様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　５　月　２６　日現在

機関番号：　３４３１６
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2016
課題番号：　２５３３０３６８

研究課題名（和文）快適なWeb検索のための検索用語の獲得支援に関する研究

研究課題名（英文）Development of Information Retrieval Support Systems

研究代表者
馬　青（Ma, Qing）

龍谷大学・理工学部・教授

研究者番号：３０３５８８８２

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：現在の検索エンジンはその性能が非常によくなってきており、適切な検索用語（キーワード）さえ与えてやればおおむね期待通りの検索結果が得られる．しかし一方、多くのユーザ、特に子どもや高齢者、外国人などにとって検索対象を表す適切な検索用語（特に専門用語など）を見つけることは往々にしてそう簡単ではない．提案研究は、このような障壁を取り払うことを目的とし、IT や医療など様々な分野において、これらの分野の関連語・周辺語またはそれらの語から構成される文を手掛かりに適切な検索用語（この場合「システム復元」）を高精度に予測する検索支援システムを、深層学習手法や最新の情報検索技術などを用いて開発した。

研究成果の概要（英文）：The current Web search engines have a very high retrieval performance as long as the proper retrieval terms are given. However, many people, particularly children, seniors, and foreigners, have difficulty deciding on the proper retrieval terms for representing the retrieval objects, especially with searches related to technical fields. In this study, we developed a support system that can highly accurately predict suitable retrieval term candidates when some clues such as their descriptive texts or relevant/surrounding words are given by the users, adopting deep learning methods and the latest information retrieval technology.

研究分野：　自然言語処理

キーワード：　検索支援　用語予測　深層学習　DBN/SdA　非構造化文書　用語抽出

１．研究開始当初の背景

Web ページの整備および検索技術に関するＩＣＴ技術が普及し、人々の生活は向上してきている。現在の Web 検索エンジンについてはさまざまな研究・開発が行われ、検索性能は非常に優れたものとなっている。ただし、これはユーザが適切な検索のための語（検索用語）を入力することを前提としている。しかし、一般人、特に子どもや高齢者、外国人などは検索対象を示す適切な検索用語（専門用語など）を知らないことが多く、検索自体が行えない、すなわち ICT を利活用できないという問題（障壁）がある。政府や文部科学省の推進事項でもある、学校教育現場を含めた一般社会に真に ICT 技術を普及させ、利活用できる社会を実現するためには、このような障壁を取り払う必要がある。

２．研究の目的

提案研究は、このような障壁を取り払うことを目的とし、適切な検索用語の提示システムの基盤形成を目指した。具体的には、IT や医療など様々な分野において、これらの分野の関連語・周辺語（たとえば「コンピュータ」、「前の状態」、「戻す」）またはそれらの語から構成される文を手掛かりに適切な検索用語（この場合「システム復元」）を予測・提示する検索支援システムの開発を目標とした。

３．研究の方法

検索用語を高精度に予測するために、近年、様々な分野で注目され、音声認識や画像認識、自然言語処理の諸課題への応用にも優れた性能を出している深層学習を用いた。深層学習の実現には Deep Belief Network (DBN)と Stacked Denoising Autoencoder (SdA)の両方を用いた。また、過学習を防ぎ、汎化能力を向上させるために、L1 正則化、L2 正則化、さらに Dropout を導入した。

機械学習を用いて関連語・周辺語から検索用語を予測・提示する場合，その学習データとして、入力（関連語・周辺語）と正解となるレスポンス（検索用語、以降略して「用語」）のペアからなるコーパスが必要となる。検索用語を説明している文書には関連語・周辺語が多く含まれると考え、インターネットからコンピュータ関連に限定した Web ページを手動と自動の二通りの方法で収集した。手動収集では人手で用語を説明する Web ページを選別し収集した。一方、自動収集では，用語の後に「とは」「は」「というものは」「については」「の意味は」の 5 語を付けて（たとえば，用語が「グラフィックボード」であれば「グラフィックボードとは」「グラフィックボードというものは」などで）Google で検索したものを説明文書として収集した。手動収集データは規模が小さい代わりに精度が高く、自動収集データは精度が低い代わりに規模が大きい。そのため、自動収集データを深層学習の教師なし学習に用い、手動収集データを深層学習の教師あり学習に用いた。また、学習データを増やし汎化能力のさらなる向上を目指すために、手動収集データに適度なノイズを加える疑似データも用いることにした。

収集したデータに対し、形態素解析など自然言語処理技術を用いて Bag of Words 方法で機械学習に必要特徴ベクトルへの変換を行った。また、機械学習の最適なパラメータはグリッドサーチと交差検証を行って決定した。

