Vol.2015-SLP-109 No.16
2015/12/2

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

森田　直樹[1,a)]　南條　浩輝[2,b)]　山本　凌紀[3]　馬　青[1,c)]

概要：意味を表す文書表現（説明文）を入力として与え，その説明文が示す語句（用語）を検索する用語検索を行う．これまでは，辞書や Wikipedia などの構造化された文書の定義文を検索対象として用語検索を行うものが主である．これに対し，本研究では構造化されてない文書（非構造化文書）を検索対象とした検索を提案する．辞書などとは異なり，非構造化文書では用語とそれに対する定義文が結びついていない．そのため説明文と意味的に類似していると思われる文書の一部（パッセージ）を見つけ，そこから関連語を抽出することで用語候補とする方法を提案する．非構造化文書（検索対象）として講演音声ドキュメントを採用し，地名とカタカナ語の説明文それぞれ 25 文を用いて用語検索の実験を行った．上位 1000 件まで用語候補を出力したところ地名とカタカナ語についてそれぞれ 22 件，17 件見つけることができた．平均逆順位（MRR）はそれぞれ 0.058，0.030 であり，改善の余地が大きいことが確認できた．

キーワード：非構造化文書，用語検索，パッセージ検索

1. はじめに

説明文からそれが示す語句（用語）を検索する用語検索について検討を行う．外国人や子供では，語が示すものは頭に浮かぶものの，それを適切に示す語句がわからないまたは思い出せないということがある．また誰でも語句を思い出せないこともある．例えば，国の名前などの固有名詞，高度な専門知識に分類される単語のような普段あまり使われることのないような単語は思い出しにくい．

単語の意味を調べたいときには辞書やインターネットを利用して，その単語の説明文を得て意味を知ることができる．しかしその逆の手順，つまり説明文から単語を検索する手法は十分に研究されていない．これまでに辞書や Wikipedia 等のように見出しとその説明文を自身の構造として含んでいる文書（構造化文書）を検索対象とするものが研究されている [1][2][3][4]．しかし新語や一般的でない専門用語は辞書にのっていないことが多く，これらの手法では探し出せない．このような新語や専門用語は，マイクロブログや SNS，論文などで用いられていることが多い．本研究はマイクロブログや SNS，論文のような「見出し－説明文」という構造が存在しない文書（非構造化文書）を検索対象とし，用語検索を試みるものである．具体的には非構造化文書から入力の説明文と意味的に類似していると思われる文書の一部（パッセージ）を選択し，そこから関連語を抽出して用語候補とする方法を研究する．すなわち説明文と似ている文の周辺に，ターゲットとなる用語が含まれていると仮定して，用語検索をする方法を研究する．

[1] 龍谷大学理工学研究科,Graduate School of Science and Technology, Ryukoku University
[2] 京都大学学術情報メディアセンター,Academic Center for Computing and Media Studies, Kyoto University
[3] 龍谷大学理工学部,Faculty of Science and Technology, Ryukoku University
[a)] t15m007@mail.ryukoku.ac.jp
[b)] nanjo@media.kyoto-u.ac.jp
[c)] qma@math.ryukoku.ac.jp

2. データ

2.1 非構造化文書

本研究は非構造化文書を検索対象とする．非構造化文書には様々なものが考えられるが，検索対象として講演音声ドキュメント [5] を採用する．これは日本語話し言葉コーパス [6] の学会講演 987 件と模擬講演 1715 件の合計 2702 件の講演を検索対象とするものである．

講演音声ドキュメントの特徴として「見出し－説明文」という構造だけでなく句読点や段落情報がないことが挙げられる．つまりどこまでが 1 文であるのかを示す手がかりもない．したがって文書の一部の意味的に似たまとまりを見つけることも難しい．本論文では，無音（息つぎのポーズ）で区切られた音声を発話と定義し，10 発話をまとめて擬似的な意味のまとまりのパッセージとする．このパッセージのうち，説明文と近いものが見つかれば，そのパッセージ中に用語が含まれていると考え，そこから用語候補を取り出す．

2.2 検索クエリ

テストデータとして，音声認識の辞書に含まれている単語リストの中から地名とカタカナ語の単語をそれぞれ 25 個ずつ用語として選択した．その用語の説明を表す 3 文からなる説明文を作成した．説明文の例を以下に示す．

