

質問応答に必要な言い換えの分析*

高橋 哲朗[†] 乾 健太郎[†] 関根 聡[‡] 松本 裕治[†]

[†] 奈良先端科学技術大学院大学情報科学研究科

{tetsu-ta, inui, matsu}@is.aist-nara.ac.jp

[‡] ニューヨーク大学

sekine@cs.nyu.edu

1 はじめに

TREC [12] や QAC [3] といった評価型ワークショップを背景に、近年自然文からの質問応答の研究が盛んに行なわれている。そこでの一般的な手法は、質問文とその解答が含まれると思われる文書間の類似度を計算し、推測した解答クラスと同じクラスでかつ類似度の高い語を抜き出すというものである。したがってこのタスクでは質問文と文書とをどのように照合するかが一つの焦点となる。

文書間の照合のためには、質問文中のキーワードとの近接性を用いる手法やパターンを用いる手法、係り受け情報を用いる手法などが提案されている。照合の制約を弱くするとノイズが多くなり、逆に強くすると表現の多様性のために多くの事例で照合に失敗するため、照合の適切なレベルを見付けるのは難しい。表層の情報だけではなく、そこから意味構造への変換を行ない表現の差異を吸収する方法も提案されている。しかしどのような意味構造を設計するか、またどのように意味構造へ変換するかなど、それ自体が大きな問題となる。

表現の差異を吸収する方法の一つに言い換えがある。言い換への適用と照合の粒度とは直交しており、質問文または文書を正しく言い換えることができれば、どのような照合方法においても柔軟にかつノイズを増やすことなく照合が可能になる。

言い換えを用いた質問応答の研究はこれまでも行なわれており [1, 4, 11]、言い換への適用による質問応答の精度の向上が報告されている。しかしここでは質問応答タスク全体の精度についてのみが議論されており、どのような言い換えが貢献したのか、または必要だったのかといった言い換え自身に関する議論は行なわれていない。そこで本稿では質問応答に必要な言い換への分析を目的とし、質問応答に必要な言い換への

パリエーションや、解かなければならない問題点について調査する。

言い換え研究の視点から見ると本研究は、質問応答という応用を実例としてその中で必要となる多様な言い換への分析を行なう研究ととらえることができる。これまでに行なわれてきた言い換への研究では、個別の事例を取り上げそれぞれの現象に対してどのような手法で解くことができるかといった議論が主だった [9, 6]。それに対し本研究は、質問応答において必要となった言い換への網羅的に集め、言い換への多様性を探る試みであると言える。言い換への分野においても、このように事例をもとにした分析はこれまでに行なわれてこなかった。

2 言い換え事例の収集

2.1 収集方法

今回の分析では、初めに言い換え事例の収集を行ない、次に得られた事例を分類し問題を選別した。本節ではまず言い換え事例の収集について述べる。

言い換え事例の収集のために質問文とその正解を含む文書(以降、対象文書)の対を用い、人手により言い換え事例を作成した。作成した言い換え事例の例を図1に示す。図1ではQ0とD0がそれぞれ質問文と対象文書である。もしQ0とD0を直接的な言い換えと見なすと、言い換への種類は組み合わせ的な数になり収束しなくなる恐れがある。そこで本稿では、Q1~Q2, D1~D5に示すようなプリミティブな言い換への分解して収集した。

図1に示すように今回の分析では質問文と対象文書の両方を言い換えた。一般的に対象文書には質問文より多くの情報があるために、質問文から対象文書に向かう言い換への情報は、情報を付加・補足する生成的なものになり、機械的な実現は難しい。そこで今回は、質問文に対しては疑問文から肯定文への言い換への標準的な形への言い換へのとどめ、対象文書を言い換へることにより質問文の標準形に近付けるという方法で言い換え事例を作成した。質問文の言い換へのそのほ

*Collecting and Analyzing Paraphrases that are Useful for Question Answering.

TAKAHASHI Tetsuro, INUI Kentaro, SEKINE Satoshi and MATSUMOTO Yuji.

