単語情報を用いたパターンマッチングに基づくテキストからのモダリティ情報の抽出

Modality Extraction from Text Based on Pattern Matching Using Morphological Features

長谷川 隆明　小原 永
Takaaki HASEGAWA and Hisashi OHARA
NTT サイバースペース研究所
Cyber Space Laboratories, NTT

Abstract

Recently, robots and computer agent with speech synthesizers have come into real world, and communicate us satisfactorily in specific domain. A problem in Text-to-Speech (TTS) technology in unlimited domain, however, is that the synthesized voice does not provide human emotion yet. In order to acquire human emotion, computer must understand text and add suitable prosody in TTS process. However, natural language understanding(NLU) is difficult. In stead of NLU, information extraction (IE) using shallow parsing is major in the field of analyzing text nowadays. Accordingly, a method for extracting modalities and emotions from text is proposed that uses pattern matching with morphological features in order to add suitable prosody in TTS process. An experiment demonstrates that the proposed method has good performance. Also it points out the controversial issue as future work.

1　はじめに

近年，音声を用いて人間と対話ができるロボットやコンピュータエージェント，あるいはメールやホームページの読み上げを行うサービスが多数社会に登場している．限定されたドメインを対象とした状況では，あらかじめ作り込まれた合成音声を駆使するロボットやコンピュータエージェントは，人間とのコミュニケーションに十分に成功していると言えるだろう．しかしながら，メールの読み上げや物語の朗読など，テキスト音声合成を用いて任意のテキストをロボットやコンピュータエージェントに喋らせるタスクでは，人間のように喜怒哀楽を表現するような豊かな抑揚の付いた合成音声を実現するレベルには至っていないのが現状である．

合成音声に豊かな抑揚を付けるためには，本来は音声に変換するテキストの内容を理解することが不可欠である．しかし，テキストを内容を深く解析し理解するのは非常に難しい．このため昨今では，表層的な解析のみを行ってテキストから必要な情報を抽出する情報抽出という手法が盛んに研究されている [6]．そこで，本稿では，情報抽出の手法を用いることにより，喜怒哀楽などのテキスト作成者の情動を表現する豊かな抑揚の付いた合成音声を実現するために必要となるモダリティ情報をテキストから抽出する手法を提案する．

本稿の構成は以下の通りである．第 2 節では，本稿で対象とするモダリティ情報について述べる．第 3 節では，テキストからモダリティ情報を抽出する手法を提案する．第 4 節で提案した手法の評価を行い，第 5 節で関連研究について述べる．第 6 節で本稿の結論を述べる．

2　抽出対象とするモダリティ情報

本稿では，テキスト音声合成の際に，テキスト作成者の情動を表現するような抑揚を付けるために必要な情報をモダリティ情報と位置づける．モダリティ情報には大きく分けて，テキスト作成者の喜怒

表 1: モダリティ情報の内訳

区分	種類	表現例
感情	喜び	うれしい, おいしい, ラッキー
	怒り	むかつく, 頭に来る, 腹が立つ
	哀しみ	悲しい, うまくいかない
	楽しみ	わくわく, 夢見る, 待ちわびる
	驚き	信じられない, 予想外, まさか
発話行為	陳謝	ごめんね, 許して
	自問	~しようなかなぁ
	勧誘	~しませんか
	祈願	~だといいんだけど
	命令	~したらどうか

哀楽などの感情とテキストの読み手への発話行為 [7] を考えている．本稿では，これらをテキストに存在する表層的な表現のみを手がかりとして抽出することを目的とする．表 1 にモダリティ情報の種類とそれらの手がかりとなる表層的な表現の例を示す．

3 モダリティ情報の抽出

本節では，表層的な情報のみを用いてモダリティ情報を抽出する手法を提案する．本稿で提案する手法の概要を図 1 に示す．まずはじめに，初期化時で抽出するモダリティ情報の定義を単語の属性情報を用いたテキストレベルで記述する．これに対して形態素解析を用いたコンパイルを行い，モダリティ情報パターンを作成する．次に，実行時では入力されたテキストに対して形態素解析を行い，テキストを単語に分割する．最後に，各単語の持つ属性情報を用いたパターンマッチングにより，テキストにモダリティ情報のタグを付与する．以下では，それぞれの処理の詳細について述べる．

