単語の共起分布を用いた文末モダリティの自動推定

Identifing Sentence End Modalities with Co-occurence of Words

中村 紘規† 玉城 伸仁† 河原 大輔† 黒橋 禎夫†
Hironori Nakamura Nobuhito Tamaki Daisuke Kawahara Sadao Kurohashi

1 はじめに

一般に、文章に記述される情報は単純な命題のみではなく、そこにはモダリティと呼ばれる、筆者・発話者の態度や判断も記述される。たとえば以下の文には、それぞれ次のようなモダリティがある。

(1) 頼むから最後までやってください。
→ 聞き手に「最後までやる」ことを働きかける発話者の態度 (勧誘)

(2) たぶん彼は来るだろう。
→「彼は来る」ことがどれくらい成立するかという発話者の判断 (推量)

文章からこれらのモダリティを解析し、認識することは質問応答や含意認識などの応用に重要である。従来は人手で書いた規則を用い、モダリティ表現に対してタグ付与を行うことでモダリティ解析を行ってきたが、人手で書ける規則には網羅性の観点で限界がある。

そこで本研究では、規則の網羅性を向上させるため、「意味が似たような語は同じ語と共起しやすい」という分布類似度 [7] の考えに基づき、モダリティを自動推定する手法を提案する。特に、モダリティは副詞、助詞などの機能的な語と結びつきが強いため、「同じモダリティを持つ語は同じ機能的な語と共起しやすい」と考えることができる。本研究では、この共起の傾向を文末ベクトル、モダリティベクトルと呼ぶベクトルで表し、類似度計算を行うことでモダリティの推定を行う。

本論文の構成を以下に示す。まず第 2 章でモダリティの自動推定に関連する研究について述べる。次に第 3 章で本研究で用いる文末ベクトルとモダリティベクトルを用いたモダリティ推定方法について述べる。第 4 章で実験設定および実験結果について述べ、最後に第 5 章でまとめを行う。

†京都大学, Kyoto University

2 関連研究

Medlock ら [1] は部分的教師あり学習を用いた、推量表現の分類を行った。文が与えられたとき、それが推量か否かを分ける 2 値分類のタスクであり、Medlock らはある単語の出現が文が推量であるのかそうでないのかを決定すると考えた。ある単語が出現したとき、それを含む文が推量を表す確率を定義して、その確率を部分的教師あり学習モデルで学習することで、分類を行った。結果、SVM と比較して良い精度で分類を行うことができたと述べている。

江口ら [4] は、〈態度表明者、時制、仮想、態度、真偽判断、価値判断、焦点〉の 7 項目から構成される独自の拡張モダリティタグ体系を構築し、さらにこの拡張モダリティを解析するシステムの開発を行った。この解析システムでは拡張モダリティタグが〈態度〉を中心に強い依存関係にあることから、条件付確率場を利用しモダリティの推定を行っている。素性には中心文節とその前後の文節および係り先の文節の形態素情報、機能語列、人手による語彙統語パタン、意志動詞の有無、分類語彙表における「人間活動の主体」に属する形態素の有無、日本語機能表現辞書つつじの意味コード、モダリティ解析用手がかり辞書の 7 種類を用いている。実験として、項目間、事象間の依存関係を考慮することの有用性および各素性の有効性を評価を行い、依存関係を考慮することは有用であり、素性は人手による語彙統語パタンが有効であったと述べている。しかしながら、人手による語彙統語パタンは作成に膨大な手間・時間がかかるため、これに代わる素性の検討が必要であると述べている。また、意志動詞の有無は重要な指標となりそうだが、今はうまく活用できていないとも述べている。

3 提案手法

本章ではモダリティの自動推定方法について述べる。本研究では、文末ベクトル、モダリティベクトルと呼ぶ

表 1: 機能表現となる形態素の品詞および品詞細分類

助詞	接続詞	形式名詞	助動詞
判定詞	連体詞	感動詞	副詞
時相名詞	副詞的名詞	指示詞	接頭辞
名詞性述語接尾辞		形容詞性述語接尾辞	
形容詞性名詞接尾辞		動詞性接尾辞	
動詞性活用語尾		形容詞性活用語尾	

