Web 文書からの用語検索における用語候補のランキングの検討

池内 省吾 [1]　南條 浩輝 [2]　馬 青 [1]
[1] 龍谷大学理工学研究科　[2] 京都大学学術情報メディアセンター
[1]t20m002@mail.ryukoku.ac.jp [2]nanjo@media.kyoto-u.ac.jp [1]qma@math.ryukoku.ac.jp

概要

適切な用語を用いないと正確な問い合わせが行えないことがある．我々は，適切な用語がわからないユーザの支援の研究を行っている．具体的には，ユーザからの用語に関する説明文を検索質問として Web 検索を行って，Web 文書から用語を抽出する「Web 文書からの用語検索」を研究している．本稿では，Web 文書群から得た用語候補のランキングについて述べる．

1　はじめに

用語検索は，用語を説明する説明文から用語を求める処理である．辞書逆引き [1] とは，説明文と近い定義文を見つけてその見出し語を求める処理であり，辞書に掲載されている語に対する用語検索は，辞書逆引きで求めることが可能である．一方，新語や専門用語は辞書形式で定義されていないことがある．そこで，我々はユーザによる説明文を検索質問として一般の Web ページを検索し，そこから用語抽出を行う「Web 文書からの用語検索」の手法を提案している [2]．本稿では，用語検索において出力する複数の用語候補のランキング方法について検討を行った結果を報告する．

2　Web 文書からの用語検索

我々が提案している BERT を用いた Web 文書からの用語検索 [2] について述べる．これは以下の 3 つの処理を順に適用して用語を求めるものである．

1. **Web 文書検索**: 与えられた説明文からなる検索質問で Web 検索を行い，Web 文書集合を取得する
2. **用語候補抽出**: Web 文書集合それぞれの文書について，与えられた検索質問と文書中のテキストを連結して BERT に入力し，質問応答の枠組みを利用して用語候補を抽出する．
3. **用語候補ランキング**: 得られた用語候補をランキングして出力する

2.1　Web 文書検索

本研究では，Web 文書として Google スニペットを用いる．Google スニペットは，Goolge 検索エンジンに検索質問を入力して得られる検索結果ページにおいて，検索結果の Web ページのタイトルや URL の下に表示される短い説明のことである[1)]．Google スニペットの取得は，Google API Client Libraries [3] を用いて行う．

2.2　用語候補抽出

用語候補抽出には BERT を用いた質問応答の枠組みを用いる．具体的には，SQuAD 2.0 [4] で用いられた方法 [5] を用いる．これは，質問 q と文書 d を連結させたもの（ **[CLS]** q **[SEP]** d ）に対して，解答可能な場合には文書中の正解の開始位置と終了位置（Start/End）を出力するように，解答不可能な場合には開始位置と終了位置として [CLS] の位置を出力するように BERT を学習（fine-tuning）するものである．本研究では，BERT の日本語 Pretraining モデルとして，京都大学提供 BERT（BERT-BASE-WWM 版）[2)] を用いる．

2.3　用語候補ランキング

得られた用語候補を何らかの方法でランキング（順位づけ）し，出力する．我々は，以前の研究 [2] において頻度に基づいたランキングを行ったものの，他のランキング手法を検討していなかった．本研究ではこの検討を行う．

1）　文書中の検索質問に関連する箇所をつなげたもので，必ずしも文書中の連続した文字列とはなっていない．
2）　https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese

3 実験のデータ

3.1 用語検索用の検索質問

本論文では，表 A に示す 23 の検索質問を用いる．これは，先行研究 [2] で本タスクのために作成されたものである．

3.2 用語抽出用 BERT の学習データ

用語抽出用 BERT の学習データとして，解答可能性付き読解データセット [6] を使用する．これは，文書読解タスク用のデータセットであり，(質問，解答，文書）の組に対して，解答可能性スコアがつけれらている．解答可能性スコアは，文書読解によって質問に答えることができるかについて人手で判断されたスコアである．このデータセットの（質問 q_i，解答 a_i，文書 d_i^k）の組について，質問 q_i に対する解答 a_i（本研究での用語）は文書 d_i^k に必ず含まれている．一つの質問・解答について複数（k）の文書が用意されている．読解タスクにおいては解答不能文書が定義されているが，これは解答を含まないという意味ではなく，解答根拠が存在しないという意味である．