3.1 モダリティ情報の定義

初期化時には，はじめに，モダリティ情報の定義として，各モダリティ情報の種類とそれに対応する表現をテキストレベルで記述する．テキストレベルで記述可能な表現は，単語の表記，品詞，読み，一般名詞カテゴリの属性を用いた表現である．本稿における一般名詞カテゴリの体系は，NTT シソーラス [2] の一般名詞意味属性の体系に準じている．

本手法におけるモダリティ情報の定義に関する記述の特長は以下の通りである．

<table>
<tr><td>[喜び]</td><td>表記 = { 美味しい | おいしい | 美味しかった | おいしかった }</td></tr>
<tr><td>[喜び]</td><td>カテゴリ = { 喜び | 成功 | 幸運 }</td></tr>
<tr><td>[哀しみ]</td><td>品詞 = 名詞:動作, 表記 = して { くれず | くれません | もらえず | もらえません }</td></tr>
<tr><td>[哀しみ]</td><td>カテゴリ = {!問題 | !悪意 | !事故 }, 表記 = { が | は | 等 | など | の }, 表記 = { 無い | 無かった }</td></tr>
<tr><td>[怒り]</td><td>表記 = いい加減にし { て | なさい | ろ }</td></tr>
<tr><td>[驚き]</td><td>表記 = {^えっ | ^えーっ | ^エッ | ^エーッ }{$| 、 | 。}</td></tr>
</table>

図 2: モダリティ情報の定義例

- 属性に表記を用いる場合は，複数の単語からなる表記を記述できるので，あらかじめ単語の分割のされ方がわからなくてもよい．
- 単語の属性はシーケンシャルに組み合わせて記述することも可能である．
- 論理和 ({a|b})，先頭マッチ (^)，末尾マッチ ($)，否定 (!) 等の正規表現を用いて記述することも可能である．

モダリティ情報の定義の具体例を図 2 に示す．

例えば，[喜び] を定義した最初の記述では，表記が “美味しい” や “おいしかった” という単語または単語の並びがテキストに存在するときに，これを [喜び] として抽出することを意図している．[喜び] の 2 番目の記述では，一般名詞カテゴリが「喜び」や「成功」,「幸運」であるような単語を [喜び] として抽出することを意図している．“美味しい” は “美味し” と “い” という複数の単語からなるが，定義を記述する際には，これを考慮しなくても良い．[哀しみ] の最初の記述例では，品詞が “名詞:動作” である単語に続いて表記が “してもらえません” という単語または単語の並びなどが続く単語の並び，例えば “ 承諾してもらえません” という表現などを [哀しみ] として抽出することを意図している．[哀しみ] の 2 番目の記述例では，一般名詞カテゴリが「問題」や「悪意」,「事故」以外の単語で，その後に表記が “の” と “無い” 等の単語または単語の並びが続く単語の並び，例えば “クーラーの無い” という表現などを [哀しみ] として抽出することを意図している．

