Vol.2019-NL-240 No.6
2019/6/13

レビュー文章集合を用いたマイクロドメインのための概念階層オントロジー構築

谷江 博昭[1,a)]　三澤 賢祐　大内 啓樹[2,3,b)]

概要：本稿では，あるエンティティに関するテキスト集合から効率的に知識ベースを構築することを目指す．より具体的には，施設の評判などを記述したレビュー文章集合から，その施設にある設備やサービスを抽出し，それらの情報を体系化する．構築した知識ベースは，検索システムや QA システムへ活用する．提案手法では，レビュー文章集合から，知識ベースに必要なドメイン個別のオントロジーを構築する．特に，解析したいテキストで学習した単語埋め込みを利用することによって，解析対象のドメインに適したオントロジーを得る．本稿では，オントロジー の質の定量的・定性的分析を行い，単語埋め込みの違いによる効果を検証する．

1. はじめに

インターネットの隆盛に伴い，ウェブ上における個人の情報発信の機会が増加し，「Amazon.com[*1]」や「じゃらん net[*2]」などのサイトに多くのカスタマーレビューが投稿されている．表 1 は「じゃらん net」に寄せられたレビュー文章の例を示している．対象施設に対する感想や，その施設が提供するサービスやグッズなどの情報が述べられている．このようなレビュー文章から対象施設に関するより詳細な情報を抽出・構造化することにより，高度な検索システムなどの応用アプリケーションに有用な知識ベースとして利用可能となる．検索で利用可能な知識ベースの例を図 1 に示す．本稿で想定する知識ベースは，対象施設に紐づく実在物であるエンティティと，そのエンティティを抽象化した概念（クラス），概念同士の上位・下位関係を定義する概念階層オントロジーからなる．

オントロジーを自動構築する研究は古くからあるが [6][7]，それらの多くは辞書や Wikipedia などに含まれる一般的な定義情報を基にしている．しかし，実世界においては概念同士の関係は利用される場面によって異なり，単純に定義することは難しい．例えば，宿泊施設は風呂や大浴場をエンティティに持ち，一般的には，風呂は大浴場の上位概念と定義できる．しかし，ミクロに見るとその関係は異なる．

温泉施設 A に対するレビュー
大浴場・**露天風呂**があり、**カフェ**や**レストラン**を揃えた温泉施設です。
熊笹うどんはモチモチで美味しく、こちらの**ソフトクリーム**もオススメです。
食事処もリーズナブルでよい。

表 1　レビュー文章の例．太字がエンティティを表す．

宿泊施設のうち，温泉旅館では“風呂”は大浴場とほぼ同義で扱われるが，ホテルでは“風呂”は宿泊部屋にある風呂を示す概念であり，大浴場とは別の概念として扱われることが多い．このような違いは些細ではあるが，実サービスにおいては極めて重要な違いとなる．一般的な定義から考えた場合，これらの違いは埋もれてしまい，また仮に定義が出来ても単一のオントロジーで全てを表現すると複雑で実利用が困難なものとなる．本稿では，概念の意味およびその関係をユニークに表現できる集合を「**マイクロドメイン**」と定義する．マイクロドメイン毎に，概念と関係を定義するオントロジーができれば，実サービスにも応用可能な有用な知識ベースとなる．

本稿では，あるマイクロドメインのレビュー文章集合を入力に，当該マイクロドメインに特化したオントロジーを構築する手法を提案する．当該手法では，レビュー文章集合から階層クラスタを自動構築し，それを元に概念階層オントロジーを生成する．当該手法で構築したオントロジーの質を評価するため，人手によって構築した正解オントロジーを用意した．正解オントロジーと自動構築した階層クラスタを比較することによって，正解オントロジーをどれだけ自動で再現できているかを評価する．結果として，正

[1] 株式会社リクルート Megagon Labs, Tokyo, Japan
[2] 理化学研究所 AIP センター
[3] 東北大学
[a)] hiroaki_t@megagon.ai
[b)] hiroki.ouchi@riken.jp
[*1] https://www.amazon.com/
[*2] https://www.jalan.net/

