

Aligning Multilingual News for Stock Return Prediction

Yuntao Wu, Lynn Tao, Ing-Haw Cheng, Charles Martineau, Yoshio Nozawa, John Hull, Andreas Veneris

Introduction & Contribution

We propose a method to align sentences in multilingual news articles using optimal transport, identifying semantically similar content across languages. Aligned sentences are sparser, more interpretable, and exhibit higher semantic similarity. Return scores constructed from aligned sentences show stronger correlations with realized stock returns, and long-short trading strategies based on these alignments achieve higher Sharpe ratios than analyzing the full text sample.

Our work:

- Analyzes **full news content** rather than aggregate sentiment scores
- Employs **sentence-level alignment** to capture semantic similarity
- Produces **sparser, more interpretable** cross-lingual alignments
- Applies the method to **financial text analysis**

Methodology

From Bloomberg, we collect an average of **17,800 stock-days** per year with both **English and Japanese** news for **3,500 Tokyo Stock Exchange stocks**.

Goal: Align English and Japanese articles at the **sentence level** for each stock-day, capturing **semantically similar content** while minimizing misalignments.

Optimal Transport (OT) maps probability mass between two distributions while minimizing transport cost. Kantorovich formulation (discrete):

$$\min_{\gamma} \left\{ \sum_{ij} c_{ij} \gamma_{ij} : \gamma_{ij} \geq 0, \sum_i \gamma_{ij} = p_{y_j}, \sum_j \gamma_{ij} = p_{x_i} \right\}$$

- x_i : English sentence embedding with empirical distribution p_{x_i}
- y_j : Japanese sentence embedding with empirical distribution p_{y_j}
- c_{ij} : cosine distance between x_i and y_j , scaled to [0,1]

After solving for γ (J→E) and γ' (E→J), compute a 1-1 sentence mapping:

$$\mathcal{A} = A * B^T * (\xi_{ij} > \xi_{thres})$$

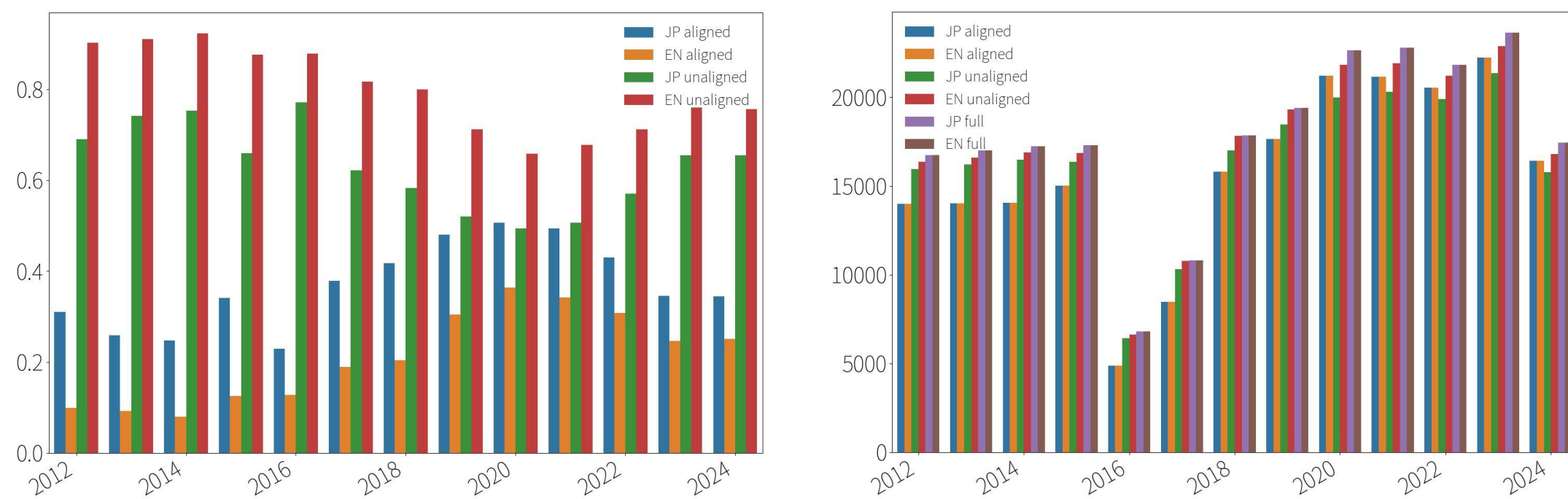
- A : J→E alignment, row max and top 5% by column
- B : E→J alignment, row max and top 5% by column
- ξ_{ij} : cosine similarity between x_i and y_j