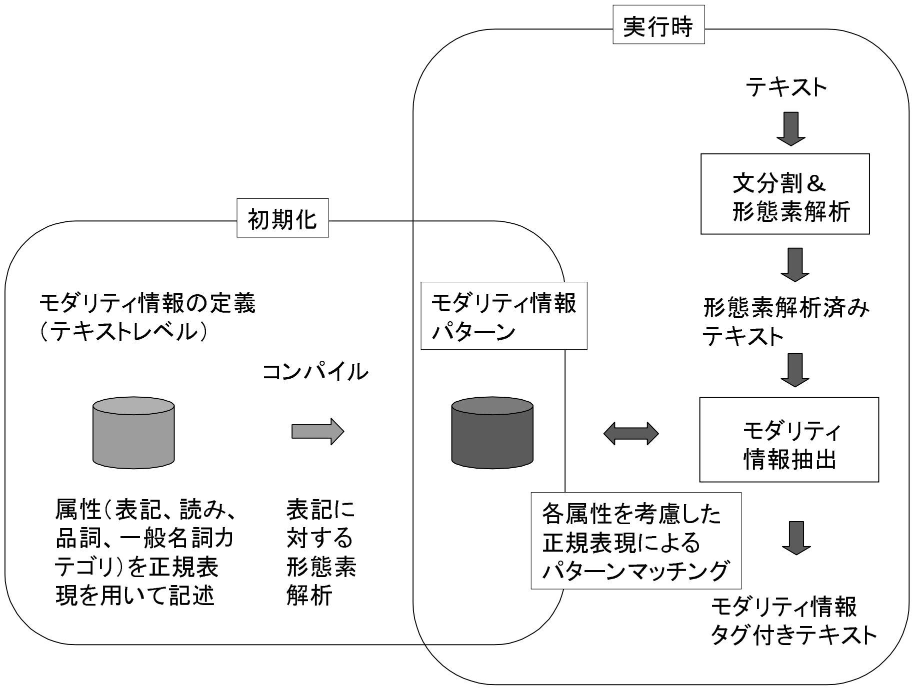

図 1: モダリティ情報抽出の手法の概要

3.2 モダリティ情報の定義のコンパイル

入力されたテキストとパターンマッチングを行うために，テキストレベルで記述されたモダリティ情報の定義に対して，パターンのコンパイルを行い，モダリティ情報パターンを生成する．パターンの生成の順序として，まず論理和などの正規表現が存在すればこれを展開する．次に，モダリティ情報の各定義の表記に対して正規表現の記号を除いた文字列に対して形態素解析を行い，正規表現の記号の意味を考慮した，各単語の表記からなるパターンに分割する．なお，形態素解析には，JTAG[1] を用いている．例えば，図 2 の [怒り] の記述例を例にとると，図 3 に示すようなパターンが生成される．

また，モダリティ情報の定義からモダリティ情報パターンへのコンパイル時に，一般名詞カテゴリについては一般名詞カテゴリ番号へ変換する．

3.3 単語情報を用いたパターンマッチング

実行時では，はじめに，入力されたテキストに対し，一文等を一行とする文分割処理を行う．次に形態素解析を行い，テキストを単語に分割し，各単語の表記，品詞，読み，一般名詞カテゴリの属性情報を得る．例えば，「クーラーの無い中学校の教室はサウナです」という文は，形態素解析の結果，図 4 に示すような各単語の表記，品詞，読み，一般名詞カテゴリ番号を得る．形態素解析 (JTAG) の形態素辞書には，一般名詞カテゴリに対応した一般名詞カテゴリ番号が格納されている．例えば，図 4 の例の “クーラー” の一般名詞カテゴリ番号 901 と 969 は，それぞれ一般名詞カテゴリ「冷暖房具」と「電力機器」である．

[モダリティ情報の定義]
表記=いい加減にし{て|なさい|ろ}
↓(論理和の展開)
表記={いい加減にして|いい加減にしなさい|いい加減にしろ}
↓(形態素解析)
[モダリティ情報パターン]
{表記=いい加減, 表記=に, 表記=して|表記=いい加減, 表記=に, 表記=し, 表記=なさい|表記=いい加減, 表記=に, 表記=しろ}

図 3: パターン生成の例

[表記]	[品詞]	[読み]	[一般名詞カテゴリ番号]
クーラー	名詞:Katakana	クーラー	901:969
の	格助詞:の:連体	ノ	
無	形容詞語幹:L く連用:L で:アウオ段	ナ	
い	形容詞接尾辞:連体	イ	
中学校	名詞:助数詞	チュウガッコウ	405:367
の	格助詞:の:連体	ノ	
教室	名詞	キョウシツ	405:367:868
は	連用助詞:は:連用	ハ	
サウナ	名詞:Katakana	サウナ	374:436:868
です	判定詞:終止	デス	

図 4: 形態素解析結果の例

モダリティ情報を入力テキストから抽出するために，入力テキストに対する文分割処理と形態素解析によって得られた単語の属性情報とモダリティ情報パターンを用いて，単語の属性情報を用いたパターンマッチングを行う．このとき，入力テキストの単語の並びとモダリティ情報パターンに存在する単語の属性の並びについて，モダリティ情報パターンで指定されている単語の属性を用いて比較する．言い換えれば，入力テキストの単語の持つ表記，品詞，読み，一般名詞カテゴリ番号のいずれかの属性情報にモダリティ情報パターンで指定されている属性情報がマッチするかを調べていく．これをモダリティ情報パターンに存在するパターンのすべての単語の並びに対して行い，最後まですべてマッチする場合にパターンがマッチしたと見なす．

モダリティ情報パターンに存在するパターンの 1 つに論理和が用いられている場合は，論理和を構成するいずれかのパターンの単語の並びがすべてマッチするときに全体の論理和のパターンがマッチしたと見なす．さらに，先頭マッチ，末尾マッチ，否定の正規表現も考慮したパターンマッチングを行う．つまり，先頭マッチの記号が存在すれば文の先頭からパターンマッチングを行い，末尾マッチの記号が存在すれば文末とのパターンマッチングを行う．否定の記号が存在する場合は，否定の記号が付けられている属性情報以外であればマッチしたとする．

