評価では平均品質と安定性の両面で優位性を示し，品質の下振れを大きく抑制できることを確認した．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics**, pp. 86–96, 2016.

[2] Ming Zheng, Kun Zhou, and Jiajun Wu. Domain adaptation of llama for it-oriented machine translation. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, 2023.

[3] Yasmin Moslem, Thai Le, and Josef van Genabith. Adaptive domain-specific machine translation with mistral 7b. In **Proceedings of the 2024 Conference of the Association for Computational Linguistics**, 2024.

[4] Rui Zhang, Shuohang Wang, and Xiaodong Liu. Qlora fine-tuning of llms for high-quality machine translation. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, 2024.

[5] Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, 2016.

[6] Bowen Tan, Yi Ren, Shang-Wen Li Zhang, Junxian He, et al. Continual knowledge distillation for neural machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1580–1597, 2021.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

A　付録：実験結果の詳細

表 1　t 平均スコアの比較：蒸留モデル vs 初期モデル

metric	$Mean_{蒸留}$	$Mean_{初期}$	Y (questions)	$mean(d_q)$	$median(d_q)$	$CI95$	Cohen's d	$tstat_p$	$wilcoxon_p$	$sign_p$	win	loss	tie
fluency_score	3.793	3.713	1025	0.080	0.060	[0.066, 0.093]	0.366	0.000	0.000	0.000	0.647	0.278	0.075
adequacy_score	3.842	3.631	1025	0.211	0.180	[0.194, 0.228]	0.768	0.000	0.000	0.000	0.809	0.129	0.062
style_score	3.865	3.799	1025	0.066	0.040	[0.053, 0.080]	0.297	0.000	0.000	0.000	0.620	0.267	0.113
coverage_score	3.953	3.865	1025	0.088	0.040	[0.076, 0.099]	0.471	0.000	0.000	0.000	0.662	0.138	0.200
overall_score	3.785	3.607	1026	0.178	0.160	[0.161, 0.195]	0.632	0.000	0.000	0.000	0.757	0.177	0.065
meteor_score	0.519	0.580	1043	-0.062	-0.055	[-0.069, -0.055]	-0.509	0.000	0.000	0.000	0.266	0.733	0.001
blue	29.685	36.500	1043	-6.816	-5.281	[-7.668, -6.016]	-0.496	0.000	0.000	0.000	0.278	0.722	0.000

表 2　分散・安定性の比較：蒸留モデル vs 初期モデル

metric	$mean(v_q)$	$median(v_q)$	CI95	Cohen's d	$tstat_p$	$wilcoxon_p$	$sign_p$	win	loss	tie	分散比の平均	分散比の 95%信頼区間
fluency_score	-2.101	-0.301	[-2.641, -1.575]	-0.239	0.000	0.000	0.000	0.265	0.674	0.060	0.122	[0.071, 0.207]
adequacy_score	-4.642	-0.714	[-5.299, -4.010]	-0.442	0.000	0.000	0.000	0.175	0.780	0.046	0.010	[0.005, 0.018]
style_score	-3.804	-0.573	[-4.507, -3.093]	-0.334	0.000	0.000	0.000	0.269	0.627	0.103	0.022	[0.011, 0.045]
coverage_score	-9.272	-1.608	[-10.139, -8.406]	-0.649	0.000	0.000	0.000	0.146	0.667	0.186	0.000	[0.000, 0.000]
overall_score	-2.887	-0.491	[-3.425, -2.347]	-0.331	0.000	0.000	0.000	0.214	0.738	0.048	0.056	[0.033, 0.096]
meteor_score	-0.424	-0.519	[-0.564, -0.276]	-0.182	0.000	0.000	0.000	0.264	0.734	0.002	0.654	[0.569, 0.759]
blue	-0.499	-0.632	[-0.692, -0.291]	-0.150	0.000	0.000	0.000	0.247	0.752	0.001	0.607	[0.500, 0.748]

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).