説明文の例

京都:
日本の関西の都市。清水寺や八坂神社といった寺や神社の名所多い。古都と呼ばれ歴史的価値のあるものが多い。
吉祥寺:
東京都武蔵野市。住みたい町ランキングに度々全国 1 位に。JR 中央線、京王井の頭線が通る。

これらの説明文が妥当であるかを調べるために 10 名の人に，説明文からもともとの用語を正しく連想できるかテストしたところ 86 %の正解率であり，この説明文は妥当であることがわかった．本研究では，これらの説明文を検索クエリとして用いる．

3. 検索システム

検索システムの全体像を図 1 に示す．これは文書選択，関連語選択，用語抽出から構成されるものである．以下それぞれについて述べる．

3.1 文書選択システム

本研究ではベクトル空間モデルに基づく文書選択システムを使用する．これは検索対象の文書の一部（パッセージ）のベクトル表現と検索クエリ (説明文) のベクトル表現の相関量を計算して関連度（スコア）の高い順にパッセージを選択するものである．文書をスコア付けするためには対象と
ル
SM
で
で

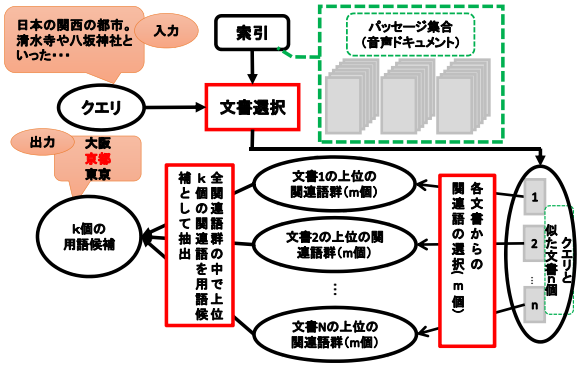

図 1 システムのイメージ図

$$SMART(Q, D_i) = \sum_{k=1}^{m} (q_{t_k} \cdot d_{i,t_k}) \tag{1}$$

ただし

$$d_{i,t_k} = \begin{cases} \dfrac{1 + \log(tf_{i,t_k})}{1 + \log(avtf)} \\ \hline (1 - slope) \cdot pivot + slope \cdot utf_i \\ 0 \end{cases} \quad \begin{array}{l} if \quad tf_{i,t_k} > 0 \\ otherwise \end{array} \tag{2}$$

$$q_{t_k} = \begin{cases} \dfrac{1+\log(qtf_{t_k})}{1+\log(avqtf)} \log \dfrac{N}{n_{t_k}} & if \quad qtf_{t_k} > 0 \\ 0 & otherwise \end{cases} \tag{3}$$

ここでは，tf_{i,t_k} は D_i 中の t_k の出現数，avtf は D_i における単語の出現回数の平均を表す．pivot は 1 ドキュメント中の異なり単語数の平均，utf_i は D_i 中の異なり単語数を表す．slope は補間係数 (0.2) である．qtf_{t_k} は Q 中での t_k の出現回数，$avqtf$ は Q に含まれる単語の出現回数の平均を表す．N は検索対象ドキュメント数を表す．n_{t_k} は t_k を含むドキュメント数を表す．

3.2 関連語の選択システム

文書選択をした結果，選択された上位 n 件の各文書 d (1≤ d ≤ n) のそれぞれをクエリ Q_d とみなして式 (3) に基づいて $q_{t_k,d}$ の値を求め，この値の降順で文書ごとに関連語を一定数（m 個）選択する．

3.3 用語の抽出システム

選択された各関連語 t_k について，$q_{t_k,d}$ の合計値 S_{t_k} を求めて（式 (4)），この値の降順で一定数（k 個）を抽出し，それを用語候補とする．

$$S_{t_k} = \sum_{d=1}^{n} q_{t_k,d} \tag{4}$$

4. 実験

4.1 実験方法

3 文からなる地名とカタカナ語のそれぞれ 25 個の説明文を検索クエリとして入力し，上位 1000 位以内の正解出現率，用語候補の中から正解となる単語が上位何番目に出力されたか, および（式 (5) で定義される平均逆順位（MRR：Mean Reciprocal Rank）を用いて評価を行う）．

$$MRR = \frac{1}{Q_N} \sum_{q=1}^{Q_N} \frac{1}{tRank_q} \tag{5}$$

$tRank_q$ は検索クエリ q に対して，正解となる答えが用語候補として出力されたときの順位であり，Q_N は検索クエリの個数である．1000 件以内に見つからなかったときは

