2V-8

複合名詞の構造パターンを考慮したスコアリング手法の改良

吉野 徹 † 福本 淳一 ‡

立命館大学大学院理工学研究科 † 立命館大学情報理工学部メディア情報学科 ‡

1 はじめに

質問応答 (Question Answering : 以下，QA) システムは，自然文による質問文に対して，Web ページ等の大量の検索文書を知識源とし，質問文に対する回答を提示するシステムである．

QA システムでは，検索文書から得られた各回答候補に対して，質問文に対する回答としてふさわしいかどうかの採点を行う (以下，スコアリング)．スコアリングでよく用いられるのは,「質問文から得られたキーワード」(以下，質問文キーワード) が検索文書に書かれた場合,「回答候補を採点するキーワード」(以下，スコアリング対象) として扱い，回答候補と「スコアリング対象」との距離で，採点する方法である．[1][2]

問題点として，QA システム RitsQA のスコアリング手法では，検索文書の「質問文キーワード」を全て「スコアリング対象」としていることで，質問文で用いられているものと「同一」,「違う」に関わらず,「スコアリング対象」として扱われていることである．

質問文の複合名詞が，検索文書では，複合名詞で記述されない時や間に助詞を含んで書かれる時がある．質問文に「日本経済」が書かれ，検索文書に「日本の経済において (省略) 非常に日本と経済構造が似ている (省略) ある種の経済的なゆがみ (省略) 日本の経済史」が書かれている場合を例にして述べる.【経済的なゆがみ】の「経済」は，複合名詞で記述されていなく,「日本経済」とは「違う」ものである．質問文の「日本経済」の間に助詞を含んで書かれているのは「日本の経済」と「日本と経済」の場合である.【日本の経済において】の「日本の経済」は,「日本経済」と「同一」である．しかし,「日本と経済」と【日本の経済史】の「日本の経済」は「違う」ものである．

質問文の複合名詞が，検索文書で複合名詞で記述されなかった時,「スコアリング対象」と扱うべきではない．また，質問文の複合名詞が，検索文書で間に助詞を含んで書かれた時,「違う」場合は,「スコアリング対象」と扱うべきではない.「同一」の場合は，間の助詞を無視して回答候補までの距離を計算すべきである．

本研究では，複合名詞の構造パターンを考慮したスコアリング手法の改良について述べる．

Toru YOSHINO† and Junichi FUKUMOTO‡
Graduate School of Science and Engineering, Ritsumeikan University†
Department of Media Technologies, Ritsumeikan University‡

2 複合名詞の違う書かれ方の分析

新聞記事を用いて，複合名詞と複合名詞の間に助詞を含んでいる文字列が，どのような助詞の場合，また，どのような品詞列の時に「同一」,「違う」かの分析について記載する．

分析データは 2004 年の毎日新聞記事データから，AxB の単語列 (AB は接尾，非自立語を除く名詞，x は助詞) のパターンを抽出．さらに，同様の新聞記事データから AB の並びが品詞情報まで一致したものを用いた．

分析方法は AxB と AB が,「同一」,「違う」になっているかを人手で分析した．

分析した結果，AxB と AB が「違う」になる x の種類は 206 種類あることが分かった．一部を以下に示す．

- より (助詞-格助詞-一般) も (助詞-係助詞)
- だけ (助詞-副助詞) でも (助詞-副助詞)

x が「より (助詞-格助詞-一般) も (助詞-係助詞)」を例にして述べる．AB が「ゲームキャラクター」と AxB が「ゲームよりもキャラクター」では「違う」ものになる．

同じ x でも A の品詞，B の品詞の並びによって，AB と AxB が「同一」,「違う」に変わる x の種類が 44 種類あることが分かった．一部を表 1 に示す．表 1 では，x が「を (助詞-格助詞-一般)」の時，A の文字列と品詞が「CD(名詞-一般)」，B の文字列と品詞が「棚 (名詞-一般)」では，AB(CD 棚) と AxB(CD を棚) は「違う」である．

A の文字列と品詞が「表彰台 (名詞-一般)」，B の文字列と品詞が「独占 (名詞-サ変接続)」の時，AB(表彰台独占) と AxB(表彰台を独占) は「同一」である．

3 スコアリングの範囲の分析

複合名詞の間に助詞を含んだ文字列 AxB(AB は自立語，x は助詞) が,「A の前の語」と「B の後ろの語」の品詞によって，複合名詞 AB と「同一」になるかどうかを分析した．