ⓒ 2019 Information Processing Society of Japan

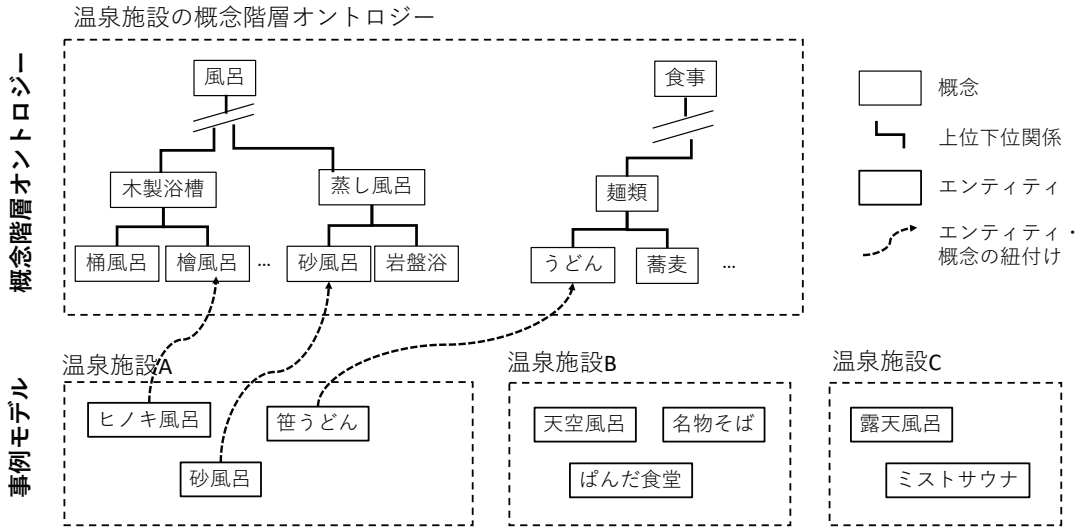

図 1 知識ベースの構成

解オントロジーに近い質の高い階層クラスタを構築可能であることがわかった．

本稿の主要な貢献は以下の 3 点である．

- 対象マイクロドメインに特化した概念階層オントロジーの構築手法の提案
- 対象マイクロドメインのレビュー文章に基づくアノテーションデータと正解階層クラスタデータ
- 自動構築した階層クラスタの質の定量的・定性的分析

2. 概念階層オントロジー

本節では，本稿で構築する概念階層オントロジーについて詳述する．

2.1 概念階層オントロジーの概要と応用例

概念階層オントロジーの例

図 1 は本稿で構築する概念階層オントロジーを示している．これは，温泉施設に関するレビュー文章から，その施設が提供するサービスやグッズに関する言及 (メンション) を抽出し，それをもとに構築したものである．

このオントロジーは階層構造を持ち，ノード (末端ノードおよび中間ノード) と，ノード間を結ぶエッジからなる．ノードは「**概念 (クラス)**」と呼ばれるものであり，温泉施設が持つエンティティは，オントロジー内のいずれかの概念と紐付けられる．エッジは概念間の上位下位関係を表し，エッジの上部にある概念が上位概念，下部にある概念が下位概念である．

検索システムへの応用

レビュー文章から抽出した各施設のエンティティ情報をオントロジーと紐付け，図 1 のような知識ベースを構築する．知識ベースを用いて，例えば「蒸し風呂のある温泉施設」を検索する際に，概念の上位下位関係を踏まえ「砂風呂」を持つ温泉施設 A を検索することが出来る．

2.2 技術的課題

前述した概念階層オントロジーを構築するには，以下の二つの技術的課題がある．

- オントロジーに登録する概念をどのように獲得すべきか？
- 概念間の階層関係をどのように獲得すべきか？

本稿では，これら二つの技術的課題に取り組む．

3. 手法

本節では，概念階層オントロジー構築のための Human-Machine ハイブリッド手法を詳述する．

3.1 概念階層オントロジー構築プロセスの概要

概念階層オントロジー構築プロセスの概要を以下に示す．

(1) 対象ドメインのレビュー文章集合からエンティティに関するメンションを抽出する [機械]
(2) 抽出したメンションをもとに階層クラスタリングを行う [機械]
(3) 自動構築した階層クラスタをもとに概念階層オントロジーを構築する [人間]

(1) と (2) は機械で自動的に行うプロセスであり，(3) は人手によるプロセスである．(1) と (2) を高精度で自動化できれば，(3) の人手による作業は短時間で完了することができる．図 2 に上記プロセスの概略を示す．以降の節で各プロセスについて詳述する．