表 1 実験結果

地名	出力順位	カタカナ語	出力順位
アメリカ	71	ユーザー	844
東京	24	コーパス	17
中国	15	キーワード	7
イギリス	20	カリキュラム	14
京都	229	アルゴリズム	42
ドイツ	9	ノード	92
千葉	30	コスト	87
広島	72	サンプル	796
スペイン	670	ビット	837
カナダ	*	ターゲット	885
群馬	8	プライド	*
八王子	*	コミュニケーション	11
エジプト	636	スピーカー	28
シドニー	8	マラソン	11
メキシコ	13	アーティスト	*
名古屋	54	パスポート	*
ラスベガス	15	サリン	8
成田	124	スターバックス	*
吉祥寺	38	デシベル	*
シンガポール	*	オリーブオイル	*
静岡	85	レントゲン	335
イラン	378	バイオリン	93
モンゴル	*	プリンター	60
高崎	11	コイル	*
熱海	466	マラリア	*
正解出現率	84 % (21/25)	正解出現率	68 % (17/25)
MRR	0.036	MRR	0.028

＊：順位が 1000 位以内に見つからなかった

$\frac{1}{tRank_q} = 0$ として計算する．

本実験では文書選択の際に上位何件のパッセージをとるかの n = 100，各文書から関連語をいくつとるかのパラメータ m = 100，用語候補の出力する数のパラメータ k = 1000 をとして，各検索クエリに対して用語候補を 1000 個出力して実験を行った．

4.2 実験結果

表 1 に地名とカタカナ語の説明文を検索クエリとした実験結果を示す．地名の解出現率は 84 %，MRR は 0.036，カタカナ語の正解出現率は 68 %，MRR は 0.028 であった．50 個の検索クエリのうち 38 個は 1000 件以内に見つかっている（Recall = 78 %）. 見つかった順位も高いとは言えず，実際に上位 10 位以内に出力された数は，地名が 3 個，カタカナ語が 2 個であった．このことから，正解出現率（Recall）と順位を向上させる必要があることがわかる．

5. 文書選択システムの改良による用語検索の性能改善

用語検索システムの精度の向上のためには，初めの文書選択の精度向上は重要である．本手法は説明文と内容が合致するパッセージに用語が含まれていると仮定するものであるので，初めの文書（パッセージ）の選択を誤ると適切な語が取り出せないためである．このため，文書（パッセージ）選択の精度向上を行った．具体的には，文書選択時にパッセージ類似度だけでなく，広域文書類似度を用いる手法 [10] を加えた．

表 2，表 3 に地名とカタカナ語を検索クエリとした場合の文書選択改良前後の結果を示す．正解出現率は地名の場合 88 %に向上した．カタカナ語は変わらなかった．また，上位 10 位以内に出力された数は，地名が 5 個，カタカナ語が 4 個であり，共に増えていることが確認できた．

検索クエリによっては，順位は上がっているものも下がっているものもあるため，平均的な検索性能を表す評価尺度 MRR を用いて評価を行った．地名とカタカナ語のそれぞれ 25 個の検索クエリ，どちらも MRR は向上しており，地名が 0.036 から 0.058，カタカナ語が 0.028 から 0.030

表 2 地名の文書選択改良前後の結果

地名	文書選択改良前	文書選択改良後
アメリカ	71	55
東京	24	122
中国	15	7
イギリス	20	25
京都	229	260
ドイツ	9	11
千葉	30	27
広島	72	10
スペイン	670	5
カナダ	*	488
群馬	8	16
八王子	*	488
エジプト	636	13
シドニー	8	6
メキシコ	13	19
名古屋	54	74
ラスベガス	15	6
成田	124	15
吉祥寺	38	39
シンガポール	*	128
静岡	85	*
イラン	378	*
モンゴル	*	*
高崎	11	8
熱海	466	29
正解出現率	84 % (21/25)	88 % (22/25)
MRR	0.036	0.058