機械学習に用いる上記データは基本的に「見出し語とその説明」という構造を持つものととらえることができる（「見出し語」＝「用語」）。しかし、多くの場合、探したい用語はこのような構造化された文書に存在するのではない。「見出し語とその説明」という構造が含まれない文書（非構造化文書）から用語を抽出するために、機械学習に基づくアプローチとは別に、最新の情報検索技術と用語抽出技術を用いた。用語抽出は、クエリ（関連語・周辺語または説明文）と最も近い文書中の一部（パッセージ）を検索し、そのパッセージに含まれる語を用語候補とし、それらを絞り込むことにより行う。用語抽出の性能向上を図るために、パッセージ選択の精度向上と用語候補絞り込みの精度向上の両面から検討した。パッセージ選択の精度向上には、広域文書の利用と最適なパッセージの決定を行った。一方、用語候補絞り込みの精度向上には、用語候補とクエリ間の近さを測るなどして用語らしさのスコアをより適切なものに補正することや、複数の説明文による用語の一括抽出の代わりに説明文ごとによる用語抽出を行った。

４．研究成果

まず、10 個の検索用語を用いた小規模な実験において、テストデータに対し 0.97 という高い予測精度が得られた。この精度はベースライン手法よりはるかに高く従来の機械学習手法の Multi-Layer Perceptron (MLP) と Support Vector Machine (SVM) のいずれよりも高かった。また、手動収集データに自動収集のデータと疑似データを加えて学習することにより予測精度は向上した。さらに，よりノイズの多い学習データを加えても深層学習 の予測精度はさらに向上したのに対し、MLP の精度向上は見られなかった。このことから、深層学習 のほうが MLP よりもノイズの多い学習データを有効利用できることが分かった。本実験では有意差検定により結果に有意差があることも確認された。なお、上記実験結果は数多い単語からなる説明文を用いた予測結果であった。少数キーワードによる予測実験も行った。表１は各用語の予測に用いた少数キーワード（関連語・周辺語

とノイズ語)を示す。表 2 は予測精度を示す。ただし、m は手動収集データ、a は自動収集データ、p は疑似データを表す。その後ろに付いている数字はデータの個数である。この結果より、提案手法の実用性が確認できた。本研究成果は自然言語処理の学術論文と国際会議にて発表した。

表 1　予測に用いるキーワード

用語	関連語・周辺語	ノイズ語
CPU	頭脳、計算、コア	管理
グラフィックボード	映像、ディスプレイ、描画	デザイン
ハードディスク	磁気ヘッド、円盤、プラッタ	読み込み
メインメモリ	作業、処理速度、アクセス	メディア
マザーボード	基盤、ソケット、チップセット	頭脳
OS	管理、Windows、基本ソフトウエア	計算
光学ドライブ	メディア、再生、レーザー	円盤
PC ケース	箱、デザイン、収納	コンセント
電源ユニット	供給、電圧、変換	箱
SSD	衝撃、フラッシュメモリ、振動	プラッタ

表 2　予測精度

学習データ数	ノイズ語なし	ノイズ語あり
m300	1.0	0.9
a300	1.0	0.9
a600	0.9	0.8
a1200	1.0	0.8
a2400	0.9	0.7
p300	1.0	1.0
p600	1.0	1.0
p1200	1.0	1.0
p2400	1.0	1.0
a300p300	1.0	1.0
a600p600	1.0	1.0
a1200p1200	1.0	1.0
a2400p2400	1.0	1.0
平均	0.985	0.931

実験の規模を 10 倍に拡大して実験も行った。その結果、自動収集データの教師なし学習への利用は予測精度の向上に大きく寄与することがわかった（図 1)。また、深層学習が従来の機械学習より精度がよいこともわかった（図 2)。本研究成果は国際会議などにて発表した。

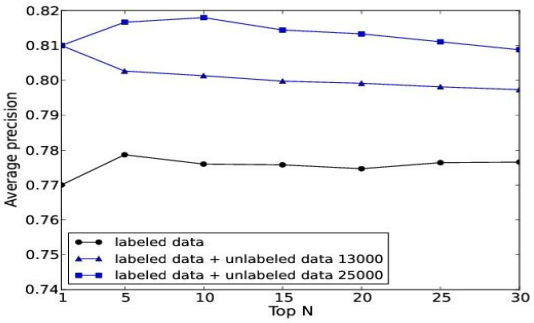

図 1 自動収集データありなしの予測精度

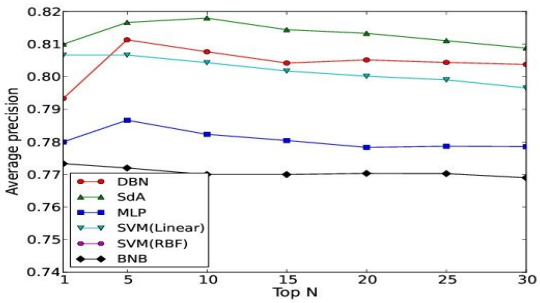

図 2 各機械学習手法の予測精度

一方、非構造化文書からの用語検索においては、検索クエリと類似度の高い文書（パッセージ）には、正解となる用語が含まれると仮定して、パッセージから関連語を選択し、その関連語のスコアを基に用語候補として出力する手法を提案し、有効性を確認した。さらなる検索精度向上のために、文書選択システムの改良、最適なパッセージ単位の検討、用語候補のスコアをリスコアリングする手法、複数の説明文からそれぞれの用語候補を抽出してスコアを統合する手法を提案し、提案手法が妥当であり、有望があることを確認した。本研究成果は情報処理学会論文誌（テクニカルノート）などで発表した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 4 件）

