

文書記述内容のメタデータ化について

斉藤孝広[†]

takahiro@flab.fujitsu.co.jp

[†] 富士通研究所

[†] 〒 211-8588 川崎市中原区上小田中 4-1-1

文書の記述内容をメタデータとして扱うために必要な技術と従来の自然言語処理技術との関連について考察を行ない、記述内容を表すメタデータを文の係り受け組とそれに付随する属性情報のセットで表現する方式を提案する。

また、実際のアンケート文からこのメタデータを作成し、これがアンケートから主な要望内容を抽出するという分析タスクにおいて有用である事を、人手による分析結果との比較で示す。

Making the meta-data of contents in a document and its application

SAITO Takahiro[†]

[†] Fujitsu Laboratories Ltd.

[†] 4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki, 211-8588, Japan

This paper discusses “meta-data of contents in a document” and its forms.

Firstly, it describes requirements of handling contents in a document as a meta-data, and related NL techniques.

Next, it proposes the method of making meta-data of contents with dependency structures of a sentence.

Finally, it reports about an experiment proving an effectiveness of meta-deta of contents for an analysis of the text in questionnaire.

1 はじめに

厳密な意味でのメタデータは、例えば書誌情報のような「文書自身に関する情報」であり、これは文書の中身を見ても獲得不能な情報である。

しかしながら、これらのメタデータが情報資源の発信・検索・アクセス・利用にいたる過程において重要な役割を演じるのと同様に、その文書の記述内容を扱いやすい形で表現したものもこれらの各過程において有用であると考えられる。本稿における「文書記述内容のメタデータ」はこのような情報を指す事にする。

第2章で文書記述内容をメタデータとして扱う方式を考察し、文書の内容を扱ってきた従来の自然言語処理技術との関連を論じる。

第3章において、この情報を文の係り受け情報により作成する方法について述べ、続く第4章では、本方式によって作成されたメタデータの係り受け関係を辿ることで、アンケート文から主な要望内容を抽出できる事を示す。

2 文書記述内容のメタデータ化

本稿においては、メタデータを本文から獲得できるかどうかに関わらず、文書の特定の属性を表現するもの(属性値)として捉えている。一方、これらの属性値が、一意性条件(同じ属性を表現する属性値は同一の値をとる)を満たしている場合、集計等の統計的処理が簡単に行なえるので便利である。そこで、この一意性条件を満たすような「文書記述内容のメタデータ」を作成する方法について考察を行なう¹。

2.1 文書記述内容のメタデータの表現形式

ある文が表現する内容は多種多様な属性を含む。例えば「携帯は便利だと思う」という簡単な文であっても、「携帯＝便利」という文の主題に関する属性値や、「記述者主観」という文の種別に関する属性値、また「現在」といった時制に関する属性値を内包している。これらの属性値の内、文書記述内容として最も重要な属性値は主題に関する

¹メタデータとしては一意性条件を満たさないものを作成し、その同義性をオントロジで定義する立場もある。ただし、今回の場合は巨大なオントロジが必要となる

る属性値であり、他の属性値はその内容のさらに詳細な部分を表現していると考えられるので、これらをまとめた形式をメタデータとする。つまり、上の例であれば、「(「携帯は便利」, 主観 / 現在)」といった形式で記述するという事である。この内、主題を表す属性値はテキスト表記を継承しているので、作成したメタデータが一意性条件を満たすためには、主題に関する属性値の統合処理を行なう必要がある。またその他の属性値については上で挙げたような文法的な属性値以外でも主題を表す属性値に出現する自立語の意味属性を属性値としてメタデータに追加する事により、より内容に踏み込んだ情報も付与することができる。例えば、「便利」という単語は肯定的な評価を意味する単語であるので、属性値としてメタデータに追加し、「(「携帯は便利」, 肯定的評価 / 主観 / 現在)」といった形式で表現する。