[†]Nara Institute of Science and Technology

[‡]New York University

Q0>広辞苑第五版はいつ発売されましたか。
 Q1>広辞苑第五版はx(いつ) 発売された。
 Q2>広辞苑第五版はx(いつ) 発売する。

D0>日本の代表的な国語辞典として知られる広辞苑(岩波書店)の第五版が11月11日に刊行される。
 D1>広辞苑(岩波書店)の第五版が11月11日に刊行される。
 D2>広辞苑の第五版が11月11日に刊行される。
 D3>広辞苑第五版が11月11日に刊行される。
 D4>広辞苑第五版が11月11日に発売される。
 D5>広辞苑第五版は11月11日に発売する。

図 1: 言い換え事例の例

とんどが質問文に特化した言い換えであり分析に偏りが生じてしまうので、今回は対象文書からの言い換えのみを分析の対象とした。事例の中には照応・省略の補完が必要となるものも多くあったが、そのような場合は照応・省略の補完も言い換えと同じように扱った。言い換え事例の作成は 1) 文の抽出, 2) 照応・省略の補完, 3) 不要個所の削除, 4) 言い換え, の順で行なった。

データには関根ら [10] によって作成された質問応答のデータと QAC [3] で使われたデータのうち、正解と十分な証拠がテキスト中に含まれる質問 384 問を用いた。

2.2 変換の分類

2.1 節で述べた方法で作成した言い換えパターンは大きく表 1 のように分類できた。

表 1: 変換の分類

活用形の変換	53 (3%)
照応・省略補完	227 (15%)
不要個所の削除	334 (21%)
言い換え	941 (61%)
合計	1555 (100%)

本稿では「言い換え」だけに焦点を当て分類を行なった。活用形の変換は照合の段階で吸収できる問題である。また照応・省略の補完については言い換えとは別の問題として異なるアプローチを用いるべきである。質問応答に言い換えを用いる場合に、不要個所を削除し言い換え元の範囲を決めるという問題は重要であるが、言い換への分析に焦点を当てるために今回は分析の対象としなかった。

3 言い換えパタンの分類

2.2 節で言い換えに分類された 941 個の変換は表 2 のように分類できた。

表 2: 言い換えパタンの分類

記号の処理	138 (15%)
構文的言い換え	338 (36%)
推論的言い換え	465 (49%)
合計	941 (100%)

言い換えは少なからず元の表現の意味を変える。その変化の大きさに明かな境界はないため言い換えと推論を明確に分けることはできないが、本稿では便宜上語彙・構文的な操作により可能となる言い換えを「構文的言い換え」と呼び、それ以外を「推論的言い換え」と呼ぶ。それぞれについて説明する。

3.1 記号の処理

記号の処理とは、以下の例のように括弧や句点などの処理を行なう言い換えである。

- (1) 昭和記念公園 (180 ヘクタール) 昭和記念公園は 180 ヘクタール
- (2) チェコが 99 年から北大西洋条約機構 (NATO) に加わる
チェコが 99 年から NATO に加わる
- (3) 「明の明星」は金星だ 明の明星は金星だ

括弧の使用の多様性などの問題もあるが、これらのうち多くはいくつかのパターンとシソーラスを用いることで規則化できる。

3.2 構文的言い換え

分類の結果を表 3 に示す。語句レベルの言い換えや構文的な言い換えについてはこれまでに分類・整理が行なわれてきている [5, 2]。本節で挙げる言い換えはその範疇に入り、その多くは構文的規則と辞書とで実現できる。辞書の構築または獲得が課題となるが、それらは言い換の種類毎に個別の問題として扱っていく必要がある。たとえば同義語・類義語への言い換えには同義語・類義語の辞書、機能動詞結合には機能動詞の辞書が必要となる。また変換後の適切性をチェックするための知識も必要である。

- (4) スタートする 始まる
- (5) 小林さんの造語だ 小林さんの作った言葉だ
- (6) 盲導犬は日本で約 800 頭いる 盲導犬は日本に約 800 頭いる
- (7) 日よけとして使われる 日よけに使われる
- (8) 初優勝を飾った 初優勝した
- (9) 英高級車メーカー 英国の高級車メーカー
- (10) ジャガイモなどを食害する ジャガイモなどにつく
- (11) 鳥と恐竜の中間 恐竜と鳥の中間
- (12) テルミンはテルミン博士によって生み出された テルミンはテルミン博士が生み出した
- (13) ペティナイフ程度しか使わない ペティナイフ程度を使う