① 森田直樹，南條浩輝，山本凌紀，馬青：音声ドキュメントを検索対象とした用語検索，情報処理学会論文誌(テクニカルノート)，査読有，Vol.58，No.3，2017，pp. 762-767

② 馬青，谷河息吹，村田真樹：Deep Belief Netwokr を用いた検索用語の予測，自然言語処理，査読有，Vol. 22，No. 4，2015，pp. 225-250
③ 吉見毅彦，小谷克則：既存対訳辞書と複数のシソーラスを用いた類義表現の生成による対訳辞書の拡充，Information，査読有，Vol.18，No.2，2015，pp. 609-624
④ K. Kotani，T. Yoshimi and H. Isahara: Application of Reading Data in an Integrated Learner Corpus，Procedia - Social and Behavioral Sciences，査読有，Vol. 95，2013，513-521

〔学会発表〕（計 14 件）
① 加藤 玲大，馬青，村田真樹：機械学習を用いた QA サイト質問文のカテゴリの類推，言語処理学会第 23 回年次大会，2017 年 03 月 14 日～2017 年 03 月 16 日，筑波大学（筑波キャンパス 春日エリア）
② 森田直樹，南條浩輝，馬青：複数の入力説明文を用いた音声ドキュメントからの用語検索，言語処理学会第 23 回年次大会，2017 年 03 月 14 日～2017 年 03 月 16 日，筑波大学（筑波キャンパス 春日エリア）
③ Q. Ma，I. Tanigawa，and M. Murata: Retrieval Term Prediction Using Deep Learning Methods，The 30th Pacific Asia Conference on Language，Information and Computation (Paclic 30)，査読有，Oct. 28-30，2016，Seoul
④ 加藤玲大，馬青，村田真樹：深層学習を用いた QA サイト質問文のカテゴリ分類，情報処理学会研究報告 Vol. 2016-NL-228，2016 年 09 月 29 日～2016 年 09 月 30 日，大阪大学（吹田キャンパス）
⑤ 森田直樹，南條浩輝，馬青：非構造化文書からの用語検索における用語候補のリスコアリングの検討，情報処理学会研究報告 SLP-111/NL-226，2016 年 05 月 16 日～2016 年 05 月 17 日，東京工業大学（大岡山キャンパス）
⑥ Q. Ma: Comparison between Deep Learning and Conventional Machine Learning Methods on Retrieval Term Prediction，The 15th China-Japan Joint Conference on Natural Language Processing (CJCNLP2015)，Oct. 18-19，Aomori
⑦ 森田直樹，南條浩輝，山本凌紀，馬青：説明文を入力とした非構造化文書からの用語検索の検討，情報処理学会研究報告 SLP-109，2015 年 12 月 02 日～2015 年 12 月 03 日，名古屋工業大学
⑧ 渡邊和弥，馬青：N グラムコーパスを用いた IT 用語の意味ベクトルの獲得，言語処理学会 第 21 回年次大会，2015 年 03 月 19 日，京都大学
⑨ 谷河息吹，馬青，村田真樹：検索語の予測における Deep Learning と従来の機械学習との比較，言語処理学会 第 21 回年次大会，2015 年 03 月 18 日，京都大学
⑩ Q. Ma，I. Tanigawa，and M. Murata: Retrieval Term Prediction Using Deep Belief Networks，The 28th Pacific Asia Conference on Language，Information and Computing (Paclic 28)，査読有，Dec. 13，2014
⑪ H. Nanjo，T. Yoshimi，S. Maeda and T. Nishio，Spoken Document Retrieval Experiments for SpokenQuery&Doc at Ryukoku University (RYSDT)，NTCIR-11 Workshop Meeting，Dec. 12，2014，Tokyo
⑫ 南條浩輝，吉見毅彦：講演音声ドキュメント検索における反復的擬似適合性フィードバックの検討，日本音響学会 2014 年秋季研究発表会，2014 年 09 月 04 日，北海学園大学
⑬ 谷河息吹，馬青，村田真樹：Deep Belief Network を用いた関連語・周辺語からの検索用語の予測，言語処理学会第 20 回年次大会，2014 年 03 月 19 日，北海道大学
⑭ 二輪和博，馬青：Stacked Denoising Autoencoder を用いた語義判別，言語処理学会第 20 回年次大会，2014 年 03 月 19 日，北海道大学

６．研究組織
(1)研究代表者
馬　青（MA QING）
龍谷大学・理工学部・教授
研究者番号：30358882

(2)研究分担者
吉見 毅彦（YOSHIMI TAKEHIKO）
龍谷大学・理工学部・准教授
研究者番号：50368031

南條浩輝（NANJO HIROAKI）
京都大学・学術情報メディアセンター・准教授
研究者番号：50388162

(3)連携研究者
(4)研究協力者