論 文 の 内 容 の 要 旨

論文題目 Category-enhanced Embedding Model for Massive Text Data
（カテゴリを用いた大規模テキストデータのための
分散表現獲得手法に関する研究）

氏　　名　　丸井　淳己

本論文は、カテゴリを用いた分散表現獲得手法によって大規模テキストデータの多様なエンティティを分析する手法を提案し、その有効性を調べた研究である。

ビッグデータの時代において、実世界エンティティはウェブ上やデータベースにテキスト形式を伴って表現されることが多くなってきた。それにつれ、電子的に記録された膨大なデータから有用な知識を抽出する必要性が近年高まっている。しかしウェブ上またはデータベース上のエンティティは構造的に整理されたデータではないことが多く、様々な形式で書かれたテキストを伴っている。そのような場合、多様なエンティティを把握・理解するために我々はカテゴリと呼ばれるようなエンティティを包含するまとまりを用いることが多い。ここで議論するカテゴリは必ずしも構造的になっている必要はなく、人間が整理のためにエンティティを共通の性質でまとめたグループのことを指す。こういったカテゴリには、ECサイトの商品カテゴリや学術論文のキーワードといった明示的につけられたカテゴリ(明示的カテゴリ)と、ソーシャルメディアのユーザのコミュニティに挙げられるような似たエンティティをまとめる手法によって得られた暗黙的なカテゴリ（暗黙的カテゴリ）が含まれる。このようなカテゴリをエンティティの情報に加えて用いることで、提案手法はカテゴリからエンティティの類似性を学習でき、逆にエンティティからカテゴリの類似性も学習できるため、エンティティの俯瞰的解析を行うのに用いることができる。

本論文において提案する手法は、カテゴリとエンティティの類似性を学習する分散表現獲得手法である。分散表現はエンティティを固定長の実数値ベクトルに射影して得られるもので、自然言語処理では近年この手法の研究が進んだ。その理由としては、自然言語コーパスに多数存在する低頻度語の表現を、その文脈として出現する高頻度語を用いて学習できるためである。現代多くの企業や研究機関が直面する大規模データには多数のカテゴリとエンティティがあり、比較的低頻度なエンティティやカテゴリは単語と

同様に従来手法により扱いにくかったことから、本論文ではこのアイデアをカテゴリとエンティティに対しても適用した。

筆者はカテゴリベクトルモデルという、カテゴリ・エンティティ・単語の分散表現を同時に学習させるモデルを提案し、その有効性を従来手法であるBag-of-words手法と既存の分散表現獲得手法である段落ベクトルモデルと比較した。ECサイトの商品カテゴリ（明示的カテゴリ）とソーシャルメディアのコミュニティ(暗黙的カテゴリ)の両方において、それぞれの手法でテキストからエンティティの表現を生成し、カテゴリを推測するタスクにその表現を用いた。それらの精度を比較した結果、提案手法であるカテゴリベクトルモデルによるエンティティの表現を用いると最も精度よくカテゴリを推測することができることがわかった。また、同じ単語がカテゴリ間でどのように異なる分散表現を得るのかソーシャルメディアのデータを用いて調べると、カテゴリベクトルの類似したカテゴリ同士で得られる単語の分散表現もまた類似することがわかり、カテゴリベクトルがユーザの書き込みの傾向を表現していることが確認できた。

本論文の提案手法はオントロジーや構造的知識を用いた既存のトップダウン的アプローチと比較すると汎用的に手に入るカテゴリを用いているためデータの制約が少ないため応用範囲が広く、Bag-of-wordsのようなボトムアップ的アプローチと比較するとカテゴリを用いることでエンティティの関係性をより良く扱えるため、それらの中間に位置付けられる研究である。また大規模データに対しても適用可能なスケーラビリティを持っているため、企業の持つ大規模データに存在する種々のエンティティを俯瞰して意思決定をするためのツールとすることができる。提案手法が、大規模データに対してカテゴリを用いる新たな分析手法の基礎にとなることを筆者は期待している。

見本 Sample

論 文 の 内 容 の 要 旨

論文題目　　□□□□□□□□□□□□□□□□□□□□

(□□□□□□□□□□□□□□□□□□□□)

↑

（※論文目録の記載と同じにしてください。）

（※論文題目が外国語の場合には、和訳を括弧書きで付けてください。）

(* The title typed here must be identical to that shown in the Thesis Table of Contents.)

(* Add a Japanese translation in parentheses if the thesis title is written in a non-Japanese language.)

氏　　名　　○○　○○

↑

（※学位記に記載される氏名と同じにしてください。）

（※漢字圏以外の外国人は、カタカナ表記となります。）

(* Type your name in the same manner as you want to have shown on your degree certificate.)

(* Type your name in katakana if you are a non-Japanese without a kanji name.)

　□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

１．**「論文の内容の要旨」は、紙媒体を２部提出してください。**
また、「PDF ファイル」及び「文書ファイル（Word 等で作成したもの）（省略可。）」の電子データも併せて提出してください。

２．**論文博士は日本文で**記入してください。（課程博士は英文でもよい。）
横書き、片面刷りとしてください。

３．大きさはＡ４判とし４ページ以内、１０ポイント程度の活字で印刷したものとしてください。
（日本語の場合は 4, 000 字以内（英語の場合は 2, 000 語以内）とする。）
４．第１ページ上部に、タイトルを「論文の内容の要旨」とした上で、論文題目及び氏名を記入し、その下から内容の要旨を記載してください。

1. **Two copies of your thesis summary must be submitted in paper form.**
   Electronic data of the thesis summary must also be submitted: a PDF file is mandatory, while submission of the original document file (MS Word or other) is optional.

2. **If you are obtaining your Doctorate degree by submitting a thesis (as a Ronpaku),** your thesis summary **must be written in Japanese. (**If you are obtaining your degree by completing the course requirements of a Doctorate program, a thesis summary in English is acceptable.)
   The thesis summary is formatted with **horizontal writing and single-sided print**.

3. The thesis summary is to be printed on **A4-size paper** and digested into **four pages or less** using **approximately a 10 point type**.
   (The restriction is **4,000 characters** for a Japanese summary and **2,000 words** for an English summary.)

4. **In the upper part of the first page, the text “論文の内容の要旨” is typed and the title of the thesis and the name of the applicant are typed on subsequent lines. The main text of the thesis summary begins below the above heading section on the same page.**