表 3: 構文的言い換えの分類

言い換えの種類	数	例文
同義語・類義語への言い換え	123	(4)
語釈文への言い換え	9	(5)
格の交替	59	(6)
機能語相当表現の言い換え	10	(7)
機能動詞の追加・削除	10	(8)
名詞句の展開・結合	38	(9)
格変化を伴う動詞句の言い換え	18	(10)
対称表現間の交替	8	(11)
態の交替	16	(12)
否定表現の肯定化	2	(13)
慣用句から一般的な句への変換	2	(14)
分裂文の標準化	5	(15)
連体節主節化	38	(16)
計	338	

(14) ディープブルーの軍門に降った ディープブルーに負けた

(15) 格助詞をとるのが名詞である 名詞は格助詞をとる

(16) 注目を集めるリナックス リナックスは注目を集める

3.3 推論的言い換え

推論的言い換えに分類された言い換えの数は 465 で言い換え全体の約半数であった。これらを詳しく見ると、「抽象化・選択」による推論と「属性・属性値」に関する推論の 2 種類に大きく分けられることが分かった。

3.3.1 抽象化・選択

465 事例の推論的言い換えの内 26% にあたる 123 事例は、「抽象化」や「並列句からの選択」を行なう言い換えだった。以下にその例を示す。

(17) イギリスダーラム州の工場 イギリスの工場

(18) 1994 年 12 月に発売される 1994 年に発売される

(19) 体臭 におい

(20) IBM のスーパーコンピューター「ディープブルー」に負けた IBM の「ディープブルー」に負けた

(21) プレイステーションとサターンが発売される プレイステーションが発売される

これらの言い換への大部分は構文的なパターンとシソーラスにより生成できる。同時に多くの誤った言い換を生成する恐れがあるが、抽象化や選択に関する言い換えにおいては意味的等価性を保証することはできないので、どこまでを正しい言い換えとするかをはっきりと定めることは難しい。言い換え後の誤り訂正検出の問題と同様に考えなければならない問題である。

3.3.2 属性・属性値

テキスト中のある対象物についてその属性と属性値を見付けることで解決できる事例が 465 事例の内 37% にあたる 170 事例あった。以下に例を示す。

(22) 全長 3910 メートルの明石海峡大橋 明石海峡大橋の全長は 3910 メートル

(23) 富士山は標高 3776 メートル 富士山の標高は 3776 メートル

(24) トリノは人口 220 万人の都市 トリノの人口は 220 万人

(25) 美空ひばりの生地の横浜 美空ひばりの生地は横浜

(26) グアテマラはスペイン語が公用語である グアテマラの公用語はスペイン語である

(27) チョコレートはカカオ豆を原料にする チョコレートの原料はカカオ豆

これらの事例では、さまざまな表現からある対象物の属性とその値を表わす表現に言い換えている。したがって { 対象物, 属性, 属性値 } の組を抽出することができれば、「対象物の属性は属性値」のようなボタンを用いることで言い換の生成は容易に可能である。

3.3.3 その他の推論的言い換え

以下にその他の推論的言い換の例を示す。

(28) アフリカマイマイは小笠原などに移入した アフリカマイマイは小笠原などにすむ

(29) 日本サブウェイの親会社はサントリー サントリーの子会社は日本サブウェイ

(30) 富士山に次ぐ日本で第二位の高峰 富士山は日本で第一位の高峰

(31) 日本第一の湖 日本最大の湖

(32) マイクロソフトは Windows を発表 マイクロソフトは Windows を作っている

これらの言い換えには語彙・構文レベルの解析や操作だけではなく、より深い意味理解や知識の獲得が必要となる。たとえば (30) では元の文で表現されている比較の関係を理解する必要がある。また (31) では名詞句あるいはそのクラスについて最大の表現方法の知識が必要となる。(32) の例のように名詞句間に支配的な関係が存在する場合は Lin ら [7] や Ravichandran ら [8] の手法で獲得できる可能性がある。

4 属性・属性値関係抽出タスクの提案

3.3.2 節で述べたように、語彙・構文的な言い換では扱えない言い換の多く (37%) は属性・属性値の表現間の言い換であった。1 節で述べたように質問応答においては質問文と対象文書との照合方法が重要な焦点となるので、これらの種類の問題については言い換を用いずに属性と属性値の間で照合する手法が考えられる。そこで我々は、