一方、メタデータを利用する立場では、どのような属性値を記述内容を区別するものとするかは、その記述内容の種別や利用目的に依存する。例えば、文の時制に関する属性値は一般的には記述内容を区別する重要な情報である。これは例えばアンケート文において「携帯は便利だと思う。」と「携帯は便利だと思ってた。」という2つの文において、時制の違いが表現内容の違いに直結することから明らかである。一方、障害管理情報の対処方法を記述する文(例えば「HDDを交換する」と「HDDを交換した」)においては時制の違いが内容に与える影響はほとんどないように思える。この例のように、記述される内容によって必要のない属性値が決まると思われるものもあるが、実際の所、その基準の不変性は保証できない。結局、種々の目的に利用可能なメタデータとしては、内容の区別に利用される可能性がある属性値は付与しておき、付与された各属性値を利用するかどうかは、利用者側に任せるのが妥当であろう。

2.2 自然言語処理における記述内容の取り扱い

文の記述内容の扱う処理は自然言語処理の基本的なテーマであり、種々の技術が各々の目的において文書の記述内容を扱っている。この内、本稿にて取り扱う文書内容のメタデータ化に関連が深い技術をいくつかとりあげて考察を行なう。

1. 文書要約

文書の記述内容を直接扱う技術の一つに文書要約がある。文書要約はその要約方針により generic な要約と user-focused な要約に分けられる [1]。

前者は文書のもつ情報量の低下を抑えつつ、その記述量を削減した要約を作成するものである。一方、後者は記述内容中のユーザの興味を反映した要約を作成するものであり、ユーザの興味対象以外の記述内容はむしろ要約には含まれない方がよい。

その意味で、user-focused な要約は、文書からある属性値に相当する記述内容を抽出するものであり、本稿における文書内容を表すメタデータの一つの表現形である。しかしながら、文書要約における出力結果は一般的には文または文の一部の形式を取るもので、一意性については全く考慮されていない。結局のところ、この結果についてはさらに本稿で述べているような処理を行なう必要がある。

2. 情報抽出

情報抽出は、記述文より特定の知識を抽出する技術である。

例えば、[2] では、人事異動に関する新聞記事より、「人名」、「会社名」、「異動前役職名」、「異動後役職名」などの属性値を抽出する例が紹介されている。

また、[3] は、営業日報の記述文から「施策」と「その効果」といった因果関係情報を抽出する。また、[4] は、プリンタ障害記述文から、「現象」、「原因」、「対策」の記述内容を複合語の形式で抽出するものである。

これらの抽出結果は、何の属性値であるかがはっきりしているので、メタデータとして非常に利用しやすい。しかしながら、その抽出ルールは抽出対象や抽出項目に強く依存し、広い文書種別に適用可能な汎用的ルールを記述するのが難しいという問題を含んでいる。

3. 文書分類

文書分類は、文書の記述内容に基づいて予め設定された定型情報を付与する技術であり、

この技術によって付与された定型情報はそのままメタデータとして利用可能である。また、この定型情報はどの属性値かがはっきりしているという意味において、情報抽出の処理結果と同様にメタデータとして利用しやすい。ただし、付与する属性値のセットを処理前に用意する必要があり、今回の「記述内容」の属性値セットを用意するのは不可能であろう。

4. 言語的アノテーション

情報抽出や文書分類のような、観点を固定した情報を抽出するという技術に対して、文書内容に関する汎用的に用いる事が可能な情報を作成する技術として言語的アノテーションと呼ばれる技術がある。

例えば GDA (Global Document Annotation) では、文中の係り受け情報などが XML のタグとして文に埋め込まれる [5]。これらのタグは、計算機による記述内容の理解を可能にすることで深い自然言語処理を実現する事を目的としている。これらのタグを利用する事により、記述内容に関する属性値を抽出することができる²。

ただし、ここで付与される情報は汎用的な利用が可能なものであるが、その利用方法自身はアプリに任されている。一方、記述内容のメタデータ化の目的は、書誌情報を利用するのと同程度の簡便さで文書の内容を取り扱う事を可能とするものであり、このギャップを埋める仕組みが必要であると考えている。

結局のところ、本稿が主題としている文書内容を表すメタデータは、文書分類や情報抽出ほど抽出内容を特化した（その分、利用は簡便な）ものではなく、言語的アノテーションほど利用の際の敷居が高い（その分、汎用的に利用可能な）ものでない。いわば、その中間を目指すものである。