3.2 エンティティのメンション抽出

レビュー文書から，対象施設に紐づくエンティティに関

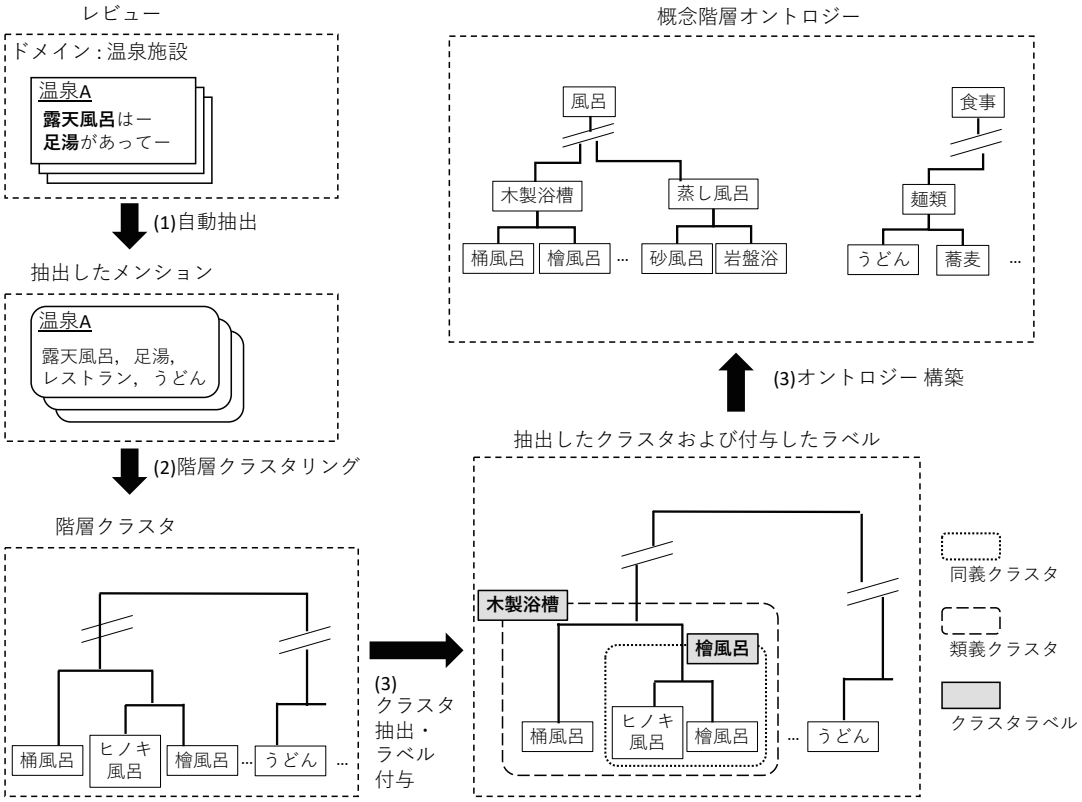

図 2 概念階層オントロジーの構築の概略.

する記述を抽出する．物理的な設備だけではなく，対象施設が所有する設備・備品，提供するサービス・食事・グッズもエンティティと捉え，抽出の対象とする．エンティティを指し示す記述をメンションと呼ぶ．

メンションの抽出は系列ラベリング問題として解く．以下に例を示す．

入力文：	露天	風呂	は	最高	でした	
正解ラベル：	B	I	O	O	O	

入力としてレビュー文書中の各文が与えられ，各単語に対して BIO ラベルを付与する．上記の例では，B と I のラベルが付与された「露天風呂」がメンションとして抽出される．

定式的には，レビュー文書の n 単語からなる各文 $X = \{w_1, w_2, \cdots w_n\}$ を入力として，各単語に対するラベル $Y = \{y_1, y_2, \cdots, y_n\}$ を予測する．

$$\textbf{入力}：X = \{w_1, w_2, \cdots w_n\}$$
$$\textbf{出力}：Y = \{y_1, y_2, \cdots, y_n\}$$

メンションの抽出には，BiLSTM-CRF [2] を利用する．BiLSTM-CRF は，複数層の双方向 LSTM レイヤーの最後に CRF を連結したアルゴリズムである．CRF アルゴリズムにより出力ラベル間のラベル依存性をモデル化できるため，BiLSTM に比べ系列ラベリング問題に適しているとされている．また，Lample らの実験では事前学習済みの単語埋め込みモデルの利用，Dropout の導入により性能向上が示されている．本研究では，Lample らの研究に従い，事前学習済みの単語埋め込みベクトルを BiLSTM-CRF の入力とし，LSTM に Dropout を導入する．

3.3 メンションの階層クラスタリング

抽出したメンションから，階層クラスタを生成する．生成する階層クラスタは 3.4 節で概念階層オントロジーの構築に利用するため，以下のクラスタを含む必要がある．

- 表記揺れや同義語のメンションで構成されるクラスタ
- 類義関係のメンション (あるいはクラスタ) で構成されるクラスタ
- 上位下位関係にあるメンション (あるいはクラスタ) とメンション (あるいはクラスタ) のペア