ベクトルを作成し、これらの類似度を計算することでモダリティの推定を行う。

3.1 文末ベクトルの作成

文末ベクトルとは、文末表現と機能表現の共起の傾向を表すベクトルである。文末ベクトルの作成は以下の手順で行う。

1. コーパス中のすべての文について、その文末表現を汎化した文末パタンに変換する
2. 文末パタンごとに、様々な機能表現との自己相互情報量を計算し、それを並べたベクトルを作成する
3. 作成したベクトルを特異値分解によって 800 次元まで圧縮したものを文末ベクトルとする

文末パタンはコーパスに含まれる文の文末文節を抽出し、その核となる動詞、形容詞、名詞を汎化したもので、データスパースネスを防ぐ目的がある。例えば「日本人っぽい」という表現は「【名詞】っぽい」という文末パタンになる。このように抽出した文末パタンと共起する機能表現との間の自己相互情報量を計算し、並べて正規化してさらにベクトルの次元を圧縮したものが文末ベクトルである。ここで機能表現とは、形態素解析器 JUMAN[6] の解析によって、品詞が表 1 のように分類される形態素のことを指し、およそ 5000 個の形態素がある。また、次元圧縮には、計算量の削減、機能表現の冗長性の削減といった目的がある。

自己相互情報量は、低頻度の要素が極端な値をとることを避けるために以下のような補完自己相互情報量 $\log'$ と適当な補完量 c を用いて計算する [3]。ここで b は文末パタン、w は機能表現であり、f_b は文末パタンの出現頻度、f_w は機能表現の出現頻度、$f_{b,w}$ は文末パタンと機能表現の共起頻度、f_{all} はすべての組の共起頻度の合計である。

表 2: 文末ベクトルの例

機能表現番号	【名詞】がある みたい	【名詞】っぽい	【動詞】ましょう
1	35.55	28.92	-0.22
2	-16.55	-17.08	1.90
3	-9.87	-14.44	-6.99
...	...	...	
10	5.92	7.99	0.05
...	...	...	
100	1.28	-3.02	-0.18
...	...	...	
200	1.00	0.54	-0.56
...	...	...	
800	0.33	-0.22	0.03

$$\mathrm{cPMI}(b,w) = \log' \frac{f_{b,w} \cdot f_{all} + c}{f_b \cdot f_w + c} \tag{1}$$

ただし、

$$\begin{aligned} \log'(x) &= \log_2\{x + (1-x) \cdot \frac{r}{1-r}\} \\ c &= \sqrt{f_b \cdot f_{all}} \\ r &= f_b / f_{all} \end{aligned} \tag{2}$$

である。

例として「【名詞】があるみたい」、「【名詞】っぽい」、「【動詞】ましょう」という文末パタンの文末ベクトルを表 2 に示す。同じモダリティを持つ「【名詞】があるみたい」、「【名詞】っぽい」は似たようなベクトルになっている一方、異なるモダリティを持つ「【動詞】ましょう」のベクトルは大きく異なっているのが見て取れる。ただ、次元圧縮を行った関係で、各機能表現番号がどのような機能表現に対応するかが判別できなくなり、どのような機能表現が特徴的であるのかを調べることはできなかった。

本研究では、コーパスとして Web コーパス 16 億文を用い、約 20 万個の文末パタンについて文末ベクトルを作成した。

3.2 モダリティベクトルの作成

モダリティベクトルとは、モダリティと機能表現の共起の傾向を表すベクトルである。基本的にはシードとな

るルールを用いて文末ベクトルを足し合わせたものであり、作成は以下の手順で行う。

1. コーパスから文を獲得し、ルールによりモダリティを付与する
2. モダリティごとに文を分け、その文に対応する文末ベクトルをモダリティごとに足し合わせてベクトルを作る

モダリティベクトルは推定を行うモダリティごとに作成する。本研究では仁田 [8] の分類に基づいたモダリティの推定を行う。図 1 に仁田の分類を示す。仁田はモダリティを大きく「発話・伝達のモダリティ」と「事態めあてのモダリティ」の 2 種に分け、さらにこれらを細かく分ける階層的な分類を行っている。本研究では分類のゆれが少なく推定が行いやすいため、図 1 の分類のうち、最も大きな「発話・伝達のモダリティ」と「事態めあてのモダリティ」の推定を行う。