3 手法

文書記述内容のメタデータ化の手法として、文節間の係り受け情報を用いた手法を提案する。

²実際の所、次章で述べる文書内容メタデータ作成方式においては、抽出対象文において係り受け解析を行ない、その結果を基にした処理を行なっている。

まず、係り受け組から元の文と同じ文意の文を構成可能であるので、係り受け組は文意を保存している分割単位であるといえる。また、係り受け組はそれ自身が、元文の持つ記述内容をうまく分割している場合が多い。例えば「携帯はとても便利」という文から、「携帯は - 便利」、「とても - 便利」という2組の係り受け組が得られるが、各々が「主題」、「程度」の属性値となっている。つまり主題を表す属性値を、適切な係り受け組の形で表現する事ができる。これらの性質が、本稿で提案する記述内容のメタデータの記述形式に適切であるからである。

1. 係り受け解析

入力文の係り受け解析を行なって、文節間の係り受け情報を抽出する。この際、一般的な文法ルールに基づく文節分割ではなく、例えば「～する事がある」といった特定のフレーズを1文節とするような文節分割ルールを採用する事で、特定のフレーズに基づく属性値を抽出するルールを簡便に記述する。

2. 属性抽出

観点に基づき、係り受け組の選定やメタデータ化する属性値を抽出する。前述の通り、抽出される属性値は、元の係り受け組の表記をそのまま継承する主題に関する属性値と、その補助情報としての属性値である。補助情報としての属性値は、「否定」や「可能」といった述語文節中の助動詞などから抽出されるもの、「～すれば良い」といったフレーズから抽出されるものがあり、それらの抽出ルールを用意する必要がある。

ただし、これらの抽出ルールは、情報抽出において用いられるものとは異なり、それ自身で特定の知識を表現するようなもの抽出するものではなく、広く文書一般に適用可能なものである。例えば「否定」属性は単に述語文節に否定の助動詞を含むかどうかでのみ判定する。勿論、これらの属性値は単独ではほとんど意味を持たないが、利用する属性値のセットを利用の際に指定することでメタデータの意味づけを行なう事を可能にする。

3. 統合化

主題に関する属性値は、係り受け組の形で表現されているが、これは入力文の表記をそのまま継承している。そのため、一意性条件を満たすようなメタデータを作成するには、この属性値の統合化が必要となる。この統合処理としては、単語に関する同義語辞書を用いた統合や複合語の抽象化([6])、述語文節の語尾の統一化などが挙げられる。

ただし、これらの単語や文節を単位にする統合処理では、一意性を完全に達成することはできず、係り受け組としての統合化が必要となる。例えば「価格を - 安くする」と「値段を - 下げる」という2つの係り受け組は、係り受け組のレベルでは同義であるが、その構成文節の「価格」と「値段」、「安くする」と「下げる」の各々には、同義性が成立していないので、単語単位の統合は行なえない。

しかしながら、このような係り受け組の統合ルールを整備する事は、単語レベルの同義情報を整備する以上にコストがかかる。まずは低コストな統合処理のみを適用したメタデータでも、ある利用目的において有用であることを示すことが重要であろう。

4 メタデータによる文書内容分析実験

本稿が提案する「文書記述内容を表すメタデータ」は、主題を表す係り受け組と、それに付随する属性値のセットの形式で表現される。係り受け組は文節間の構文情報を表現しているので、係り受け組を辿る操作で文節間の修飾関係を見ることが、段階的に詳細な記述内容を獲得する事が可能である。

そこで、自由記述形式のアンケート文³ 2665文(元データが複数文で構成されている場合は分割した)の分析に関し、本手法で作成されたメタデータの検索や集計といった単純な処理で、主要要望内容を抽出可能かどうかでその有効性を検証する。

³マイボイスコム株式会社発売のMyVoiceレポート“航空会社に対するイメージ分析調査”の自由記述項目である「航空会社に対する意見・要望」フィールド。ただし、「飛行機を利用した事はない」といった、本来の趣旨とは外れた内容の文も含まれている。

4.1 メタデータの作成

まず、メタデータを以下のように作成する。この処理結果は、前章で述べたように一意性条件を完全に満たすものではない。ただし、今回の実験の目的は、このような低コストな処理で作成したメタデータの有用性を検証する事にある。

1. 係り受け組の選択

「名詞文節 - 述語文節」の形式の係り受け組(以後、「述語係り受け組」)を扱う。また、補助的な情報として、その係り元の名詞文節を係り先とする「修飾文節 - 名詞文節」の形式の係り受け組(以後、「名詞係り受け組」)からもメタデータを作成する。

