3Q-02

単語ベクトルに基づく記録文書の概念検索

三澤 虎遊汰† 靳 展‡ 柴田 千尋† 田胡 和哉†
東京工科大学コンピュータサイエンス学部†
東京工科大学大学院バイオ・情報メディア研究科コンピュータサイエンス専攻‡

1 背景

近年，単語の同義語や上位下位語などの概念情報を登録したデータベースを用いた検索など様々なあいまい検索が存在する．そういった検索手法は概念情報によって類語を判断してしまう．そのため，ユーザの考える類語だけでなくそれ以上の類語で検索してしまう．これは複数人が編集する文書における検索手法には同義語や類語の判断が困難になるため適さない．そのため類語をクラスタリングしユーザの考える類語のみで検索することが必要となる．

本稿では検索時に複数の類語をクラスタリングすることで情報検索を行うツールを提案する．

2 提案手法

2.1 全体の流れ

図 1 に概念構成図を示す．ユーザの入力した単語に対し，複数の類語を抽出し，抽出された類語をグループ単位で分ける．例として「言語」という単語から「日本語，英語，Java，C」という 4 つの単語が類語として抽出された場合，国の言葉に関係する「日本語，英語」とプログラミング言語に関連する「Java，C」のグループに分ける．そして複数のグループをユーザに提案し，選択したグループに含まれる全ての単語で検索を行いユーザに提供する．これによりユーザの求める情報のみの提供が可能となる．

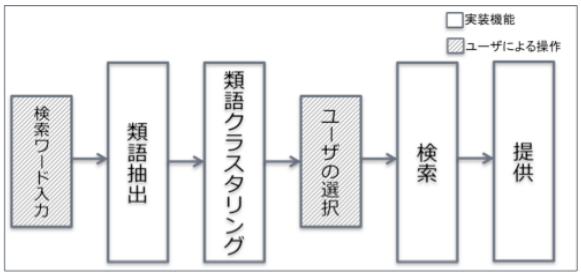


図 1 概念構成図

Concept search using vector representations of words from archive texts
† Koyuta MISAWA, Chihiro SHIBARA, Kazuya TAGO
School of Computer Science, Tokyo University of Technology
‡ Kin TEN
Bionics, Computer and media science, Entrepreneurship program, Tokyo University of Technology Graduate school.

2.2 必要となる機構

図 1 の実現には以下の機構が必要となる．

- 単語のベクトル化機構
  文書データに出現する単語を文脈による単語の関係性からベクトル表現を獲得する．検索時に入力された単語はこの作成した単語ベクトルのリストから選択される．
- 類語の抽出機構
  機械が類語や同義語を評価し抽出する機構である．抽出された単語ベクトルと数値的に近いベクトルを抽出することで，類語や同義語として評価することが可能となる．（図 1「類語抽出」）
- 類語クラスタリング機構
  語群からそれぞれ関連する単語同士でクラスタを分ける機構である．抽出された全ての単語ベクトル同士の距離によりクラスタを設定する．（図 1「類語クラスタリング」）
- 検索機構
  選択したクラスタに含まれる全ての単語に対してキーワード検索を行い，マッチした箇所をユーザに提供する．（図 1「検索」）

3 実装

3.1 実装対象

東京工科大学のクラウドサービスセンター[1]では古谷によって開発された「板」[2]が運用されている．板はプロジェクト内での情報共有を行うテキストエディタであり複数人での同時編集が可能である．本提案手法は板上の検索システムとしてここに実装した．

3.2 各機構の実装方法

「単語のベクトル化機構」と「類語の抽出機構」は word2vecd で実装することで，単語を質の良い低次元ベクトルへと写像することができ，結果の高精度化および処理速度の低下の防止を実現できる．また，「クラスタリング機構」については K-means++法を用いて実装した．これにより類語として抽出されたベクトルを自動で複数クラスタに分ける．

3.3 word2vec

word2vec[3]は Tomas Mikolov らによって提案されたニューラルネットワークである CBOW モデル， Skip-gram モデル[4]を扱うツールであり，オープンソースとして公開されている．