モダリティの付与は構文解析器 KNP[5] を用いて行う。KNP には、文字列マッチによりモダリティを付与する約 100 個のルールがある。しかし、KNP の付与するモダリティは粒度が細かいので、表 3 に示すように KNP のモダリティと仁田のモダリティの対応付けを行う。ただし、KNP の「意志」および「疑問」については、仁田の分類のうちどちらにも出現するので、今回は扱わないことにした。また、「〜してはならないだろう (禁止/認識-推量)」のように、KNP が複数のモダリティを付与する文末パタンのうち仁田の分類に対応させたときに曖昧性が生じる文末パタンについても、今回は扱わないことにした。

そして、モダリティの付与された文末パタンをモダリティごとに分け、足し合わせることでモダリティベクトルを作成する。

モダリティベクトルの例を表 4 に示す。比較しやすいよう、文末ベクトルを足し合わせたあと正規化してある。文末ベクトルと同様、次元圧縮を行った関係で、どのような機能表現が特徴的であるかを調べることはできなかった。

3.3 モダリティの推定

モダリティの推定は k-Nearest Neighbor 法とモダリティベクトル法の 2 種類の方法で行った。

k-Nearest Neighbor 法は、モダリティを求めたい文末ベクトルに近い k 個の文末ベクトルのモダリティから、求めるモダリティを決定するものである。本研究では、

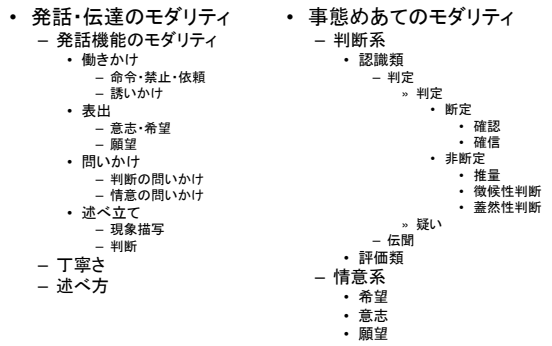

図 1: 仁田の分類

表 3: KNP のモダリティと仁田のモダリティの対応付け

KNP のモダリティ	仁田のモダリティ
勧誘	発話・伝達
禁止	発話・伝達
命令	発話・伝達
依頼 (A・B)	発話・伝達
評価 (強・弱)	事態めあて
認識 (推量・証拠性・蓋然性)	事態めあて
疑問	———
意志	———

ベクトルの類似度は以下のコサイン類似度で計算する。$\mathbf{v}_{b_1}$、$\mathbf{v}_{b_2}$ は文末パタン b_1、b_2 の文末ベクトルである。

$$\begin{aligned} \mathrm{sim}(\mathbf{v}_{b_1}, \mathbf{v}_{b_2}) &= \cos(\mathbf{v}_{b_1}, \mathbf{v}_{b_2}) \\ &= \frac{\mathbf{v}_{b_1} \cdot \mathbf{v}_{b_2}}{||\mathbf{v}_{b_1}|| \cdot ||\mathbf{v}_{b_2}||} \end{aligned} \quad (3)$$

また、k としては 1 及び 10 を選択した。k=1 の場合は最も類似度が高かった文末ベクトルのモダリティを推定結果とする。k=10 の場合は、上位 10 件でモダリティごとにベクトルの類似度を足しあわせたものをスコアとし、スコアの高いモダリティを推定結果とする。

モダリティベクトル法は、モダリティベクトルと文末ベクトルの類似度計算によりモダリティを推定する方法で、モダリティを求めたい文末ベクトルに最も近いモダリティベクトルのモダリティを、求めるモダリティとするものである。類似度計算は k- Nearest Neighbor 法と同様、コサイン類似度を使用する。

モダリティ推定の例を表 5 から表 7 に示す。「【名詞】っぽい」という文末パタンについて推定した例で、正解は仁田のモダリティでは「事態めあてのモダリティ」、KNP のモダリティでは「認識-証拠性」である。1-

表 4: モダリティベクトル (正規化済) の例

機能表現番号	事態	発話
1	0.834	0.790
2	0.479	0.241
3	-0.023	0.002
···	···	···
10	0.038	-0.029
···	···	···
100	-0.007	-0.012
···	···	···
200	0.009	0.019
···	···	···
800	0.006	0.009

