※この欄には何も記入しないこと

2025 年 4 月・2024 年 9 月入学東京工業大学大学院修士課程・専門職学位課程

志望理由書

氏名	勅使川原　奏大	志望する系	経営工学系

(1) GPA

GPA	3.9	所属大学	東京工業大学

（小数第 3 位は四捨五入し小数第 2 位まで表示）

算定方式　GPA は大学により算定方法が異なります。所属する大学の GPA の算定方法のうち、当てはまるものを選択してください。「その他」の場合には、その算定方法における満点も記入してください。

□ 5 点満点
✔ 4.5 点満点（東京工業大学の算定方法）
□ 4 点満点
□ その他（ _____点満点 ）

(2) 志望理由（500 字程度）

私が、経営工学系および中田先生の研究室を志望する理由は 2 つある。

1 つ目は、数学を応用して様々な問題を解決するという分野に関心があるからである。具体的には、数理最適化や機械学習について関心があり、これらの分野で活躍できるような能力を身に着けたいと考えている。そして、中田先生の研究室は、データ解析コンペティションへの参加など、そのような分野に集中して取り組むことができるような機会・環境が整っていると考えている。よって、修士課程ではこの研究室に所属することを希望している。

2 つ目は、機械学習を活用できる分野の 1 つである自然言語処理の研究に携わりたいと考えているからである。特に、テキストの類似性評価とその活用に関する研究に取り組む予定である。この類似性評価は、関連テキストの検索やテキスト生成など様々な活用先があるものとなっている。そこで、中田先生の研究室にて、最適化や機械学習に関する能力を高めつつ、この類似性評価に関する研究を発展させていきたいと考えている。

以上 2 つの理由から、私は経営工学系および中田先生の研究室で学修を行うことを志望する。

(3)研究計画

研究テーマ	最適輸送を用いたタグ(単語)の新類似度の開発

- このページを 1 ページ目として A4 用紙 3 ページ以内（参考文献のページは含まない）で書くこと。
- 適切に先行研究を引用しながら以下を記述すること。なお、テーマの選択に当たっては第一志望の指導教員の研究室に所属した場合を想定してよい。

(a) 研究の目的・新規性・内容・計画
(b) その研究とこれまでの学修との関連
(c) その研究に対する自身の能力・適性

(a) 研究の目的・新規性・内容・計画

本研究の目的は、2 つのテキストの類似性の測定を行うための、より優れた手法を開発することである。具体的には、求職者のレジュメと求人票の内容を比較する際に利用できるような類似性測定手法の開発を行う。性能の良い類似性測定手法を開発することで、求職者と求人者間でのより適切なマッチングを行えるようになることが期待される。

テキスト間の類似性を測定するにあたっては、まず単語埋め込みという技術を用いて、単語が持つ情報を実数ベクトルで表現する必要がある。有名な手法としては Sentence-BERT[1]などが存在する。これにより、各単語を定量的に取り扱うことが可能となる。また、単語の情報が埋め込まれたベクトル空間においては、意味が近い単語は近い位置に配置されるようになっている。例えば、“apple”という単語を表すベクトルは、国名である“Japan”を表すベクトルよりも、同じ果物を表す“orange”という単語を表すベク

トルにより近い場所に位置しているはずである。

さて、このように単語をベクトル空間上の点として表すと、単語の集まりであるテキストはそのベクトル空間における点群とみなせる。つまり、比較したい 2 つの文$\boldsymbol{a}, \boldsymbol{b}$は次のように表現できる。

$$\boldsymbol{a} = (w_1, \dots, w_n), \qquad \boldsymbol{b} = (w'_1, \dots, w'_m)$$

よって、テキストの類似性比較という問題は、2 つの点群をどのようにして比較するかという問題に帰着する。2 つの点群を比較する手法として、点群を構成する単語ベクトルの平均ベクトルを考え、それを比較することで点群間のを比較するという単純な手法が考えられる。しかし、この手法では、単語がもつ情報が平均化の段階で欠落してしまっている可能性が高い。

そこで、一方の点群を他方の点群まで移動させる最適輸送問題を考え、その最小輸送コストを類似度の指標として使用する、WMD(Word Mover's Distance)というアイデアが考え出された[2]。このアイデアでは、2 つの文$\boldsymbol{a}, \boldsymbol{b}$を一様な重みづけをした離散分布$\mu_a, \mu_b$とし、各単語間のコスト$c_{ij}$をユークリッド距離を用いて定義する。

$$\mu_a = \{(w_i, 1/n)\}_{i=1}^{n},\ \mu_b = \{(w'_j, 1/m)\}_{j=1}^{m},\ c_{ij} = \|w_i - w'_j\|_2$$

このようにして、2 つの文の類似性評価を 2 つの点群間の最適輸送問題に帰着させ、その最適輸送コストを文$\boldsymbol{a}$と文$\boldsymbol{b}$の類似度の指標にするというものである。

しかしこの手法も完全ではない。というのも、文中の各単語ベクトルに対する重み付けが全て同じであることや、コストをユークリッド距離で算出することが必ずしも適切とは限らないからである。この点を踏まえた改善手法として、単語ベクトルのノルムと方向ベクトルを分けて活用する WRD(Word Rotator's Distance)という手法が開発されている[3]。本テーマでも、WMD を参考に、データの特性などを踏まえた適切な手法を開発したいと考えている。

また、最適輸送問題で厳密解法を利用すると、最悪の場合入力サイズの 3 乗に比例する計算時間を要してしまい[4]、類似度の指標として実用性が低下してしまう。しかし、最適輸送問題時代は長く研究されており、様々な近似解法が存在する。それらも活用しながら、実用性の高い手法を構築したい。

なお、研究に利用するデータセットとしては、共同研究先のデータを利用するほか、一般に使われている web サイトからデータセットを作成することも検討している。

(b) その研究とこれまでの学修との関連

自分はこれまで、特に最適化問題について関心をもって学修に取り組んでいたため、自然言語処理という分野自体を直接学んだ経験はなかった。しかし、一見あまり関係がなさそうな最適輸送問題に自然言語処理の類似性に関する問題を帰着させるというアプローチ方法が非常に興味深く感じたため、研究テーマとして取り組むことにした。

(c) その研究に対する自身の能力・適性

python によるデータ処理や最適輸送問題に関する基礎的な内容はある程度習得しており、今後も学修を続けていくつもりである。また、自然言語処理分野はこれまであまり学修してはいないものの、必要に応じて論文や言語処理学会の機関紙を見るなどして、学修を進めている。したがって、研究テーマを進めていくにあたって、能力面・適性面で問題はないと考えている。

参考文献

[1] N. Reimers, and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In EMNLP-IJCNLP, 2019

[2] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From Word Embeddings To Document Distances. In 32nd International Conference on Machine Learning, PMLR. 2015.

[3] S. Yokoi, R. Takahashi, R. Akama, J. Suzuki, and K. Inui. Word Rotator's Distance. In Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

[4] 佐藤竜馬. 最適輸送の理論とアルゴリズム. 講談社. 2023