本研究での用語検索では解答根拠は不要である．さらに，求めた Web 文書に用語が存在しない場合もある．そこで，読解タスクのデータセットそのものを学習に使うのではなく，次のように修正して学習データとする．

1. **解答可能データ集合**: 文書読解タスク用データセットの（q_i，a_i，d_i^k）の組を全て解答可能データ集合とする．a_i の d_i^k における開始位置と終了位置を BERT に学習させる．
2. **解答不能データ集合**: 文書読解タスク用データセットの（q_i，a_i，d_i^k）の組に対して，d_i^k を別の q_j に対応する文書 $d_j^l (j \neq i)$ に置き換えたものを解答不能データ集合とする[3]．開始位置と終了位置を共に [CLS] の位置として BERT に学習させる．

4 用語候補ランキング

Web 文書からの用語検索では，検索質問と各 Web 文書を連結した文字列を BERT に入力して用語候補を求め，その候補をランキングして出力する（図 1）．

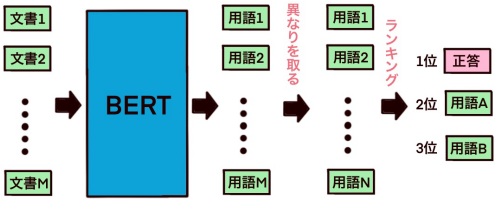

図 1 複数文書からの用語候補の抽出とランキング

最も単純なランキング手法として頻度に基づくもの（多数決）がある．実際に，我々も以前に多数決方法による用語検索を行った [2] ものの，他のランキング手法は十分に検討できていなかった．このような背景に基づき，本研究では多数決方式と他の 4 種類ランキング方法（下記）を検討する．

1. **Maj (Majority):** 多数決（頻度）に基づいてランキングを行う．
2. **T-TSum (Sum of Term-Term similarities):** 用語候補間で類似度をとり，その和に基づいてランキングを行う．これは用語候補の内でもっとも中心となるような用語が目的の用語であることを期待するものである．
3. **Q-T (Qery-Term similarity):** 検索質問（単語列）と用語候補（単語）との類似度に基づいてランキングを行う．正解用語と検索質問は同じ意味を表していると仮定し，両者の類似度を直接求めることで目的の用語が求められると期待するものである．
4. **Q-T$_{sni}$ (Query-Term_snippet similarity):** 検索質問（単語列）と用語候補のスニペット（単語列）との類似度に基づいてランキングを行う．正解用語による検索で得られるスニペットを正解用語の説明とみなし，これが検索質問と近いと期待するものである．
5. **Q$_{sni}$-T$_{sni}$ (Query_snippet-Term_snippet similarity):** 検索質問のスニペット（単語列）と用語候補のスニペット（単語列）との類似度に基づいてランキングを行う．両者ともにスニペットとすることで，正解用語と検索質問間で高い類似度が得られることを期待するものである．

Maj 以外のランキング手法についての概要を図 2 から図 5 に示す．

本研究では，「X のスニペット」として，X を検索クエリとして，Google API Client Libraries を用いて 1 件だけ取得したものを採用した．

本研究では，用語候補（単語）や検索質問（単語

3) 偶然に a_i が含まれていないことも確認する．

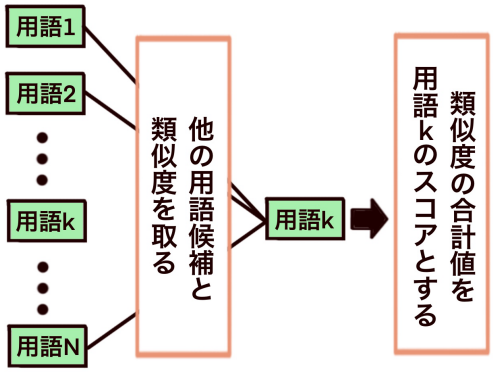

図 2 T-Tsum (Sum of Term-Term similarilites)

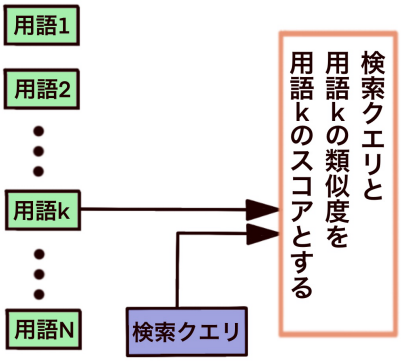

図 3 Q-T (Query-Term similarity)