分析データとして，表 1 の分析結果の x の文字列と品詞が「の (助詞-連体化)」の時，AB と AxB が「同一」と判断した AxB(例 : 和風のスープ) が，2004 年 1 月の毎日新聞記事データで AxB と一致した 12912 文 (例 : とても美味しい和風のスープです．) を抽出した．

表 1: A と B の品詞によって AB と AxB が「同一」,「違う」に変わる助詞 x(一部抜粋)

x の文字列と品詞	A の品詞	B の品詞	AB と AxB の関係	AxB の例
の (助詞-連体化)	名詞-固有名詞-地域-国	名詞-一般	同一	日本の経済
を (助詞-格助詞-一般)	名詞-一般	名詞-一般	違う	CD を棚
を (助詞-格助詞-一般)	名詞-一般	名詞-サ変接続	同一	表彰台を独占
と (助詞-並立助詞)	名詞-固有名詞-地域-国	名詞-一般	違う	日本と経済

分析手法は「抽出した文中の AxB」と「AxB」が「同一」かどうか人手で分析した．分析した結果,「A の前の語」または「B の後ろの語」の品詞が「名詞」の時,「抽出した 1 文中の AxB」(例 : BSE の検査対象となる．) と「AxB」(例 : BSE の検査) は「違う」である.「A の前の語」の品詞が,「接頭詞」の時,「抽出した 1 文中の AxB」(例 : 準々決勝の組み合わせの抽選) と「AxB」(例 : 決勝の組み合わせ) は「違う」である.

4 改良スコアリング手法

質問文の複合名詞 AB(AB は自立語) が検索文書で,複合名詞で記述されなかった時,「スコアリング対象」としない.

また，検索文書で複合名詞中に助詞を含んだ文字列 AxB(x は助詞) が表れた時,「同一」の場合，間の助詞を無視し，回答候補までの距離を計算する.「違う」の場合,「スコアリング対象」としない．複合名詞と複合名詞の間に助詞を含んだ単語列が「同一」,「違う」の判別には，2 章の分析結果を用いる．また AB と AxB が「同一」の場合でも「A の前の語」または「B の後ろの語」の品詞が「名詞」,「A の前の語」の品詞が「接頭詞」の時,「スコアリング対象」としない.

質問文に「日本経済」が書かれ，検索文書に「日本の経済において (省略) 非常に日本と経済構造が似ている (省略) ある種の経済的なゆがみ (省略) 日本の経済史」が書かれている場合を例にして述べる.

【ある種の経済的なゆがみ】の「経済」は,「スコアリング対象」としない.【日本の経済において】の「日本の経済」は，表 1 より,「日本経済」と「同一」と判断し，間の助詞を無視し，回答候補までの距離を計算する.「日本と経済」は表 1 より,「日本経済」と「違う」と判断し,「スコアリング対象」としない.【日本の経済史】の「日本の経済」は，表 1 より,「日本経済」と「同一」と判断されるが,「日本の経済」の後ろの語が「史 (名詞)」のため,「スコアリング対象」としない.

5 実験・評価

実験では NTCIR-3 の QAC タスクで使用された複合名詞を含む 10 の質問文毎に，Google によって上位 10 記事の検索記事を取得した．RitsQA で得られた回答と改良したスコアリング手法で得られた回答から，正解の回答の順位とスコアがどのように変動したかを比較した.

記事毎に RitsQA の回答と改良したスコアリング手法で得られた回答を比較した結果を表 2 に示す.

表 2: 記事毎の回答の比較結果

比較結果	記事数
正解の回答の順位上昇	20
正解の回答の順位変動なし&スコア増加	4
正解の回答の順位変動なし&スコア減少	1
正解の回答の順位下降	8
正解の回答の順位，スコア変動なし	67

表 2 では,「正解の回答の順位変動なし&スコア増加」の記事が 4 件に対し,「正解の回答の順位変動なし&スコア減少」の記事が 1 件で同じ件数である.「正解の回答の順位上昇」の記事が 20 件に対し「正解の回答の順位下降」の記事が 8 件のため，改良したスコアリング手法の有効性が認められる.

6 おわりに

本研究では，複合名詞の構造パターンを考慮したスコアリング手法の改良について述べた．改良したスコアリング手法の方がより正解の回答を得やすいことが分かった.

改良したスコアリングで正解の回答の順位やスコアが減少した記事もあるため，さらに改良が必要である.

参考文献

[1] 奥村学，礎崎秀樹，東中竜一郎，永田昌明，加藤恒昭 : “質問応答システム (自然言語処理シリーズ 2)”, コロナ社,2009

[2] Clarke, C.L.A., Cormack, G.V. and Lynam, T.R. : “Exploiting Redundancy in Question Answering”, In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.358-365(2001)