Construct group-level embeddings by averaging sentence embeddings

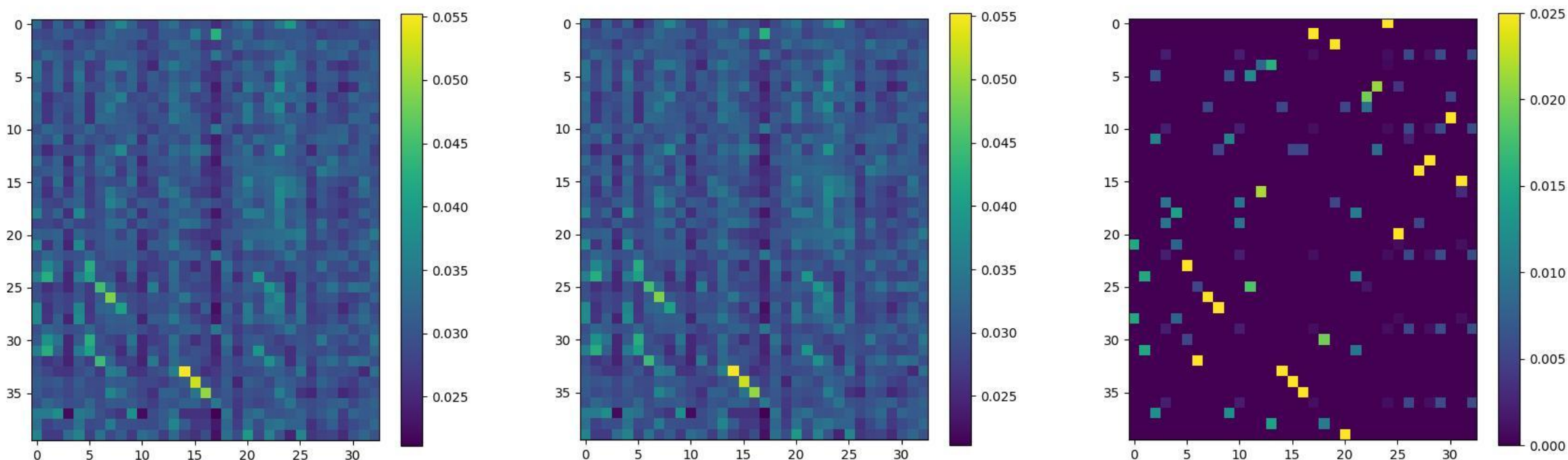
- Unaligned: average over sentences with $\mathcal{A} = 0$
- Aligned: average over sentences with $\mathcal{A} = 1$
- Full: average over all sentences in the article

Alignment Results

- Most stock-day articles contain a mix of **aligned** and **unaligned** sentences
- In recent years, about **40–60%** of sentences are **mutually aligned** across languages
- Produces 1–1 sentence alignments based on **semantic similarity**



Proportion of aligned sentences and article counts with aligned/unaligned sentences



Sparsity of alignments computed by Softmax (left), Entmax (middle), Optimal Transport (right) on news articles discussing the Bank of Japan (8301.T) on 2023-01-04. Sample alignments for this article can be found in the paper.

Type	Article
English	<p>The market potential in China remains high and the company will continue investing despite the anti-corruption drive in the country's health sector and Beijing's plan to nurture domestic players in advanced medical devices manufacturing, said Olympus CEO Stefan Kaufmann Monday.</p> <p>Co.'s profit margins in China are expected to be at the same level as in other markets, said CFO Tatsuya Izumi</p> <p>Co.'s training program for endoscopists in China over the years have made its preferred choices for many doctors in China: Izumi</p>
Japanese	<p>The impact of Donald Trump's potential re-election is uncertain, but co will need to restructure supply chains affected by the de-globalization trend: Kaufmann</p> <p>オリンパスのシェフファン・カウフマン社長兼最高経営責任者（CEO）は、中国事業について同市場のポテンシャルは大きく、注力する姿勢は変わらないとした。</p> <p>カウフマン氏や最高財務責任者（CFO）の泉竜也氏がブルームバーグなどの取材に応じた。</p> <p>泉氏は中国市場の競争が激しくなる可能性はあるが、利益率が他の市場に比べて下がることを現時点では想定していないと述べた。</p> <p>オリンパスは長年中国で内視鏡医の育成に貢献しており、競争力があるとしている。</p> <p>中国での反腐敗運動や自国製品を優先的に購入する「バイ・チャイナ」の方針、購入量ベース調達制度（VBP）などで、医薬品や医療機器を扱う外資企業は課題に直面している。</p> <p>ドナルド・トランプ氏が大統領に返り咲いた場合について、カウフマン氏はアメリカの政策がどのように変化するかは分からないとしたうえで、ここ数年続くグローバル化の巻き戻しを受けた供給網（サプライチェーン）再構築の必要性は今後も続くとした。</p> <p>これまでは効率性や生産性が重視されてきたが、今後はレジリエンスが重要になると述べた。</p>
Translated	<p>Olympus President and CEO Stefan Kaufmann stated that the Chinese market has great potential and the company remains committed to its business.</p> <p>Kaufmann and CFO Tatsuya Izumi were among those interviewed by Bloomberg and other media outlets.</p> <p>Izumi said that while competition in the Chinese market may intensify, he does not currently expect profit margins to fall compared to other markets.</p> <p>Olympus has long contributed to the training of endoscopists in China, and cited this as a competitive advantage.</p> <p>Foreign companies dealing in pharmaceuticals and medical devices face challenges due to China's anti-corruption campaign, the "Buy China" policy, and the volume-based purchasing system (VBP).</p> <p>Regarding a return of Donald Trump to the presidency, Kaufmann said it is unclear how US policy would change, but noted that the need to restructure supply chains following the recent rollback of globalization will continue.</p> <p>He said that while efficiency and productivity have been emphasized in the past, resilience will become more important going forward.</p>

