講演番号：57

講義検索での分散表現によるクエリ拡張の試作

Prototype of query expansion based on distributed representation in lecture search

向後ジェフリー [1]　小林亜樹 [1]
Jeffrey Kougo　Aki Kobayashi

工学院大学情報学部情報通信工学科 [1]
Department of Information and Communications Engineering, Faculty of Informatics, Kogakuin University[1]

1　はじめに

大学進学希望者を対象者として, 学びたいことから推奨学科を提供する検索システムの構築を行った. 大学シラバステキストを用いた講義検索では，利用者の語彙とシラバス中の語彙の不一致が問題である．そこで本研究室では，自然言語処理で一般的に使われるようになった単語や文章の分散表現 [1][2] を用いてクエリ拡張を行うことで，利用者の意図に沿う検索を行うシステム開発に取り組んでいる．本論文では，実用的に動作するシステム開発の一環として，シラバステキストデータ自体や検索用分散表現などを格納するデータベースや検索 API を試作する．

2　講義検索システム

2.1　概要

開発している講義検索システムは，利用者が検索語を入力すると，対応する講義の適応度を表す科目スコアが得られる．ここで，科目スコアの算出には，入力語である検索クエリを分散表現を用いたクエリ拡張を施した上で，分散表現を用いた文書間類似度を用いる．また，学科スコアはこの科目スコアを用いて算出される．複数大学のシラバステキストを用いるため，統一されたデータベーススキーマと，高速な分散表現の取得，類似度算出を実現する機能を提供する．システム全体は，シラバステキストを収集する Web クローラ，本検索データベースとスコア算出部，利用者インタフェースによって構成される．

2.2　前処理

Web クローラより，1 講義=1 json オブジェクトとして渡されるシラバステキスト（例 : 図 1）を RDB の講義情報テーブルへ格納しておく．Bag of words として取り扱うため，形態素解析を行い単語に分け，品詞などの付随情報と共に単語テーブルに格納する．このデータベースの ER 図を図 2 に示す．分散表現への変換，類似度算出を行う Word2Vec，Doc2Vec は，日本語版 Wikipedia コーパスを用いてモデルを構築しておく．モデル構築に用いたパラメータは，表 1, 2 に示す．講義毎の Bag of words を文書と見做し，構築した Doc2Vec モデルで検索するたび分散表現を得ることができる．

```
syllabus_data = {
    "university" = 大学名""
    "faculity" = 学部名""
    "department" = 学科名""
    "subject" = 講義名""
    "grade" = 学科学年コード""
    "code" = 履修コード""
    "document1" = 授業の狙い""
    "document2" = 具体的な到達目標""
    "document3" = 授業計画""
}
```

図 1　シラバステキスト

表 1　Word2Vec のパラメータ

パラメータ	数値	説明
size	300	ベクトルの次元数
window	15	コンテキスト周辺の単語数
min_count	5	学習に使う単語の最低出現数

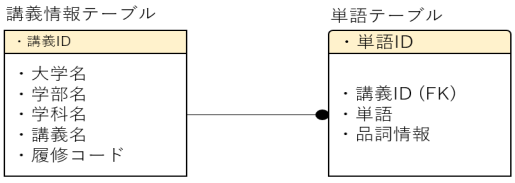

図 2　データベースの ER 図

表 2　Doc2Vec のパラメータ

パラメータ	数値	説明
dm	0	Doc2Vec の学習手法
vector_size	300	ベクトルの次元
window	15	コンテキスト周辺単語数
alpha	0.025	学習率
min_count	5	学習に使う単語の最低出現数
sample	1e-5	単語を無視する際の頻度の閾値
epochs	20	イテレーション回数
dbow_words	1	Skip-gram を追加

2.3　クエリ拡張

クエリ拡張は Word2Vec モデルを用いる．検索クエリを分散表現に変換した後，類似度上位 N 件（N は別に定めるパラメータ）を検索クエリ語に加えて，拡張されたクエリ語集合であるとする．

2.4　スコア

拡張クエリ語集合を文書と見做して，Doc2Vec モデルで分散表現を得る．これと，各講義との分散表現ベクトル空間でのコサイン類似度を各科目スコアとして算出する．

また，学科スコアは，当該学科に設置された全科目スコアの総和を用いる．（式 (1)）

$$score(g) = \sum_{\boldsymbol{d} \in \boldsymbol{D}_g} sim(\boldsymbol{q}, \boldsymbol{d}) \tag{1}$$

$\boldsymbol{q}$ は拡張クエリ語集合の分散表現, $\boldsymbol{d}$ は各講義の分散表現, $\boldsymbol{D}_g$ は学科 g における講義の分散表現の集合を表す.

3　クエリ拡張効果の評価

作成した試作検索システムを用いて，クエリ拡張の有無を含めた拡張語数 N の違いがスコアに及ぼす影響について評価実験を行う．検索クエリには大学案内の案内ページ中のキーワードを用いる．

4　おわりに

分散表現を用いたクエリ拡張をシラバステキストを用いた講義検索に適用するシステム構成について設計し，試作した．適切なパラメータを決定するための評価結果について取り纏める予定である．

参考文献

[1] Tomas Mikolov, Kai Chen,Greg Corrado, Jeffrey Dean ”Efficient Estimation of Word Representations in Vector Space”, ICLR, 2013

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrad, Jeffrey Dean(2018), ”Distributed Representations of Sentence and Documents”, Cornell University Library arXiv:1405.4053.

Copyright © 2020 IEICE