＊：順位が 1000 位以内に見つからなかった

表 3 カタカナ語の文書選択改良前後の結果

カタカナ語	文書選択改良前	文書選択改良後
ユーザー	844	164
コーパス	17	174
キーワード	7	9
カリキュラム	14	6
アルゴリズム	42	31
ノード	92	517
コスト	87	267
サンプル	796	*
ビット	837	808
ターゲット	885	843
プライド	*	257
コミュニケーション	11	16
スピーカー	28	116
マラソン	11	9
アーティスト	*	*
パスポート	*	*
サリン	8	6
スターバックス	*	*
デシベル	*	257
オリーブオイル	*	*
レントゲン	335	*
バイオリン	93	36
プリンター	60	96
コイル	*	*
マラリア	*	52
正解出現率	68 % (17/25)	68 % (17/25)
MRR	0.028	0.030

＊：順位が 1000 位内に見つからなかった

になり，精度が高くなっていることがわかった．

今回，用語候補を上位 1000 件まで出力したところ，地名とカタカナ語それぞれについて求める正しい答えを 21 件，17 件見つけることができた．しかし，候補として出力された順位が低く実用性はまだ低い．文書選択を改良すると文書選択の段階で，よりよいパッセージを見つけられることができ，用語検索の精度は良くなっていることが確認できた．しかし，用語候補の順位としては十分ではなく改善する余地が大きい．

6. 結論

非構造化文書を検索対象として用語検索をした．説明文に類似したパッセージを見つけ，そこから関連語を選択することで用語を見つける方法を検討した．初期のパッセージ選択の精度向上が重要であることおよび，まだ十分な用語検索精度が得られないことがわかり，改善の余地が大きいことを確認した．文書選択の改良も必要であるが，現時点では順位は低いものの正しい解答（用語）はある程度見つかっているため今後は用語候補の出力段階で求める用語を上位にする方法についても研究していく予定である．

謝辞

本研究は科研費（課題番号 25330368）の助成を受けた．文書選択システムの構築には GETA[11] を使用した．

参考文献

[1] 粟飯原俊介, 長尾真, 田中久美子. : ”意味的逆引き辞書『真言』”, 言語処理学会第 19 回年次大会 発表論文集, pp.406–409, 2013.
[2] 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，北海道大学，pp. 547–550，2014 年 3 月
[3] Qing Ma, Ibuki Tanigawa, and Masaki Murata:” Retrieval Term Prediction Using Deep Belief Networks”, The 28th Pacific Asia Conference on Language, Information and Computing (Paclic 28), pp. 338–347, Phuket, Thailand, December 12-14, 2014.
[4] 谷河息吹, 馬青, 村田真樹：”検索語の予測における DeepLearning 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，京都大学，pp. 684–687，2015 年 3 月
[5] Tomoyosi Akiba and Kiyoaki Aikawa and Yoshiaki Itoh and Tatsuya Kawahara and Hiroaki Nanjo and Hiromitsu Nishizaki and Norihito Yasuda and Yoichi Yamashita and Katunobu Itou：”Construction of a test collection for spoken document retrieval from lecture audio data”,IPSJ-Journal,vol50,No2,pp.501–513,2009.
[6] 前川喜久雄：”言語研究における自発音声”， 日本音響学会研究発表会講演論文集 (春季) , pp.19-22, 2001.
[7] 西尾友宏，南條浩輝，吉見毅彦：”講演音声ドキュメント検索のための擬似適合 性フィードバック”，情報処理学会論文誌，Vol.55, No.5, pp.1573–1584, 2014.
[8] 北　研二，津田和彦，獅々堀正幹：”情報検索アルゴリズム”，共立出版株式会社，ISBN4-320-12036-1 (2002).
[9] 小作浩美，内山将夫，井佐原均，河野恭之，木戸出正継：”WWW 検索における複数検索結果の結合処理とその評価”，情報処理学会論文誌，Vol. 44, No. SIG 8（TOD 18）,pp. 78-91 (2003).
[10] 南條浩輝，弥永裕介，吉見毅彦：”広域文書類似度と局所文書類似度を用いた講 演音声ドキュメント検索”, 情報処理学会論文誌, Vol.53, No.6, pp.1654–1662, 2012.
[11] ”汎用連想計算エンジン GETA”,http://geta.ex.nii.ac.jp