ただし、通常の文節の定義に基づく係り受け組をそのまま利用するのではなく、述語文節の連続(ただし、その間に副詞をはさんでも良い。このような係り受け組は述語文節間の係り受け関係として獲得される)においては、それらをまとめて1文節とし、その先頭の述語のみを主題属性値の述語文節とした。これはフレーズによる属性値を抽出するものである。例えば、通常の文法においては2文節となる「～してもよい」というフレーズを1文節とする事で、「～する」という述語と述語属性値「許可」を抽出する。

また、その文節自身の意味が弱い場合は、それに直前の文節を接続して1つの文節とした。この処理は、名詞文節が「事」、「時」、「場合」といった形式名詞で構成される場合や、述語文節がサ変動詞語尾である場合に行なう。

なお、文節数が1の文は係り受け組が抽出できないので、今回の抽出対象から外した。

2. 抽出属性

述語文節に含まれる助動詞やフレーズによって判定される表1の属性を付与する。

なお、表中の属性値「要望フレーズ」は、要望を表しているとみなせるフレーズに付与される属性値である。具体的には、まず別のアンケートデータ⁴に出現した動詞、形容詞や

形容動詞について、その意味属性が要望を表しそうな単語のリストを作成し、述語文節にこのリスト内の単語が含まれている場合にこの属性値が付与される。一方、属性値「要望」は、述語文節に助動詞「～た(い)」が含まれている時に抽出される。

属性値	典型的表現
「否定」	“～しない”
「疑問」	“～するの?”
「受身」	“～られる”
「可能」	“～できる”, 可能形語尾
「使役」	“～させる”
「要望」	“～したい”
「使役要望」	“～して欲しい”
「許可」	“～してよい”
「義務」	“～しなければならない”
「禁止」	“～してはならない”
「十分」	“～すればよい”
「要望フレーズ」	“～すると便利”, “～だと面白い”

表 1: 抽出属性値

3. 表記の統合化

述語係り受け組においては、助詞を削除する。削除された助詞がほとんどの場合で推測可能であり、助詞を含めた統合化を厳密に扱うには文の並列構造や態の解析といった深い処理が必要であるからである。また、同義語の統合処理については、表記の揺れといった同義性が確実なもののみに対して行なう。

一方、述語文節については、その先頭の自立語に終止形語尾を付与して言い切りの形に変換した。ただし、「否定」属性については、係り受け組の表記にも反映させた。

以上より、例えば、「家族向けサービスが増えてくれば便利だと思う。」という文からは、「家族向けサービス - 増える」という主題に関する属性値と、属性値「要望フレーズ」が抽出される⁵。

ドコモ「iモード」の利用に関する調査-2”の自由記述項目である「携帯電話を活用した新しいサービスへの意見・要望」フィールド。

⁵一般的な意味での文節、「増える」「便利である」「思う」は述語文節の連続により1文節に統合され、その文節に

⁴ マイボイスコム株式会社発売の MyVoice レポート “NTT

4.2 メタデータによる要望分析

回答者の要望意見を表しそうな文の属性値としては、「要望」、「使役要望」、「義務」、「十分」、「要望フレーズ」などが挙げられる。まず、これらの属性値で検索を行なった結果である 1032 文のみを分析対象とする。

A: キーワードによる話題把握

述語係り受け組（「名詞 - 述語」の形式の係り受け組）における係り元文節（名詞）の異なり数は 766 個であった。そこで、この単語の出現頻度上位 20 位（同頻度のものがあるので 23 個のキーワード）は以下の通りである。なお、意味の似た単語はまとめた。また、括弧内の数字はその出現頻度である。

1. サービス (137)
2. マイレージ (97), マイレージサービス (18)
3. 料金 (49), 運賃 (36), 価格 (27), 値段 (17), 航空運賃 (15)
4. ポイント (43), マイル (14)
5. 飛行機 (18)
6. 有効期限 (16), 期限 (9)
7. 航空会社 (15)
8. 人 (14)
9. 割引 (11)
10. それ (10)
11. 商品券 (9), 商品 (8)
12. 航空券 (9)
13. ツアー (9)
14. 機内食 (8)
15. 安全性 (8)

