7ZH-02

ベクトル情報が付与された単語による検索支援システムの研究

北野太陽† 大島千佳† 中山功一†
佐賀大学理工学部知能情報システム学科†

1. 研究背景

私たちは日常的に動作や様子，ものやことなどについて，共通認識された単語として扱うことで，読み書きや会話などのコミュニケーションができる．しかし，同じ単語であっても，状況によって，内包する意味が異なる場合がある．さらに，時代の変化とともに，単語の意味も変わっていく．

私たちは，意味の不明な単語に接した時，前後の文脈によって，その単語の意味を推測しようとする．キーワード検索においても同様に，その単語の文字列だけではなく，その単語が使用されている文脈によって，その単語の意味をより適切に特定できれば，より適切な情報が得られると期待される．

本研究では，単語の意味をベクトルとして定量的に定める fastText というライブラリを使用し，より効率よく情報取得するための支援を目的とする．キーワードとして用いられる単語の辞書的な意味ではなく，その単語が使われている文脈上での意味が表現されたベクトルとして利用することで，より適切な検索結果が得られると期待する．

2. 想定するシステムの使用方法

ユーザが PC 上でテキストを読んで，不明な単語(以下，未知単語とする)を見つけた時に，本システムが使用される．使用方法を，以下に図 1 と併せて説明する．

(1) ユーザは，Web ページ内の未知単語をクリップボードにコピーする．

(2) 本システムは，コピーされた未知単語を含む Web ページ内の全てのテキスト情報に基づき，未知単語のベクトルを作成する

(3) 本システムは，未知単語の単語ベクトルと，検索対象となる Web ページに存在する未知単語と同じ文字列である単語の中から，近い単語ベクトルを持つ単語を検索結果として表示する．

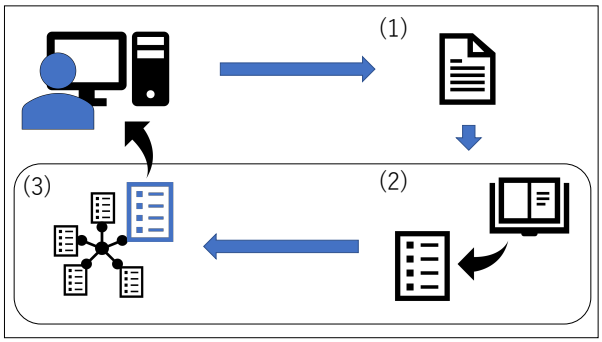

図 1：想定する本システムの使用方法

以上の処理により，自分が調べたい単語の意味に合致した意味で使われている Web ページを検索できるようになると期待する．

3. システムで利用する関連技術

3.1 fastText[1]

fastText は Facebook AI Research が 2016 年に開発した自然言語処理向けアルゴリズムであり、GitHub にてオープンソースとして公開している単語のベクトル化とテキスト分類をサポートした機械学習ライブラリである．fastText は，他のアルゴリズムと比較して動作が軽く速いのが特徴である．

3.2 MeCab[2]

MeCab は，京都大学情報学研究科−日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジンであり，言語，辞書,コーパスに依存しない汎用的な設計を基本方針としている．パラメータの推定に Conditional Random Fields(CRF)を用いており，ChaSen が採用している隠れマルコフモデルに比

A Search Support System Based on Words Represented by Vectors
†Hiroaki Kitano, Chika Oshima, Koichi Nakayama
†Department of Information Science, Saga University

べ性能が向上している。また、平均的に ChaSen, Juman, KAKASI より高速に動作する．

4. システムの構成

本研究では，fastText によりベクトル化された単語を使用して，Wikipedia からの情報取得を支援するシステムの開発を試みる．単語のベクトル化には，fastText を使用する．また，日本語に対して fastText を使用するにあたり，分かち書きという単語ごとに空白を書き込む処理が必要である．その分かち書き処理には，形態素解析エンジンである MeCab を使用する．ベクトルの類似度比較には gensim を使用する．システムの手順は以下の通りである

(1) 検索単語を 5 つ以上含む Wikipedia の記事ごとに単語のベクトル化を行い，モデルを作成する．

(2) (1)のモデルから，検索単語のベクトルのみを抽出し，検索単語のみのモデルを作成する．このベクトルには番号付けする．

(3) 検索単語の意味が一意(以下，一意的検索単語)になるようなテキストを集め，そのテキストに対して，単語のベクトル化を行い，モデルを作成する．

(4) (3)のモデルから，一意的検索単語のベクトルを抽出し，(2)で作成したモデルに合成する．

(5) (4)のモデルにおいて，一意的検索単語のベクトルと比較して，ベクトルが類似する順に検索単語ベクトルの番号を取得する．

(6) (5)で取得した番号をもとに，該当する Wikipedia 記事の URL を取得する．

5. 評価実験

被験者計 10 人を 5 人ずつの A グループと B グループに分けて，実験を行う．各グループの検索単語は共通にする．また，検索単語の意味が異なるようにテキストを用意し，システムを使用する．表 1 は，検索単語と，テキスト中に含まれる検索単語の簡易的な意味の一覧である．表 1 における意味の解釈は，実験に使用するテキストを読んだ著者の解釈である．

検索単語 1 つにつき，システムから取得した URL 上位 5 つと，Wikipedia で検索し表示される URL 上位 5 つを使用する．その 10 の URL を使用したページ(以下，検索ページとする)をランダムで見せて，以下の手順で評価する．

(1) 被験者は，検索単語を含むテキストから，検索単語を含む一部を抜粋したものを読む．

(2) 被験者は，5 つの検索単語に対して，それぞれ 10 の検索ページを読んで，それぞれ検索結果としての適切さを 5 段階で主観評価する．

以上の実験手順を A グループ，B グループの被験者 1 人ずつ行い，ベクトル情報が付与された単語による検索の有効性を検証する．実験結果と考察については，発表の際に紹介する．

表 1：実験で用いる検索キーワード

検索単語	意味（A 群）	意味（B 群）
マフラー	襟巻き．防寒具.	エンジンの消音装置.
チェック	格子模様.	検査すること
ノリ	リズムにのる	海苔（食品）
ブロック	ゲーム用語	区分．地域
ATM	非同期転送モード.	現金自動預け払い.

6. 分析方法

(1) 実験結果を Wikipedia 内の検索ページとシステムで得られた検索ページに分類する．

(2) 分類した検索ページそれぞれに対して，グループごとに検索単語の評価を集計し，平均を求める．

(3) A グループ，B グループの実験結果を併せて，分類した検索ページごとに評価を集計し，平均を求める．

以上の集計結果を踏まえ，ベクトル情報が付与された単語による検索が有効であったか，また，なぜそのような結果になったのかを考察する．

7. おわりに

本研究では，ベクトル情報が付与された単語による検索支援システムを提案した．検索支援システムの有効性については，これから評価実験を行う予定である．

また，その実験結果を考察し，本研究内において，より扱いやすいシステムの開発に取り組む予定である．

参考文献

[1]https://nissenad-digitalhub.com/articles/facebook-fasttext/
[2] http://taku910.github.io/mecab/