複数マッチするパターンが存在する場合に対しては，最長一致などいくつか方法が考えられるが，現在では処理速度の点から，あらかじめモダリティ情報に付けた優先度とパターンの登録順に従って，モダリティ情報のパターンが入力テキストにマッチするか調べ，マッチした時点で終了としている．モダリティ情報パターンがマッチした場合には，入力テキストのマッチした単語の並びの部分に該当するモダリティ情報のタグを付与する．モダリティ情報の優先度とパターンの登録順にパターンマッチングが行われるため，テキストに付与されるタグが，入れ子になったり交差したりすることはない．

4 評価

提案した手法を評価するために，電子メールのテキストを対象とした評価実験を行った．以下では，実験結果について報告し，実験結果に対する考察を述べる．

4.1 実験結果

評価実験のため用意したモダリティ情報の種類は，[喜び]，[怒り]，[哀しみ]，[楽しみ]，[驚き] の 5 つである．モダリティ情報の定義をコンパイルした後のモダリティ情報パターンの数は，図 2 に示した定義例を含めて，それぞれ 167，104，340，113，117 で計 828 個である．これらは論理和が用いられているパターンの数については，すべて展開したパターンの数として計算している．実験に用いた電子メールの数は，モダリティ情報の定義を記述するために用いたクローズドデータが 77 通，用いなかったオープンデータが 63 通である．これらはビジネスのメール，プライベートのメール，商用サービスのメーリングリストからの配信等が混在したものである．

電子メールのテキストから 5 種類のモダリティ情報を抽出した実験の結果を表 2 に示す．実験の結果から，モダリティ情報のすべてを対象としたときの F-measure の値がクローズドデータで 87.8，オープンデータで 79.0 となり，提案した手法が有効であ

表 2: モダリティ情報抽出の評価

	クローズドデータ					
	該当数	抽出数	正解数	再現率 (%)	適合率 (%)	F-measure
喜び	130	167	123	94.6	73.7	82.9
怒り	9	9	9	100	100	100
哀しみ	118	130	108	91.5	83.1	87.1
楽しみ	102	107	99	97.1	92.5	94.7
驚き	7	5	5	71.4	100	83.3
すべて	366	418	344	94.0	82.3	87.8
	オープンデータ					
	該当数	抽出数	正解数	再現率 (%)	適合率 (%)	F-measure
喜び	107	101	88	82.2	87.1	84.6
怒り	1	2	1	100	50.0	66.7
哀しみ	67	56	38	56.7	67.9	61.8
楽しみ	90	71	69	76.7	97.2	85.7
驚き	1	0	0	0	0	–
すべて	266	230	196	73.7	85.2	79.0

ることを検証した．

4.2 考察

クローズドデータとオープンデータとの比較での大きな違いは，適合率は維持しているものの，全体的に見て再現率が低下していることである．これはクローズドデータには存在しないものが多くあったためで，モダリティ情報の定義が足りないことを意味している．この問題は，さらに大きなデータセットを使ってモダリティ情報を定義することができれば，再現率は大幅に改善されると考えている．しかしながら，単純に表記でのみ定義を追加していくことは，その分だけパターンの数が増えて，処理速度に影響を及ぼす恐れがある．このため，品詞や一般名詞カテゴリを組み合わせて，効率的にモダリティ情報の定義を記述することが重要である．

一方，適合率を下げている原因には，次のようなものが考えられる．

- 形態素解析における同音異義語の単語選択の誤り
- 意図しない一般名詞カテゴリの単語へのマッチ
- 文脈に合わない単語の表記のマッチ

例えば，1 番目の具体例として，次のようなものがあげられる．“つまり” という表記は連用詞と名詞というふたつの異なる品詞の単語として形態素辞書に登録されており，名詞の “つまり” に一般名詞カテゴリの「遮断」が登録されている．テキストを正しく解析できず，本来ならば連用詞の方が選択されるところを名詞の方が選択されてしまった場合，モダリティ情報の定義で [哀しみ] として一般名詞カテゴリ「遮断」を記述しているならば，誤って “つまり” を [哀しみ] として抽出してしまう．また，2 番目の具体例としては，次のようなものがあげられる．“素敵なプレゼントが当たる” という文に存在する単語の動詞語幹 “当た” には「成功」という一般名詞カテゴリが登録されている．モダリティ情報の定義に [喜び] として一般名詞カテゴリ「成功」が記述されていれば，[喜び] としてこの単語が抽出される．しかしながら，“サービス再開に当たって” という文では，同じ動詞語幹 “当た” が用いられるため，誤って [喜び] というモダリティ情報を抽出してしまう．さらに，3 番目の具体例としては，次のようなものがあげられる．“昨日の帰りは渋滞にはまって大変だった．” という文には，“はまって” という表記が存在する．この表記がモダリティ情報の定義に [哀しみ] として記述されている場合，“最近すっかり犬好きで犬にはまっている私” という文での “はまって” に誤って [哀しみ] というモダリティ情報を抽出してしまう．