これだけでは具体的な要望内容が不明なものも多いが、少なくとも何に関連した要望が多いかは掴む事ができる。まずはこれらのキーワードを、さらに分析を進めるキーワードのセットとする。

「便利」という単語が含まれるので、属性値「要望フレーズ」が抽出される。

B: 詳細内容の獲得

キーワードのみで具体的な要望が把握できないものについては、そのキーワードを係り元または係り先とする係り受け組を見ることでその記述内容を知ることができる。

例えば、出現頻度 1 位の「サービス」について、まずそれを係り元とする係り受け組で検索を行ない、その係り先の述語文節を集計した所、その異なり数は 59 とあった。また、その述語文節は、「ある」、「充実する」、「欲しい」といった、望まれるサービス内容を記述している文に用いられると思われるグループと、「やめる」、「いらない」といった、必要のないサービス内容を記述している文に用いられると思われるグループに分けられた。

そこで、前者の述語グループでさらに絞り込みを行ない、その中で今度は「サービス」を修飾する文節を集計した所、その異なり数は 44 となり、内容的には以下のように分類できた。

● 料金に対するサービス

文節例: 「安くなる」、「無料になる」、「格安な」、「割引チケットの」

● その他のサービス内容

文節例: 「景品交換などの」、「ペット関連の」、「ビジネスクラスのような」

● 「～向け」サービス

文節例: 「子供向けの」、「家族向けの」、「単発客への」

● この時点では情報が不十分なもの

文節例: 「溜る」、「利用する」

情報が不十分なために内容の推測ができなかった文節 (23 種類) は、例のような述語の連体修飾文節であるものがほとんどであった。これについてもさらにこの述語文節を係り先とする係り受け組を見ることにより、「買物でも / マイレージが / 溜る」、「期間限定で / 割安に / 利用する」といった文節が得られ、その内容を把握することができた。

一方、「やめる」といった後者の述語グループに対して同様の操作を行なうと、「過剰な」、「ごまかしのよう」といった文節が得られた。

C: 人手による分析結果との比較

本実験に使用したデータには、データ発売元が人手で分析を行なった結果が分析レポートとして附属している。そこで、この分析レポートで挙げられた要望項目が、メタデータを用いた機械的な検索及び集計操作で獲得可能かを検証する。

この分析レポートには、自由回答中の主要望として、18 データ (元文) を挙げている。その内容を要望項目としてまとめ、それらの要望項目を記述している文に含まれる単語やフレーズを小項目キーワードとして挙げると以下の様になる。

- マイレージの有効期限の廃止・延長
- ポイント獲得範囲の拡張
各社共通化, クレジットカード, 買物, ツアー
- ポイント利用範囲の拡張
少ないポイントの利用, 交換商品の充実
- 分かりやすいサービス
料金体系, マイレージサービス
- サービスの充実
エコノミー, 座席, 機内食, 子供向け
- インターネットを利用したサービス
割引, 予約, 取消
- 値下げ・割引
子供割引, 家族割引, キャッシュバック
- 安全性

一方、メタデータを利用して文に記述された要望内容を獲得する操作を、以下の手順で規定する。

準備操作: 係り元文節の名詞がキーワードである述語係り受け組を検索する。今回は、ここで検索された係り受け組の、頻度 2 以上かつ上位 10 位のものを操作 1,2 の対象とする。

操作 1: 準備操作の検索結果である係り受け組において、その名詞文節を係り先文節とする名詞係り受け組を検索する。今回はこの検索結果中、上位 10 位までが検索されたとする。

操作 2: 準備操作の検索結果の上位の係り受け組において、その述語文節を共有する述語係り受け組を検索する。今回はこの検索結果中、上位 10 位までが検索されたとする。

例えば、この操作を「マイレージ」をキーワードとして行なうと、以下の様な結果が得られる。

- 準備操作の検索結果 11 件中に、「マイレージ - 溜る」, 「マイレージ - 貯めやすくなる」といった係り受け組が出現する。
- 操作 1 の検索結果 10 件中に「航空会社共通の - マイレージ」といった係り受け組が出現する。
- 操作 2 の検索結果 6 件中に、「ショッピング - 溜る」, 「食事 - 溜る」, 「格安チケット - 溜る」といった係り受け組が出現する。

