3C-7

文章構造を付加した日本語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

テキストの検索法として，キーワードやタグを用いた検索法が普及しており[1]，より高度な検索法として，文脈情報を用いた検索も研究されている[2]．筆者は，先に文献[3,4]において，日本語テキストに簡単な文法構造を人手で付加し，これを用いてテキストを検索する方法を提案した．本稿では，この検索法を拡張し，人手で付加した文脈情報を検索に利用する方法について検討する．

２．説明文と質問文

説明文，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，自分の写真を友達に送っている．

S2：昨年の写真は病院に入院している写真であった．

S3：太郎は足にギプスをし，元気がなかった．

S4：今年の写真は市民マラソンに参加した写真であった．

S5：太郎は元気になった．

（２）質問文（原文）

Q1：太郎は元気か？

回答：太郎は昨年は元気がなかったが，今年は元気だ．

Q2：太郎はギブスをしている写真を友達に送ったか？　回答：はい

３．文脈情報の検索への利用

人は，文章中の各文の意味を理解する際に，今読んでいる１文に記述されている情報だけでなく，それ以前の文に書かれた情報および常識を用いて，各文の意味を理解している[2]．本稿では，各文の意味の理解に必要となる，他の文に記された情報を文脈情報と呼び，文脈情報の検索への利用について考える．

例として，２の文 S3 と S5 を考える．S3 には「太郎が元気でない」ことが，S5 には「太郎が元気である」ことが書かれており，これらは一見矛盾に見える．しかし，S2 の情報「昨年の」から S3 は昨年の太郎についての，一方，S4 の情報「今年の」から S5 は今年の太郎についての記述と推測され，S3 と S5 が矛盾なく正しく理解される．

計算機によって文脈情報を処理するために，本稿では予め，次の情報を人手で作成して説明文に付加する．

・関連性が高い複数の文を集めてブロック（段落）とする．２では，block1={S1}，block2={S2,S3}，block3={S4,S5}．

・異なるブロックの間の関係を記述する．２では，block2（block3）が block1 の「例/詳細化」であることを記す（block1→例/詳細化→block2，block1→例/詳細化→block3）．

４．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，本検索法では，原文の簡単な文法構造（形式文と呼ぶ）を人手で作成し付加し，計算機はそれを使って質問に答える[3,4]．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，複雑，繊細な表現（時制，様相，量化子等）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 は次の２つの単文に分解する．

A1-1：太郎は毎年，写真を友達に送る．

A1-2：写真は太郎のものである．

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．

（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．たとえば上の単文 A1-1 に対して次の形式文 B1-1 を作成する．

B1-1：{(主,太郎,は),(動,送る),(直目,写真,を),(間目,友達,に),(補語[when],毎年,)}

ここで，品詞を次のように略記し，補語には 5W1H の区別も記した．

主:主語，動:動詞，目:目的語，直(間)目:直接(間接)目的語，補[5W1H]:補語[5W1H]．

説明文 S1,…,S5 のための形式文（B1-1 を除く）を次に示す．

Japanese text retrieval by using manually attached text structure

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

B1-2: {(主,写真,は),(動,ものである),(補[who],太郎,の)}
B2-1: {(主,写真,は),(動,ものである),(補[when],昨年,の)}
B2-2: {(主,写真,は),(動,ものである),(補[who],太郎,の)}
B2-3: {(主,太郎,は),(動,入院する),(目,病院,に)}
B3-1: {(主,太郎,は),(動,ギプスをする),(補[where],足,に)}
B3-2: {(主,太郎,は),(否・動,である),(補[how],元気)}
B4-1: {(主,写真,は),(動,ものである),(補[when],今年,の)}
B4-2: {(主,写真,は),(動,ものである),(補[who],太郎,の)}
B4-3: {(主,太郎,は),(動,参加する),(目,市民マラソン,に)}
B5-1: {(主,太郎,は),(動,である),(補[how],元気)}

B3-2 の(否・動,である)は動詞「でない」を表す.

５．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．計算機は質問文のための形式文と説明文のための形式文を照合することで質問に答える．

（１）質問文 Q1 の処理

検索者は質問文 Q1 を表す形式文 C1 を作成して計算機に入力する．

C1: {(主,太郎,は),(動,である),(補[how],元気)}

C1 の中のより多くの組(品詞,主要語,付属語)が Bi-j に含まれるほど，C1 と Bi-j の一致度が高いと見なす．この質問文では，B3-2 が否定の形で，B5-1 が肯定の形で C1 と一致する．これは矛盾であるので，計算機は矛盾の解消を試みる．

・文の「補語」は同じブロック内にある他の文を修飾できる可能性が高い

と仮定する．B3-2 を修飾できる可能性がある補語を B3-2 を含むブロック block2={B2-1,…, B3-2} の中から探す．同様に，B5-1 のための補語を B5-1 を含むブロック block3= {B4-1,…B5-1} の中から探す．それらの中から，種類（5W1H）が同じで値が異なる補語を探し，(補[when],昨年,の)を B3-2 のための，(補[when],今年,の)を B5-1 のための修飾と考え，それらを B3-2，B5-1 に付加した形式文を作成する．

B3-2': {(主,太郎,は),(否・動,である),(補[how],元気),(補[when],昨年,の)}
B5-1': {(主,太郎,は),(動,である),(補[how],元気),(補[when],今年,の)}

計算機は B3-2'，B5-1'，それらの元になった形式文 B2-1，B3-2，B4-1，B5-1，原文 S2，S3,S4,S5 を検索者に返す．検索者はそれらを見て回答を判断する．

（２）質問文 Q2 の処理

検索者は質問文 Q2 を表す形式文 C2-1,C2-2,C2-3 を作成して計算機に入力する．

C2-1 : {(主,太郎,は),(動,送る),(直目,写真,を),(間目,友達,に)}
C2-2: {(主,写真,は),(動,ものである),(補[who],太郎,の)}
C2-3: {(主,太郎,は),(動,ギプスをする)}

ここで，C2-1 と C2-2 の普通名詞「写真」は同じインスタンスである必要がある．

計算機は，C2-1，C2-2，C2-3 の各々とよく一致する形式文として，B1-1，B2-2，B3-1 を見つける．ここで，C2-2 との一致は，B4-2 より，B3-1 と同じブロックにある B2-2 を優先した．B1-1，B2-2 の原文は S1，S2 であって，異なる．一般に，原文が異なる場合，それらに現れる普通名詞は同じインスタンスを指すと限らないので，計算機は写真が同じインズタンスか否か調べる．普通名詞が指すインスタンスの同一性について次を仮定する．

・同じブロック内に現れる普通名詞は同じインスタンスを表す可能性が高い．
・特別な関係にある複数のブロックに現れる普通名詞は同じインスタンスを表す可能性が高い．

計算機は，上の２番目の仮定より，block1→詳細・例→block2 である block1 と block2 の写真，すなわち B1-1 と B2-2 の写真が同じインスタンスである可能性が高いと判断する．その結果，計算機は C2-1，C2-2，C2-3 の全てが満たされる可能性が高いこと，および，その元になった形式文 B1-1，B2-2，B3-1，原文 S1，S2，S3 を検索者に返す．

参考文献

[1] 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.
[2] 言語処理学会 編，デジタル言語処理学事典，共立出版，2010 年.
[3] 鈴木晋，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，情報処理学会第 74 回全国大会講演論文集，pp.1-535―536，2012 年 3 月.
[4] 鈴木晋，指示詞を考慮したテキスト検索方法の検討，愛知工業大学総合技術研究所研究報告 第 14 号，pp.99--104，2012 年 9 月.