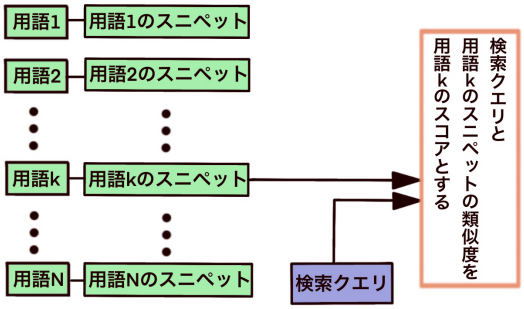

図 4 Q-T$_{sni}$ (Query-Term_snippet similarity)

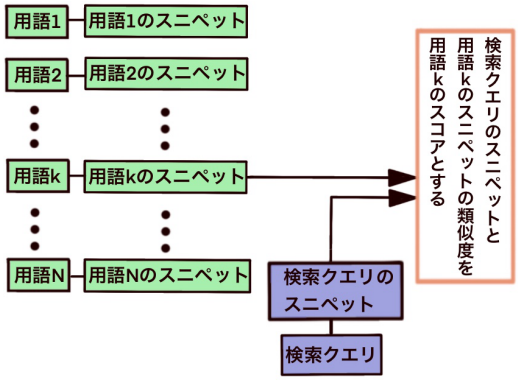

図 5 Q$_{sni}$-T$_{sni}$(Query_snippet-Term_snippet similarity)

列），それらのスニペット（単語列）間の類似度を求めるために，これらを同一次元のベクトル（分散表現）で表現する．分散表現は，用語候補，検索質問，スニペットを同一に扱い（それらを *string* とする），**[CLS]** *string* **[SEP]** を BERT に入力して求める．具体的には，各 token に対応する BERT の出力 embedding の平均値（average pooling）を求める．ベクトル間の類似度はコサイン距離とする．

5 実験結果と考察

23 件の検索質問を用いて用語検索を行った．図 1 に示す通り各質問について M 件（最大 100 件）の Web 文書を検索し，各 Web 文書から 1 件の用語候補を取り出して重なりをとり，用語候補の集合（図 1 の N 件）を得た．この用語候補中に正解の用語を含むものは 17 件，含まないものは 6 件であった．この 6 件に関してはどのようなランキング手法を適用しても正解を上位にランクすることはできない．求める Web 文書数を多くし，各文書から複数の用語候補を出力するなどが必要であることがわかる．

はじめに多数決ランキングを行った．用語候補中に正解を含むもの 17 件のうち 11 件で正解の用語が 1 位であった．残りの 6 件は，3 位，3 位，4 位，5 位，6 位，13 位であった．結果は表 1 の Maj に示されている．23 件に対する平均逆順位（MRR）は 0.54 であった（表 2）．1 位ではない 6 件については，正解の用語に代わって検索質問に関係はするものの別の語が上位に来ていることが多い．例えば，正解用語が「バナナ」である時には，「マンゴー」や「みかん」が上位にランクされていた．4 位以降も「パパイヤ」「ぶどう」「パイナップル」と続き，バナナと同じカテゴリの果物が続く．検索質問が曖昧で，抽出する用語が絞りきれない時には，このようなことになると考えられる．用語「アイシャドー」に対しては，「目元」「アイ」「ニベア」「油分」「アイスクリーム」が上位にランクされており，化粧品とは別カテゴリの食品カテゴリの語が含まれていた．用語「コールドゲーム」に対しては，「野球」「ヤクルト」「3 月 21 日」「甲子園球場」「8 試合」「順延になった」「9 月 20 日」などが上位にランクされており，野球に関する用語が抽出されていると考えられるが，日付などの別のカテゴリと考えられる語も含まれていた．これらに対しては，用語間の類似度や用語と検索質問の類似度を求めることで明らかに異なる候補

表 1 用語ごとの検索結果

用語	Maj	T-Tsum
猫	1	33
バナナ	3	37
ウインドブレイカー	1	20
アイドリング	1	46
海賊版	3	18
カバディ	1	1
サブスク	1	57
GPU	＊	＊
GPS	1	11
GPA	4	14
GDP	＊	＊
GNP	＊	＊
CPU	1	19
鬼滅の刃	1	7
ティンパニー	＊	＊
ドーピング	＊	＊
コールドゲーム	13	7
マンホール	5	11
湯たんぽ	1	25
エキストラ	1	11
アイシャドー	6	5
メトロノーム	＊	＊
夏至	1	2