これらより、「マイレージポイントの共通化」や「マイレージ獲得範囲の拡大」といった要望項目や小項目キーワードを獲得することができる。

この操作を A で生成したキーワードリストの 23 キーワードについて行なった結果、1 キーワード当たり平均 6.73 回の検索操作で上で挙げた要望項目やその小項目キーワードが抽出された。その抽出例を以下に挙げる。

- 「有効期限 - なくす」& 「マイレージの - 有効期限」
- 「ポイント - 交換する」& 「クレジットカードの - ポイント」
- 「ポイント - 交換する」& 「少ない - ポイント」
- 「サービス - 向上する」& 「座り心地などの - サービス」
- 「料金 - 下げる」& 「子供の - 料金」
- 「安全性 - 競争する」

また、今回の操作で洩れてしまったものは、2 つの要望項目 (「分かりやすいサービス」, 「インターネット利用したサービス」) と 1 つの小項目キーワード (「キャッシュバック」) であった。つまり、人

手による分析で抽出された要望 6 項目の中の 4 項目、また、小項目キーワードにおいては 18 個中の 12 個が得られた事になる。

洩れてしまった要望項目及び小項目キーワードについて、その原因を調査した所、体言止め文において検索対象とする述語属性値が付与されないため、始めの絞り込み時点で洩れてしまっていた事が分かった。これについては、体言止め文については今回のアンケートの質問文を考慮し、「～を望む」といった述語を付与したデータで処理を行なうことで、1 つの小項目キーワード(「料金体系」)以外の全ての要望項目及び小項目キーワードが抽出されることを確認した。

また、分析レポートの小項目キーワードにはないが、「夏休みなどのトップシーズンにおける料金への要望」、「レンタカーや宿泊施設へのマイレージの利用」といった要望も今回の操作で獲得できていた。

4.3 今後の課題

タスクの性質上、今回は特に問題とならなかったが、記述内容の正確な集計を必要とするタスクに、このメタデータを利用するには以下の課題を解決する必要がある。

1. 抽出属性値と内容属性のギャップの問題

記述内容の属性値である「意見・要望」を、単に文法的な属性値のみで代用しているため、精度に問題がある。

2. 一意性条件の問題

一意性条件を満たすには、係り受け組としての同義性を判定して統合化する必要がある。

前者に関しては、[7] で述べられているような、より詳細な処理を行なって、判定結果を属性値としてメタデータに付与する事で解決される。ただし、このような内容に踏み込んだ属性値は、属性値間の関係を記述するオントロジを定義して、体系的に扱う必要があると考えている。

また、後者に関しては、完全な自動化は不可能であると思われる。システムとしては、統一候補を出力し、利用者がその正誤を判定して統合ルールを記述するような支援ツール(このようなツール

については、例えば [8] で報告されている)を提供するのが現実的であると考えている。

5 まとめ

文書の内容を表すメタデータとその作成方式について考察を行ない、既存の自然言語処理との比較を行なった。

係り受け情報を用いた方式で実際にアンケート文を対象としてメタデータを作成し、それをユーザの要望分析に利用する手法で、人手で抽出した要望項目が抽出できることを示した。

参考文献

- [1] Mani,I. and Bloedorn,E:Machine Learning of General and User-Focused Summarization,Proc.15th National Conference on Artificial Intelligence,pp.821-826,1998
- [2] 関根聡: テキストからの情報抽出, 情報処理 Vol.40,No.4,pp370-373,1999
- [3] 市村由実 他: 営業日報を対象としたテキストマイニング - 成功事例および機会損失情報の抽出 -, 第 14 回人工知能学会全国大会, pp532-534,2000
- [4] 斉藤孝広 他: 障害情報からのマイニング, 情報処理学会研究報告,FI-61,pp145-152,2001
- [5] Koiti Hasida:Global Document Annotation(GDA), <http://www.i-content.org/gda>
- [6] 斉藤孝広 他: 連想検索における属性語の抽出方式, 第 14 回人工知能学会全国大会,pp171-172,2000
- [7] 乾 裕子 他: 文末表現に着目した自由回答アンケートの分類, 情報処理学会研究報告,NL-128,pp181-188,1998
- [8] 市村由実 他: 日報分析システムと分析用知識記述支援ツールの開発, 電子情報通信学会論文誌,Vol.J86-D-II,No.2,pp310-323,2003