文書集合である学習データの文脈を元に単語の関係性をベクトル化し扱うことを可能とする．そのため，単語のベクトル情報には単語の前後の文脈情報が埋め込まれている．

3.4 K-means++法

K-means++法[5]は David Arthur らによって提案された K-means 法の初期値設定を改良した手法である．K-means は非構造型クラスタリング手法の一種であり，現在幅広く用いられている．K-means++の手順は以下のようになっている．

1. 1つ目のクラスタ中心をランダムに選ぶ．
2. それぞれのデータ点とその最近傍のクラスタの中心を求める．
3. その距離の 2 乗に比例した確率分布を用い，新たなクラスタ中心に選ぶ．
4. 全てのデータを各データ点に最も近いクラスタに割り当てる．
5. 各クラスタごとにクラスタ中心を求める．

ステップ 1-3 は K-means++の初期値設定であり，ステップ 4-5 は K-means の初期値設定以降に行う手順である．また，指定した個数のクラスタが選ばれるまでステップ 2-3 を行う．

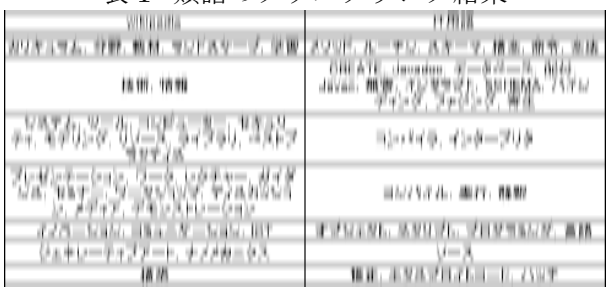
3.5 使用したデータ

word2vec では 2 つの文書データを使用し比較した．Wikipedia の大規模で一般的な語句の多く含まれる文書データと，クローラで収集した IT 用語に特化した文書データの 2 種類である．

4 結果

本ツールで「ソース」と検索した際のクラスタを表１に示す．

表１ 類語のクラスタリング結果



5 実験・評価

5.1 実験設定

数人に IT 関連の検索ワードを入力してもらい，抽出された類語がいくつ正しいかを調べる．次にクラスタに含まれる単語が正しいかを 5 段階で評価してもらう．

5.2 実験結果および評価

Wikipedia と IT 用語のデータを用いて各ワードについて抽出された類語の正解率を表 2 に示す.Wikipedia よりも、IT 用語に特化して収集した文書データからの類義語の方が、精度が極めて高いことがわかる．一方，クラスタの評価は「正しい，やや正しい，どちらともいえない，やや異なる，異なる」の 5 段階評価で行った．Wikipedia のクラスタでは「7 件,12 件,12 件,3 件， 2 件」IT 用語のクラスタでは「8 件,7 件,12 件,5 件,3 件」となり，平均の評価値はそれぞれ 3.52，3.34 となり，Wikipedia のクラスタの方がわずかに良い評価となった．

表 2 類語の正解率

検索ワード	wikipedia	IT用語
ディスプレイ	47%	87%
メモリ	20%	97%
iPhone	47%	67%
JPEG	40%	93%
JavaScript	37%	67%
平均	38%	82%

6 おわりに

学習データに IT 用語に特化した文書を用いることで IT 関連の情報のみを類語として抽出するため正解率が高くなった．また，Wikipedia のように一般的な用語を多く含む文書を用いることで IT 以外の情報がノイズとなる．そのため正解率が低下したと考えられる．IT 用語のクラスタの方が Wikipedia と比べ低い評価が多かったのは，情報量が少ないためにベクトルが偏ってしまったためだと考えられる．その為，板には IT 用語の学習モデルが向いていると考えられる．今後は，学習データの情報量を増やしクラスタの精度の向上させる．

参考文献

[1]http://www.teu.ac.jp/cloud/ (2015 年 1 月 8 日)
[2]古谷文弥.”知的財産作成を支援するクラウドアプリケーション IWB の実装と評価”. 情報処理学会第 77 回全国大会,予稿集, 2015
[3] Google code. https://code.google.com/p/word2vec/
[4] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. “Efficient estimation of word representations in vector space”. In Proceedings of Workshop at ICLR, 2013.
[5]David Arthur, Sergei Vassilvitskii. “k-means++: The Advantages of Careful Seeding”. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 1027-1035, 2007.