正解となる用語の出現順位を表す
＊：順位が 100 件以内に正解がなかった

であるもの（食品や日付など）を除くことができる可能性がある．

次に，類似度に基づく 4 種類のランキングを行った．MRR は 0.05〜0.13 と低いものであった．この結果も表 2 に示されている．

T-TSum では，多数決で上位にランクできるものについては低いランクをつけてしまったものの，多数決ランキングで低いランクのものについて，より高いランクに出力できていた．実際に多数決で 6 位のアイシャドーは 5 位に，13 位のコールドゲームについては 7 位にランクすることができた（表 1 T-TSum）．

Q-T では，単語列（検索質問）と単語（用語候補）との間で類似度をとろうとしており，うまく働かなかったものと考えられる．

Q-T$_{sni}$ では，用語候補からスニペット（単語列）を取り出して，単語列同士の類似度を測った．**Q-T** よりも類似度の求め方としては適切と考えられるが，課題が大きいことがわかる．

表 2 各手法による MRR の値

手法	MRR
Maj	0.54
T-Tsum	0.11
Q-T	0.05
Q-T$_{sni}$	0.13
Q$_{sni}$-T$_{sni}$	0.06

Q$_{sni}$**-T**$_{sni}$ では，用語候補と検索質問の双方からスニペット（単語列）を取り出して，単語列同士の類似度を測った．検索質問のスニペットが正解用語のスニペットと類似し，かつそれ以外の用語候補のスニペットと類似しないことを期待したが，そのようなスニペットの取得に課題があったと考える．

6 まとめ

用語検索において出力する複数の用語候補のランキング方法について検討を行った．多数決によるランキングが精度が高かった．類似度に基づく手法それ自体は高精度ではなかったが，多数決手法が苦手とするようなものに対して高いランクに出力できる可能性があることがわかった．今後は多数決手法と類似度に基づく手法の併用による用語検索の高精度化を検討していきたい．

謝辞: 本研究は科研費基盤研究 (C) 19K12241 の補助を受けた．

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] 池内省吾, 南條浩輝, 馬青. BERT を用いた Web 文書からの用語検索. 情報処理学会研究報告 Vol.2021-NL-249, pp. 1 – 6, 2021.
[3] Google. Google API Client Libraries. https://developers.google.com/api-client-library.
[4] Pranav Rajpurkar, Robin Jia, , and Percy Liang. Know What You Don’t Know: Unanswerable Questions for SQuAD. *ACL*, Vol. 2, p. 784–789, 2018.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*, Vol. 1, p. 4171–4186, 2019.
[6] 鈴木正敏, 松田耕史, 岡崎直観, 乾健太郎. 読解による解答可能性を付与した質問応答データセットの構築. 言語処理学会第 24 回年次大会, pp. 702–705, 2018.

A 付録

表 A 作成した検索質問 23 件

用語	検索質問
猫	ペットの動物。部屋で飼うことが多い。かわいい。
バナナ	黄色の果物。細長くて甘い。皮をむいて食べる。
ウインドブレイカー	寒い時に着る服。風を通しにくい。
アイドリング	止まっているのに、エンジンをかけたままのこと。
海賊版	偽物のこと。DVD とかに使われる。
カバディ	インドのスポーツ。鬼ごっこみたいなもの。
サブスク	定期的にお金を払うことで、サービスを使い放題になること。
GPU	パソコンのグラフィックをきれいにする。研究とか仮想通貨にも使われる。
GPS	人工衛星を使って自分の位置を知る。
GPA	学校などの成績の一種。ポイントの平均点みたいなもの。
GDP	国の中で作られた価値。外国が国内で作ったものも含む。
GNP	国として作った価値。外国で作ったものも含む。
CPU	パソコンの計算部分。速さとコア数で計算速度が変わる。
鬼滅の刃	鬼を倒す漫画。映画の中で一番お金を稼いだ。
ティンパニー	大きめのたたく楽器。置いて使う。
ドーピング	薬を飲んで能力をあげること。悪い意味で使う。
コールドゲーム	雨で中止になった試合。
マンホール	道路にある穴のふた。
湯たんぽ	寒い時にお湯を入れて使う道具。
エキストラ	テレビや映画に出る人で、その他一般の人。
アイシャドー	目の上に塗る化粧品。青が多い気がする。
メトロノーム	音楽の時にリズムをとる装置。揺れて音が鳴る。
夏至	一年の内で一番昼が長くなる日。