「文書中から任意の対象物についてその属性と属性値の組を可能な限り抽出する」

というタスクを提案する．たとえば，(33) のような文書があったときに，

(33) ... 自動車道を走り、全長 3 9 1 0 メートルの明石海峡大橋を通過、淡路サービスエリア付近で ...
ここから { 明石海峡大橋, 全長, 3 9 1 0 メートル } という情報を抜き出すタスクである．これは (22) の言い換えに対応するが，ここで提案するタスクでは言い換えによる照合ではなく { 対象物, 属性, 属性値 } の組の間での照合を前提とし，質問文は考慮せず文書集合だけを対象とする．

このタスクを情報抽出 (IE) と比較すると，IE は抽出すべき情報のカテゴリーを人手で与えるのに対し，このタスクではカテゴリーやそれを抽出するためのパターンを与えることなく，ドキュメント全体から属性・属性値の関係を抽出する．その中から特定の属性・属性値を選択することで IE と同様の出力が得られるので IE の一般化ととらえることができる．また Logical Form などの意味構造への変換と比較すると，それらがテキスト中のすべての情報を意味構造に変換するのに対し，今回提案するタスクでは属性とその値だけを取り出すことに焦点を当てている．そうすることで問題を限定することができ，これまでに提案されている IE の手法などを部分的に適用できるようになると考えられる．

現在，質問応答の精度は質問文のセットに対してどれくらいの精度で正解を出力できるかによって求められる．しかし質問応答システムは一般的に質問文解析，ドキュメント検索，パッセージ抽出，解答候補のランキングなどさまざまなモジュールからなっており，上記の評価方法では問題の所在がクリアにならない．そこで今回提案したタスクを質問応答から分離した一つのサブタスクとして，単体での実験や精度の評価をしていきたい．

5 まとめ

本研究では質問応答に必要な言い換えを抽出しそれらを分析してきた¹．分析の結果，質問応答に必要な記号的処理や構文的言い換えを洗い出すことができた．これらの知識は言い換えを用いる質問応答システムだけでなく，言い換えを用いないシステムにとっても，照合を行なうための知識として有用なものである．また語彙・構文的な言い換えの範囲を越えた言い換えについても様々なバリエーションを抽出

できた．そしてそれらの分析を通して，問題の一部については属性と属性値を抽出することによって解決できる見込みが得られた．そこで我々は質問応答のサブタスクとして，文書中の任意の対象物の属性とその値を抽出するタスクを提案した．実際に質問応答の多くの事例でこの情報が必要となっていたことから取り組む価値のあるタスクであると言える．今後はこのタスクに取り組んでいきたい．

参考文献

- [1] Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. Web question answering: Is more always better? In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 291–298, 2002.
- [2] 藤田篤, 乾健太郎, 松本裕治. 言い換え知識の類型化と例文集構築の試み. 言語処理学会第 10 回年次大会発表論文集, 2004.
- [3] Jun'ichi Fukumoto, Tsuneaki Kato, and Fumito Masui. Question answering challenge (qacl): Question answering evaluation at ntcir workshop 3. In *Working Notes of the Third NTCIR Workshop Meeting: QAC1*, 2002.
- [4] Ulf Hermjakob, Abd Echibahi, and Daniel Marcu. Natural language based reformulation resource and web exploration for question answering. In *the 2002 edition of the Text REtrieval Conference (TREC)*, 2002.
- [5] 乾健太郎. 言語表現を言い換える技術. 言語処理学会第 8 回年次大会チュートリアル資料, pp. 1–21, 2002.
- [6] 近藤恵子, 佐藤理史, 奥村学. 格変換による単文の言い換え. 情報処理学会論文誌, Vol. 42, No. 3, pp. 465–477, 2001.
- [7] Dekang Lin and Patrick Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, 2001.
- [8] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.
- [9] 佐藤理史. 論文表題を言い換える. 情報処理学会論文誌, Vol. 40, No. 7, pp. 2937–2945, 1999.
- [10] 関根聡. 百科事典を対象とした質問応答システムの開発. 言語処理学会第 9 回年次大会発表論文集, pp. 637–640, 2003.
- [11] Tetsuro Takahashi, Kozo Nawata, Sinya Kouda, and Kentaro Inui. Seeking answers by structural matching and paraphrasing. In *Working Notes of the Third NTCIR Workshop Meeting: QAC1*, 2002.
- [12] Ellen M. Voorhees. Overview of the trec 2002 question answering track. In *Proceeding of the The Text REtrieval Conference (TREC)*, 2002.

¹分析の詳細は下記 URL に掲載した．