Sentence alignments between English and Japanese news for Olympus (7733.T) on 2024-07-23. Translation of the Japanese news is provided by Google Translate.

The 2nd Workshop on LLMs and Generative AI for Finance



UNIVERSITY OF
TORONTO



Link to Paper

Trading Strategy Performance

Apply Ridge regression to link stock returns with text embeddings, generating return scores $\text{Soft}^{l,k}$:

$$\arg \min_w \left\| X_{c,t}^{l,k} w - \text{Ret}_{c,t}^{OC} \right\|^2 + \lambda \|w\|^2$$

- $X_{c,t}^{l,k}$: embedding in language l and alignment type k , using articles published between 8:30 am day $t - 1$ and 8:30 am day t for each firm c
- $\text{Ret}_{c,t}^{OC}$: open-to-close return from Compustat Global
- λ : regularization parameter chosen by cross-validation
- Rolling window: 6-year training, 1-year out-of-sample prediction
- Evaluation Period: 2018-2024
- Trading strategy: long top 25% and short bottom 25% of stocks by return score (days with ≥ 20 traded stocks), **equally weighted**

	Ret	Soft ^{EN,A}	Soft ^{EN,UA}	Soft ^{EN}	Soft ^{JP,A}	Soft ^{JP,UA}	Soft ^{JP}
Ret	1.00	0.02	0.03	0.03	0.02	0.01	0.03
Soft ^{EN,A}	0.02	1.00	0.41	0.67	0.67	0.38	0.57
Soft ^{EN,UA}	0.03	0.41	1.00	0.76	0.49	0.47	0.50
Soft ^{EN}	0.03	0.67	0.76	1.00	0.62	0.41	0.62
Soft ^{JP,A}	0.02	0.67	0.49	0.62	1.00	0.44	0.72
Soft ^{JP,UA}	0.01	0.38	0.47	0.41	0.44	1.00	0.71
Soft ^{JP}	0.03	0.57	0.50	0.62	0.72	0.71	1.00

Return Score Correlations

Alignment	Lang	Geo Mean	Mean	Std	5%	50%	95%	Sharpe	Ann. Sharpe
Aligned	JP	0.35%	0.36%	1.32%	-1.76%	0.39%	2.31%	0.27	4.36
	EN	0.28%	0.29%	1.34%	-1.86%	0.23%	2.46%	0.22	3.42
Unaligned	JP	0.17%	0.18%	1.34%	-1.98%	0.16%	2.29%	0.13	2.12
	EN	0.23%	0.24%	1.29%	-1.66%	0.16%	2.45%	0.18	2.91
Full	JP	0.30%	0.31%	1.23%	-1.59%	0.27%	2.38%	0.25	3.98
	EN	0.24%	0.25%	1.16%	-1.65%	0.17%	2.20%	0.21	3.40

Strategy Summary

Acknowledgements

This research was carried out at Rotman School of Management, FinHub Financial Innovation Lab, University of Toronto. We gratefully acknowledge financial support and insightful discussions from Royal Bank of Canada (RBC) Capital Markets. New York, NY and Toronto, ON, Canada.

References

- Yifei Chen, Bryan Kelly, and Dacheng Xiu. 2022. Expected Returns and Large Language Models. <https://ssrn.com/abstract=4416687>
- Zi-Yi Dou and Graham Neubig. 2021. Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 2112–2128. <https://doi.org/10.18653/v1/2021.eacl-main.181>
- Robert B. Durand, Joyce Khuu, and Lee A. Smales. 2023. Lost in translation. When sentiment metrics for one market are derived from two different languages. Journal of Behavioral and Experimental Finance 39 (2023), 100825. <https://doi.org/10.1016/j.jbef.2023.100825>
- Chenrui Wang and Yixuan Qiu. 2025. The Sparse-Plus-Low-Rank Quasi-Newton Method for Entropic-Regularized Optimal Transport. In Forty-second International Conference on Machine Learning. <https://openreview.net/forum?id=WCKMkMcqpb>