

文法構造を付加したテキストに対する情報検索方法の検討

鈴木 晋†

愛知工業大学情報科学部†

1. はじめに

テキストを検索する方法として、キーワード検索やタグを用いた検索方法が普及している。本稿では、より高度な検索を目指して[1, 2, 3], 日本語テキストに簡単な文法構造を手で付加し、この文法構造を用いてテキストを検索する方法について検討する。

2. 説明文と質問文

日本語の説明文と質問文の例を次に示す。これらの自然な日本語文を原文と呼ぶ。

(1) 説明文 (原文)

S1: 太郎は毎年、正月に、自分でとった写真を友達に送っている。

S2: 一昨年の写真は電車の中で漫画を読んでいる大人の写真であった。

S3: 昨年の写真はサラリーマンについてであった。

(2) 質問文 (原文)

Q1: 太郎は正月に友達に写真を送りますか？

Q2: 太郎は写真を誰に送りますか？

Q3: 太郎は漫画を読んでいる大人の写真を友達に送りましたか？

3. 原文の形式的表現

計算機が2の自然な日本語文(原文)を読んでその内容を理解し、質問に答えるのはまだ難しいように思われる。そこで、本稿では、原文の簡単な文法構造(形式文と呼ぶ)を手で作成し、計算機がそれを使って質問に答える方法を検討する。2の原文を例に形式文の作り方を説明する。

(1) 簡単な単文への分解

原文は複文であることが多い。また、その他にも複雑、繊細な表現(時制、様相、量子等)をもつことが多い。これらを計算機で処理するのは難しいので、原文を複数の簡単な単文に分解する。たとえば、S1を次の単文A1-1とA1-2に分解する。

A1-1: 太郎は写真を撮る。

A1-2: 太郎は正月に写真を友達に送る。

ここで、A-1とA-2は共通な名詞「写真」を含ん

でおり、これらは集合「写真」の中のある同じ要素を表すと考える。すなわち、A-1とA-2は次を表すと考える。

$\exists x \in \text{集合「写真」}$,

太郎は x を撮る。

太郎は正月に x を友達に送る。

一般に、原文を簡単な単文の集合に分解すると原文の意味の一部が失われるので、単文集合は原文の近似になる。

原文 S2 と S3 のための単文を次に示す。

A2-1: 写真は一昨年のものである。

A2-2: 写真は大人のものである。

A2-3: 大人が電車の中で漫画を読む。

A3-1: 写真はサラリーマンについてである。

A3-2: 写真は昨年のものである。

(2) 単文の形式的表現 (形式文)

計算機の処理をさらに容易にするために、各単文から組(品詞, 主要語, 付属語)の集合{(品詞, 主要語, 付属語), ...}を作成する。この集合を形式文と呼ぶ。たとえば単文 A1-1 に対して次の形式文 B1-1 を作成する。

B1-1: {(主語, 太郎, は), (動詞, 撮る), (目的語, 写真, を)}

品詞を次のように略記する。

主: 主語, 動: 動詞, 目: 目的語,

直目: 直接目的語, 間目: 間接目的語,

補: 補語, 他: その他

説明文 S1, S2, S3 (単文 A1-1, ..., A3-2) のための形式文を次に示す。

B1-1: {(主, 太郎, は), (動, 撮る), (目, 写真, を)}

B1-2: {(主, 太郎, は), (動, 送る), (直目, 写真, を), (間目, 友達, に), (他, 正月, に)}

B2-1: {(主, 写真, は), (動, ものである), (補, 一昨年, の)}

B2-2: {(主, 写真, は), (動, ものである), (補, 大人, の)}

B2-3: {(主, 大人, が), (動, 読む), (目, 漫画, を), (他, 電車, の中で)}

B3-1: {(主, 写真, は), (動, ものである), (補, サラリーマン, についての)}

B3-2: {(主, 写真, は), (動, ものである), (補, 昨年, の)}

テキストデータベース作成者は各原文に対して形式文を作成し、それらを原文に付加する。

Text retrieval by using manually attached grammatical structures

† Susumu SUZUKI, Faculty of Information Science, Aichi Institute of Technology

4. 質問の処理

データベース検索者は、各質問文について、それを表す形式文を作成し計算機に入力する。計算機は質問文を表す形式文（形式質問文と呼ぶ）と説明文を表す形式文（形式説明文と呼ぶ）を照合することで、質問に答える。以下に、2の質問文 Q1, Q2, Q3 について説明する。

(1) 質問文 Q1 について

質問文が正しいか否か（質問文の内容が説明文の中に記述されているか否か）を回答する質問である。検索者は質問文 Q1 を表す次の形式文 C1 を作成して計算機に入力する。