そこで，今後の課題として，適合率をさらに上げるためには，以下を対応策を検討する必要がある．

- 同音異義語に対する形態素解析の精度の向上
- 一般名詞カテゴリと表記や品詞との組み合わせの利用
- 前後の文字列に対する一般名詞カテゴリの利用

例えば，2 番目の対応策は，先にあげた具体例の問題に対応するために，一般名詞カテゴリを指定する単語の直前に品詞が名詞と表記が “が” であるような 2 つの単語が共起するようにモダリティ情報の定義を記述することである．評価実験のモダリティ情報の定義において，図 2 に示したように，一般名詞カテゴリと表記の組合わせを一部用いているが，このような記述をさらに増やすことが必要だと考えている．また，3 番目の対応策は，先にあげた具体例の問題に対応するために，“はまって” という表記の前に一般名詞カテゴリ「停滞」や「混乱」を持つ単語を同時に定義しておくことである．これらの対策により，適合率は大きく向上するであろうと考えている．

5 関連研究

本稿で提案したモダリティ情報の抽出手法に関連する研究を以下にあげる．諸橋ら [4] は，コールセンターへの問い合わせ事例の分析を行うために，特定の単語と述語などに着目することによって，モダリティ情報を抽出している．モダリティ情報の抽出のために構文情報が使われているのに対して，本稿で提案した手法は，各単語の持つ属性情報によるパターンマッチングを用いている．また，乾ら [3] は，自由回答アンケートを分類する際に重要な働きを担うモダリティについて分析している．モダリティは文末表現に現れて，助詞や助動詞相当の語の他に，動詞句や形容詞句を伴うと報告している．しかしながら，モダリティ情報の自動抽出にまでは触れていない．

6 おわりに

本稿では，単語の持つ属性情報を用いたパターンマッチングにより，テキストからモダリティ情報を抽出する手法を提案した．電子メールのテキストを対象とした実験結果の F-measure の値から，提案手法は良好な精度が得られることを検証した．また，今後の課題として，さらにモダリティ情報の抽出の精度を高めるために，モダリティ情報の定義を増やすとともに，単語の持つ属性情報をうまく組み合わせることによって緻密にモダリティ情報の定義を記述していく．現在，本稿で提案した手法と，音韻毎のピッチ周波数や振幅値（パワー）などの韻律特性を韻律制御コマンドとして記述できる言語である MSCL[5] とを結合させせることにより，入力されたテキストからモダリティ情報に応じた抑揚付きの合成音声を出力するソフトウェアを開発している．このソフトウェアによって，任意のテキストから適切な抑揚付きの合成音声が生成できるので，様々な応用に利用することが可能である．

参考文献

[1] Fuchi,T. and Takagi,S.: Japanese Morphological Analyzer Using Word Co-occurrence, *Proc. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics(COLING-ACL'98)*, pp.409–413, 1998.

[2] 池原，宮崎，白井，横尾，中岩，小倉，大山，林：日本語語彙体系 1 意味体系，岩波書店，1997.

[3] 乾，内元，井佐原：モダリティ分析に基づく自由回答アンケートの分類，言語処理学会第 4 回年次大会発表論文集, pp.540–543, 1998.

[4] 諸橋，那須川：テキストマイニング：膨大な文書データからの知識獲得—意図の認識—，情報処理学会第 57 回全国大会 5K-03, 1998.

[5] 水野，中嶌：合成音声制御のための階層型記述言語 MSCL，人工知能学会研究会資料 SIG-SLUD-9703, pp.7–12, 1998.

[6] *Proc. 6th Message Understanding Conference* (*MUC-6*), Morgan Kaufmann, 1995.

[7] 山梨正明：発話行為〈新英文法選書 12〉，大修館書店, 1986.