Vol.2021-MPS-132 No.17
2021/3/1

多様な類似観点を反映するテキスト検索ランキングのための学習データ補正方法

藤城 真祥[†, a)]　小池 大地[†,b)]　末永 高志[†,c)]

概要：テキスト間の類似性をもとにした検索タスクにおいては，表層の一致度合いや概念の類似といった様々な観点が存在する．これらの観点はそれぞれに有用性があり，複数の観点を加味した検索結果の提示が期待される．このような複数の観点を加味するためには，それぞれの観点で計測された類似度のスコアを加重平均したものが考えられるが，同義語のように表層は異なるが概念的には類似する事例を上位に挙げることが難しい．これに対して，特定の観点で類似性が高い事例も優先するようなスコアリングのロジックを，非線形な機械学習手法を適用することで作成することが考えられる．しかしながら，テキスト検索といったタスクでは，明確に負例を定義することが困難で，正例以外の事例が膨大に存在する．また，これらのデータの分布は正例のデータを包含するような関係性になるため，正例以外のデータを負例として見なして学習することには課題がある．本研究では，テキスト検索のタスクを対象に，特定の観点でも類似性の高い事例を優先することを目的とした，学習データの補正方法を提案する．

キーワード：テキスト検索，スコアリング手法，機械学習

1. はじめに

テキスト検索のタスクにおいて，単語の一致による候補を抽出するキーワードマッチ検索だけではなく，同義語の検索を可能とする概念類似の観点や，書き損じを含む場合の検索を可能とする表層一致の観点などの，多様な類似観点を考慮した検索に対する期待が高まっている．テキスト間の類似観点として挙げた表層一致や概念類似は，それぞれに異なる有用性があるため，複数観点を加味した検索結果の提示が期待される[1,2]．本研究では，複数の類似観点を反映したテキスト検索技術に関する検討を行う．

一般に，表層一致および概念類似それぞれの観点における類似度を算出し，得られた類似度を加重平均により統合したスコアに基づく検索は，単独観点での検索と比較して精度の改善が期待できる．これは前述した通り，それぞれの観点に異なる有用性があり，統合することでそれぞれの観点を補完するためである．しかしながら，クエリと正解が同義語関係となる事例においては，表層一致の類似度が著しく低いため，概念類似との加重平均によるスコアも低くなり，概念類似単独と比較して精度が低下する．

本稿では，上記の問題を解決するべく 2 つの課題に取り組む．1 つ目は，一方の観点の類似度のみ著しく低い事例も統合後のスコアを高くするため，非線形な機械学習手法を用いたスコアリングロジックを検討する．2 つ目は，機械学習に適用する学習データに対して，スコア値の分布を考慮したデータの生成方法を検討する．

2. 提案法

2.1 非線形な機械学習手法によるスコアリングロジック

一方の観点の類似度が著しく低い事例であっても，統合後のスコアを高くするためには，各観点の類似度に応じて統合時の重みを変更する必要がある．そこで，類似度を非線形統合するロジックを，学習データの分布から決定する機械学習手法を用いることを考える．学習を行う際には，テキスト間の表層一致と概念類似それぞれの類似度を入力として，類似，非類似の二値判別の確率を推定する手法が適用可能である．検索の際には，検索対象とのスコアをもとに求められた類似判別の確率を統合スコアと見立て，スコア順に検索対象を出力する．

2.2 学習データの生成方法

二値判別のロジックを機械学習の手法で学習させるためには正例と負例が必要となる．一方で，テキスト検索のようなタスクにおいては，人が判断した適切な検索結果を正例として設定できるが，明確に負例を設定することが困難である．例えば，正解以外の事例を負例とする場合，正例の近傍にも当該データが多く存在し，正例を包含するようなデータ分布となる．図 1 は後述する本検証データにおける正例と正例以外のデータに対するスコアの分布である．この図から，横軸の概念の類似度が高く，縦軸の表層一致の類似度が低い領域において，正例以外のデータが正例のデータを包含しており，このデータで学習すると正解以外のデータを優先する判別ロジックが作成されることになる．

この課題を解決するため，負例とするデータを人為的に作成した分布に従うよう生成する．これにより，一方の観点の類似度が高く，もう一方の観点の類似度が著しく低い場合であっても，統合スコアリングは大きな値となる非線形統合ロジックを作成する．

3. 検証

3.1 検証データ

本検証では，商標における指定商品名の検索を対象とする．指定商品名とは，商標登録出願した商標が対象とする商品の権利範囲を定めるものである．

商標の出願人は，指定商品名を記載する必要があるが，

[†]株式会社 NTT データ 技術開発本部 AI 技術センタ
[a)]Misaki.Fujishiro@nttdata.com [b)]Daichi.Koike@nttdata.com
[c)]Takashi.Suenaga@nttdata.com

自由記述も可能であり，曖昧な表現となることも多い．権利範囲を明確にするために内容の特定と適切な表現への修正が必要となり，審査に時間を要する．審査の効率化のためには出願人が効率的に修正が不要な基準となる商品名（以下，基準商品名）を選定できるサービスが期待されており，今回検討するような検索技術の重要性が高まっている領域である．