C1: {(主, 太郎, _), (動, 送る), (間目, 友達, _), (直目, 写真, _), (他, 正月, _)}

ここで、_は don't care を表す。

C1 の中のより多くの組(品詞, 主要語, 付属語)が B_{i-j} に含まれるほど、C1 と B_{i-j} の一致度が高いと見なす。この例では、C1 の中の全ての組が B_{1-2} に含まれているので、C1 は B_{1-2} と完全に一致している。計算機は形式説明文の中から、C1 とよく一致する B_{i-j} (B_{1-2}) を探し、 B_{i-j} および B_{i-j} を含む原文 (S_1) を利用者に返す。C1 と B_{i-j} が完全に一致するとは限らないし、形式文は原文の近似であるので、どのような質問においても、最終的には、検索者が計算機が返す原文 (S_1) をみて、質問文が正しいか判断する。

(2) 質問文 Q2 について

質問文の中の変数に適切な語句を代入するとき、質問文が正しくなるような語句を回答する質問である。検索者は質問文 Q2 を表す形式文 C2 を作成して入力する。

C2: {(主, 太郎, _), (動, 送る), (間目, X?, _), (直目, 写真, _)}

変数 X?に「友達」を代入すると、C2 と B_{1-2} が一致するので、計算機は $X=友達$ と、 B_{1-2} を含む原文 S_1 を利用者に返す。

(3) 名詞の同一性

3に述べたように、同じ原文から作成された各形式文中の同じ名詞 N (たとえば、 B_{1-1} と B_{1-2} の写真) は、その N が表す集合の中のある同じ要素を表すと考える。しかし、異なる原文から作成された形式文の中の同じ名詞 N' については同じ要素をとることができるとは限らない。たとえば、 B_{1-1} , B_{1-2} の中の写真と B_{2-1} , B_{2-2} の中の写真は同じ要素（「太郎が正月に友達に送っている、自分で撮ったある写真で、かつ、電車の中で漫画を読んでいる大人の、一昨年のある写真」）をとることができるが、一方、 B_{2-1} , B_{2-2} の中の写真と B_{3-1} , B_{3-2} の中の写真は、一昨年と昨年が矛盾するので、同じ要素をとるこ

とができない。

(4) 質問文 Q3 について

検索者は質問文 Q3 を表す形式文 C_{3-1} , C_{3-2} , C_{3-3} を作成して入力する。

C_{3-1} : {(主, 太郎, _), (動, 送る), (直目, 写真, _), (間目, 友達, _)} (太郎は写真を友達に送る)

C_{3-2} : {(主, 写真, _), (動詞, ものである), (補, 大人, _)} (写真は大人のものである)

C_{3-3} : {(主, 大人, _), (動, 読む), (目, 漫画, _)} (大人が漫画を読む)

質問文 Q3 は複数の形式質問文からなる。そのため、計算機は各形式質問文 C_{3-j} に対して、それと一致度が高い形式説明文を探す。同じ原文から作成した形式説明文の中の名詞は同じ要素を表すので、計算機は初め、 $\{C_{3-1}, C_{3-2}, C_{3-3}\}$ と各原文 S_i のための形式説明文集合 $\{B_{i-1}, B_{i-2}, \dots\}$ を比較して、各 C_{3-j} が同じ S_i から作られたある B_{i-k} に一致するか調べる。この場合、 C_{3-2} は B_{2-2} と、 C_{3-3} は B_{2-3} と一致するが、しかし、 C_{3-1} が B_{2-1} とほとんど一致しないので、そのような形式説明文集合 $\{B_{i-1}, B_{i-2}, \dots\}$ はない。計算機はその旨を検索者に回答する。

上に述べたように、異なる原文から作られた形式説明文の中の名詞であっても同じ要素をとることができる場合もあるので、次に、計算機は検索者からの指示があれば、各形式説明文の元になっている原文が同じであるか気にすることなく、すべての形式説明文 $B_{1-1}, B_{1-2}, \dots, B_{3-1}, B_{3-2}$ の中から $C_{3-1}, C_{3-2}, C_{3-3}$ とよく一致するものを探す。この場合、 $C_{3-1}, C_{3-2}, C_{3-3}$ は各々、 $B_{1-2}, B_{2-2}, B_{2-3}$ に一致するので、計算機はその旨と、 $B_{1-2}, B_{2-2}, B_{2-3}$ を含む原文 S_1, S_2 を検索者に返す。質問が正しいか否かは、検索者が S_1 と S_2 をみて判断する。

5. おわりに

人手で作成した文法構造（形式文）を利用して日本語テキストを検索する方法を検討した。今後、文法構造の作成の容易さ、計算効率を考慮しながら、原文のもつ情報の損失を少なくするように改良したい。

参考文献

- [1] オールドウド, アンデルソン, ダール 著, 公平珠躬, 野家啓一 訳, 日常言語の論理学, 産業図書, 1979年.
- [2] アントニウ, ハルメレン 著, CD-ROM で始めるセマンティック Web, ジャストシステム, 2005年.
- [3] 磯崎秀樹, 東中竜一郎, 永田昌明, 加藤恒昭 著, 質問応答システム, コロナ社, 2009年.