表 5: モダリティの推定例 (1-NN)

推定するパタン	【名詞】っぽい	
正解	事態 (認識-証拠性)	
似ている文末パタン	モダリティ	類似度
【名詞】があるみたい	事態	0.549
推定結果	事態	

NN、10-NN、MV はそれぞれ 1-Nearest Neighbor 法、10-Nearest Neighbor 法、モダリティベクトル法を表し、事態、発話はそれぞれ事態めあてのモダリティ、発話・伝達のモダリティを表す。

4 実験

4.1 設定

KNP がモダリティを付与できる文末パタン 100 パタン (closed)、KNP がモダリティを付与できない文末パタン 100 パタン (open) についてモダリティを推定する実験を行った。100 パタンの内訳は、closed は KNP の付与するモダリティのうち「意志」と「疑問」を除いた 10 種類のモダリティから、各 10 パタンずつ文末パタンを取り出したものであり、open は正解となるモダリティを KNP の付与するモダリティに照らし合わせて、10 種類が各 10 個ずつになるように選んだものである。

これら 200 パタンについて、前節で述べた 2 種類の方法を用いてモダリティの推定を行う。

表 6: モダリティの推定例 (10-NN)

推定するパタン	【名詞】っぽい	
正解	事態 (認識-証拠性)	
似ている文末パタン	モダリティ	類似度
【名詞】があるみたい	事態	0.549
【名詞】があるような	事態	0.524
【名詞】があったような	事態	0.482
【名詞】があるようで	事態	0.481
【名詞】がするな	発話	0.479
【名詞】があるみたいです	事態	0.479
【名詞】になってるらしい	事態	0.461
【名詞】があるみたいですね	事態	0.457
【名詞】になってるな	発話	0.456
【名詞】もあるみたい	事態	0.452
	事態合計	3.885
	発話合計	0.935
推定結果	事態	

表 7: モダリティの推定例 (MV)

推定するパタン	【名詞】っぽい
正解	事態 (認識-証拠性)
モダリティベクトル	類似度
事態	0.374
発話	0.341
推定結果	事態

4.2 結果

表 8 に各手法の精度を比較した結果を示す。1-NN、10-NN、MV はそれぞれ 1-Nearest Neighbor 法、10-Nearest Neighbor 法、モダリティベクトル法を表す。

結果、モダリティベクトル法が open、closed とも Nearest Neighbor 法を大きく上回っていることがわかる。これは、モダリティが付与されていない表現が大半を占め、モダリティを推定したい表現に最も似ているのが、モダリティが付与されるべきだがされていない表現となっている場合があるのが精度を下げる原因になっていると考えられる。他の原因として、文末ベクトルにはモダリティ特有の機能表現の情報だけでなく、口調や文体に特有の機能表現の情報も含まれ、これにより、モダリティよりも口調が似ている表現の類似度が高くなったことも考えられる。モダリティベクトル法では、同じ

表 8: Nearest Neighbor 法と提案手法の推定精度

	1-NN	10-NN	MV
closed	81.0	79.0	85.0
open	75.0	71.0	86.0

表 9: 新たに獲得されたモダリティ表現例

モダリティ表現	実表現例	モダリティ
【形容詞】だろう	いいだろう	事態
【動詞】ぞ	出るぞ	発話
【名詞】やろ	いかんやろ	発話
【名詞】せよ	ＧＥＴせよ	発話
【動詞】ネ	来てネ	発話
【動詞】ねばならない	待たねばならない	事態
【形容詞】かもしれない	面白いかもしれない	事態
【動詞】と思われる	原因と思われる	事態

上 4 つは、提案手法のみで獲得された表現の例。
下 4 つは、提案手法および NN 法両方で獲得された表現の例。

モダリティを持つ様々な口調、文体の文末ベクトルを足しあわせているので、結果としてこれらの影響が少なくなり、精度に差が出たと考えられる。

また、例として新たに獲得されたモダリティ表現を表 9 に示す。表の上 4 つは NN 法では獲得できなかったが、提案手法では獲得できた表現の例、下 4 つは NN 法、提案手法両方で獲得できた表現の例である。

