Title	整っていないビッグデータを対象とした文書解析法の開発
Sub Title	Language processing of big data in versatile formats
Author	斎藤, 博昭(Saito, Hiroaki)
Publisher	慶應義塾大学
Publication year	2018
Jtitle	学事振興資金研究成果実績報告書 (2017.)
JaLC DOI	
Abstract	整っていない文書ビッグデータとして, 国立情報学研究所が研究用に配布している25万件余りの不満調査データセットを用いた。これは多人数の書込みにより作成された, 書式, 文体, 文長が不統一なさまざまな種類の不満・苦情を集めた日本語データである。このデータを対象として, 語やフレーズを指定して, それに該当する文書を取り出すタスクを設定した。このような検索処理を実時間で行なうには, インターネット上の検索エンジンで行なわれているように, 文書を前もってベクトル表現に変換しておく必要がある。過年度の研究においては, 各次元が特定の内容語に相当するような数万次元の単語ベクトルを用いた。また, 例えば「セーター」という語から服装品に関係する不満であることを認識する汎化手法を過年度は試した。今年度は検索の際にさらに自由度をもたせ, 入力語に関連する単語をユーザに提示し, 指定された関連語も含めた検索ができるようにした。関連語の提示に関しては二種類のモデルを用意した。一つはWikipediaで学習したもの, もう一つは不満調査データから作成したモデルである。「政治」という単語を例に取ると, 前者のモデルからは経済, 社会, 財政, 外交, 改革などの関連語が取得でき, 後者のモデルからは自民党, 格差, 政府, 与党, 首相といった関連語が取得できた。一般的な傾向として, 後者のモデルからは具体的で時事的な語が取得できることがわかった。このモデル作成においては, 深層学習のフリーソフトウェアであるword2vecを採用し, 200次元のベクトルを用いた。関連語も検索に加えることで, 再現率が向上する一方, ユーザからの入力を要求する煩雑さにつながる場合もあり, この手法が常に最善のアプローチであることを主張するものではないが, 形式が整っていないデータを対象とした文書処理手法として一定の成果を示すことはできた。 Big document data are often written in versatile formats with peculial wordings, jargons, colloquial expressions, to name a few. We chose "Complaint data set" collected in various situations from many people, whose data were made available by NII for research purposes. We need to convert these data into some vector representation to attain realtime search, just like regular search engines on the Web. Each dimension of the vector is assigned to a content word, thus the number of dimensions exceeds tens of thousands. This year we enabled the user to add related words interactively in addition to the original words/phrases. The candidates of the related words are hypothesized from two models, one is made from Wikipedia and the other is made from the Complaint data set. The latter model tends to produce topical words, while the former reflects common usages. This kind of interactive approach might impose a burden on the user, but more prolific search results can be obtained.
Notes	
Genre	Research Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=2017000001-20170039

2018 年 4 月 4 日

2017 年度　学事振興資金（個人研究）研究成果実績報告書

研究代表者	所属	理工学部	職名	准教授	補助額	200 （B）千円
	氏名	斎藤 博昭	氏名（英語）	Hiroaki SAITO		

研究課題（日本語）

整っていないビッグデータを対象とした文書解析法の開発

研究課題（英訳）

Language Processing of Big Data in Versatile Formats

１．研究成果実績の概要

整っていない文書ビッグデータとして、国立情報学研究所が研究用に配布している 25 万件余りの不満調査データセットを用いた。これは多人数の書込みにより作成された、書式、文体、文長が不統一なさまざまな種類の不満・苦情を集めた日本語データである。このデータを対象として、語やフレーズを指定して、それに該当する文書を取り出すタスクを設定した。このような検索処理を実時間で行なうには、インターネット上の検索エンジンで行なわれているように、文書を前もってベクトル表現に変換しておく必要がある。過年度の研究においては、各次元が特定の内容語に相当するような数万次元の単語ベクトルを用いた。また、例えば「セーター」という語から服装品に関係する不満であることを認識する汎化手法を過年度は試した。今年度は検索の際にさらに自由度をもたせ、入力語に関連する単語をユーザに提示し、指定された関連語も含めた検索ができるようにした。関連語の提示に関しては二種類のモデルを用意した。一つは Wikipedia で学習したもの、もう一つは不満調査データから作成したモデルである。「政治」という単語を例に取ると、前者のモデルからは経済、社会、財政、外交、改革などの関連語が取得でき、後者のモデルからは自民党、格差、政府、与党、首相といった関連語が取得できた。一般的な傾向として、後者のモデルからは具体的で時事的な語が取得できることがわかった。このモデル作成においては、深層学習のフリーソフトウェアである word2vec を採用し、200 次元のベクトルを用いた。関連語も検索に加えることで、再現率が向上する一方、ユーザからの入力を要求する煩雑さにつながる場合もあり、この手法が常に最善のアプローチであることを主張するものではないが、形式が整っていないデータを対象とした文書処理手法として一定の成果を示すことはできた。

２．研究成果実績の概要（英訳）

Big document data are often written in versatile formats with peculial wordings, jargons, colloquial expressions, to name a few. We chose ″Complaint data set″ collected in various situations from many people, whose data were made available by NII for research purposes. We need to convert these data into some vector representation to attain realtime search, just like regular search engines on the Web. Each dimension of the vector is assigned to a content word, thus the number of dimensions exceeds tens of thousands. This year we enabled the user to add related words interactively in addition to the original words/phrases. The candidates of the related words are hypothesized from two models, one is made from Wikipedia and the other is made from the Complaint data set. The latter model tends to produce topical words, while the former reflects common usages. This kind of interactive approach might impose a burden on the user, but more prolific search results can be obtained.

３．本研究課題に関する発表

発表者氏名 （著者・講演者）	発表課題名 （著書名・演題）	発表学術誌名 （著書発行所・講演学会）	学術誌発行年月 （著書発行年月・講演年月）
原田裕生，末廣駿，斎藤博昭	分散表現を用いた対話型不満調査データセット検索システム	言語処理学会第 24 回年次大会	2018 年 3 月