今回利用する検証データは，審査済みの出願商標，1240 件を対象とし，基準商品名約 1.6 万件の中から適切な商品名を検索することとした．クエリとなるデータは，出願時点で自由記述された指定商品名であり，登録時に修正された基準商品名を正解とした．

評価方法としては，検索対象の基準商品名 1.6 万件を統合スコアでランキングし，正解となる商品名が N 位以内に提示される N 位順位率を用いた．

3.2 評価

表層一致の観点の類似度算出には編集距離を採用した．概念類似の観点の類似度算出には，FastText で商品名のテキストを分散表現したベクトル間のコサイン類似度を採用した．それぞれの手法で算出された類似度の平均をとることで，N 位順位率が向上することを確認した．一方で，表 1 に示す通り，クエリに対する正解が同義語関係にあり，表層一致の類似度が著しく低いために，スコアの平均を取ることで単独観点での検索と比較して順位が低下する事例が存在した．

このように，検索の上位に挙げたいものは，双方のスコアが高いだけでなく，一方のスコアが著しく高いものも含まれる．このようなタスクに対して，二値判別を行う機械学習に適した負例のデータを考察すると，双方の観点のスコアが低いものは多数存在し，スコアが高くなるにつれて少数となるものが適切であるといえる．本検証では，このようなデータの分布を表現するために，原点を中心とした多次元正規分布に基づいて負例を生成した．学習データの生成に利用する多次元正規分布の標準偏差は，事前検証としてそれぞれ[0.15, 0.2, 0.25, 0.3]の間で変化させ，10 位順位率が最も高くなった 0.2 を採用した．生成された学習データの分布を図 2 に示す．非線形な機械学習手法には，ニューロン数を 8，隠れ層は 2 層とするニューラル・ネットワークを採用した．

提案法により正解が提示される順位が向上した事例を表 2 に示す．平均を用いた場合，表層一致のスコアが非常に高くなる「ブラシ」を含む事例のみが上位となっている．一方で，提案法では，表層一致のみが高い事例だけでなく，「パンフレット」や正解事例などの編集距離のスコアが小さい事例を上位としている．

各方式で得られた N 位順位率のグラフを図 3 に示す．平均によるスコアリングでの順位と提案法のスコアリングの順位を組み合わせることで，1 位に正解を提示する割合は，平均と比較して 3.5 ポイント，10 位までに正解を提示する割合は 1.4 ポイント改善し，検証データの 87%を 10 位までに提示可能であった．

以上の結果より，提案法により平均では上位とすることのできない事例を補完し，検索結果の多様性を向上させる効果があるといえる．

4. おわりに

本稿では多様な類似観点を反映するテキスト検索ランキングのための学習データ補正方法を提案し，商標の指定商品名の類似性判断のタスクにおいて有効性を確認した．

参考文献

[1] M. Al-Asa'd, et al.: “Question to Question Similarity Analysis using Morphological, Syntactic, Semantic, and Lexical Features,” 2019 IEEE/ACS 16th AICCSA, pp. 1-6 (2019).

[2] B.Bejuk, et al.: “Solving Community Question Answering Ranking Problem Using LightGBM, ” Text Analysis and Retrieval 2018 Course Project Reports, 15(2018).

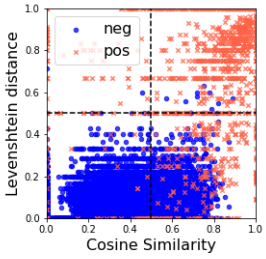

図 1 実データの分布

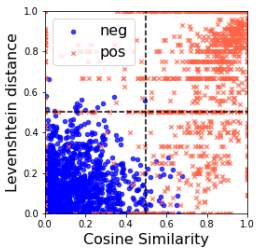

図 2 生成されたデータ分布

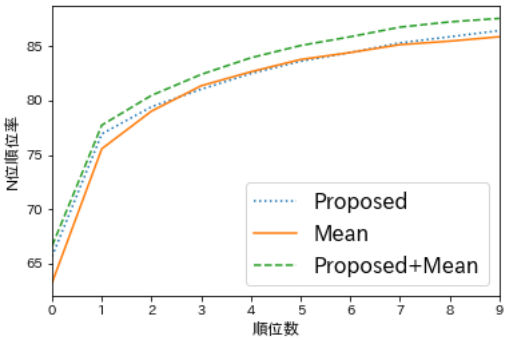

図 3 各手法の N 位順位率の積み立てグラフ

表 1 平均で上位とできない事例の各手法での順位

クエリ	正解	編集距離	FastText	平均
チラシ	広告・販売促進用印刷物	9217 位	3 位	47 位

表 2 正解の順位が向上した事例

クエリ	正解	手法	1 位	2 位	3 位	4 位
チラシ	広告・販売促進用印刷物	平均	ブラシ	歯ブラシ	靴ブラシ	金ブラシ
		提案手法	ブラシ	ポスター	パンフレット	広告・販売促進用印刷物