4.3 考察

各手法のモダリティごとの正解率を表 10 に示す。事態は「事態めあてのモダリティ」、発話は「発話・伝達のモダリティ」であり、表中の数字は精度 (%) である。この結果から、Nearest Neighbor 法は「事態めあてのモダリティ」を当てやすく、提案手法は「発話・伝達のモダリティ」を当てやすいことが見て取れる。以下、各手法の誤りについて具体的に見る。

Nearest Neighbor 法での誤りの多くは、類似度上位に適切な候補がなく、低い類似度で推定してしまっているパタンである。おおよそ、正しく推定できているものは 0.5〜0.7 程度の類似度の文末ベクトルを用いてモダリティの推定を行っているのだが、誤りの多くは類似度 0.3 前後の文末ベクトルを用いてモダリティを推定している。これは前節でも述べたように、モダリティが付与

表 10: 各種法のモダリティごとの正解率

	1-NN		10-NN		提案手法	
	事態	発話	事態	発話	事態	発話
closed	98.0	64.0	100	58.0	78.0	92.0
open	96.0	54.0	96.0	46.0	76.0	96.0

されている表現が少ないのが一因であると考えられる。これを改善するには、類似度で足切りをした上で、何度も推定を繰り返し、徐々にモダリティの付与されている表現を増やしていくといった手法が考えられる。

ただ、それにも関わらず、「事態めあてのモダリティ」で精度が下がらないのは、これと共起する機能表現は口調や細分類によらずほぼ一定であるからではないかと思われる。共起する機能表現が口調や細分類に大きく影響されると思われる「発話・伝達のモダリティ」と比べて、データの少なさが推定に影響することがないのではないように思われる。

提案手法での誤りは、KNP の分類における「評価」の表現の誤りが大半を占めている。「評価」は「〜するべきである」や「〜したほうがよい」に代表されるモダリティであるが、これらは、「発話・伝達のモダリティ」に含まれる、「命令」や「勧誘」と似たような使われ方をされる場合があるのが原因ではないかと考えられる。これの改善方法は現時点では思いつかないが、益岡 [2] の分類では「評価のモダリティ」がその他のモダリティとして別個に扱われているのを鑑みるに、「評価」を「事態めあてのモダリティ」から外してみるのもよいかもしれない。

5 まとめ

本論文では、少数のシードからモダリティを自動推定する方法として、モダリティと機能表現の共起に着目し、ベクトル演算によって推定する手法について提案を行った。提案手法では仁田の分類のうち大分類の推定を行い、結果 85〜86 %の精度で推定を行うことができることを示した。さらに、他手法と比較して、高い精度で推定できることを示した。

今後は、4.3 節で述べた、繰り返し推定を行う手法の実装や、「評価」の分類について考えていきたい。同時に、現在扱いを保留にしている「意志」や「疑問」といったモダリティについてもどう分類するのか考えていく必要があり、これと並行してさらに細かい分類の推定手法について検討していきたい。

参考文献

[1] Ben Medlock and Ted Briscor. Weakly supervised learning for hedge classification in scientific literature. In *the 45th Annual Meeting of the ACL*, 2007.

[2] 益岡隆志. 日本語モダリティ探究. くろしお出版, 2007.

[3] 玉城伸仁. 会話文生成を指向する日本語言い換えの研究. Master's thesis, 京都大学大学院 情報学研究科, 2009.

[4] 江口萌, 松吉俊, 佐尾ちとせ, 乾健太郎, 松本裕治. モダリティ、真偽情報、価値情報を統合した拡張モダリティ解析. 言語処理学会第 16 回年次大会, pp. 852–855, 2010.

[5] 黒橋禎夫. 日本語構文解析システム KNP version 3.0 使用説明書. 京都大学大学院 情報学研究科, 9 2009.

[6] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version 6.0 使用説明書. 京都大学大学院 情報学研究科, 9 2009.

[7] 柴田知秀, 黒橋禎夫. 超大規模ウェブコーパスを用いた分布類似度計算. 言語処理学会第 15 回年次大会, pp. 705–708, 2009.

[8] 仁田義雄. 日本語のモダリティとその周辺. ひつじ書房, 2009.