定式的には，抽出されたメンションの集合 $M = \{m_1, m_2, \cdots m_n\}$ から，クラスタの集合 $C = \{c_1, c_2, \cdots c_j\}$ を探索する．クラスタ集合 C 内の要素である c_i は，下記のいずれかのクラスタに該当する．

同義クラスタ：クラスタの要素は $m \in M$ で構成され，クラスタ内の要素に対して同義関係が成り立つ

類義クラスタ：クラスタの要素は $m \in M$ あるいは $c \in C$ で構成され，クラスタ内の要素に対して類義関係[*3]が成り立つ

上位下位クラスタ (ペア)：2 つの要素からなるクラスタ (ペア) であり，要素 $\{x, y\}$ はそれぞれ $m \in M$ あるいは $c \in C$ である．x は y の上位概念[*4]にあたる

本手法では，抽出されたメンションに対して，単語分散表現間のコサイン類似度を用いて，階層クラスタを構成する．クラスタ間の距離測定方法にはウォード法を用いる．解析対象のマイクロドメインに特化した単語埋め込みを用いることで，対象マイクロドメインにおける同義・類義・上位下位クラスタを多く含む階層クラスタを得る．

なお，クラスタリングの対象とするメンションは頻度で足切りを行なった．足切りは，複数の施設で複数回登場することを条件とし，これに満たないメンションを除外した．

3.4 概念階層オントロジーの構築

獲得した階層クラスタから概念階層オントロジーを再構成する．階層クラスタには対象マイクロドメインにおける単語の類似度に基づくクラスタが含まれるが，単語間の関係性（同義・類義・上位下位）は含まれない．図 2 に示すように，階層クラスタから同義・類義・上位下位クラスタを抜き出し，ラベルを付与することで，概念の上位下位関係を規定する概念階層オントロジーを構築する．

具体的には，以下の作業を行う．

(i) 階層クラスタ内から同義クラスタに該当するクラスタを抜き出す．同クラスタ内のメンションを 1 つの概念にまとめてオントロジーの末端ノードとして追加する．

(ii) 階層クラスタ内から類義クラスタに該当するクラスタを抜き出し，同クラスタに人手でラベルを付ける．同クラスタ内のメンションをそれぞれ概念としてオントロジーの末端ノードに追加する．クラスタにつけたラベルは，中間ノードとしてオントロジーに追加し，上位下位関係のエッジをつける．

(iii) 階層クラスタ内から上位下位クラスタに該当するクラスタを抜き出し，下位にあたるメンションを概念としてオントロジーの末端ノードに追加する．上位にあたるメンションは，中間ノードとしてオントロジーに追加し，上位下位関係のエッジをつける．

なお，上記の (i)(ii)(iii) の全てにおいて，クラスタ内に誤りが存在する場合には，人手で修正を行なった上でオントロジーへ反映させる．

階層クラスタリングで同義・類義・上位下位が適切にクラスタリングされている場合，上記の作業が容易に行える．

[*3] 共通の上位概念を持つ要素同士の関係を類義関係と呼ぶ
[*4] y is a x の関係が成り立つ

３００円で/足湯/E が楽しめて、...
始めに/内湯/E にはいり、その後/露天風呂/E にも入りました
/脱衣所/E で/鍵付きロッカー/E を使いたい場合は、...
湯上りの/ビール/E は最高。

表 2 アノテーションの例 (/,/E で囲まれた箇所がアノテーションされたメンション)

対象のレビュー件数	6,000
対象施設の数	487
アノテーションされたメンションの数	15,546
メンションの異なり数	2,941
足切り後のメンションの異なり数	130

表 3 正解データの統計量.

4. データ

対象のマイクロドメインを温泉施設とし，温泉施設のレビュー文章集合から，人手により正解クラスタおよび概念階層オントロジーを作成した．なお，利用するレビューは全て「じゃらん net」のレビューを用いた．

正解データの作成は次のように行なった．レビュー文章集合として，温泉施設に関するレビューから抽出した 6,000 件を利用する．抽出したレビューに対して，ターゲット施設が所有するエンティティへのメンションを表 2 のようにアノテーションする．アノテーションされたメンションから 3.3 節と同一のルールにより，メンションの足切りを行う．残ったメンションから，温泉施設において同義関係，類義関係，上位下位関係と見なせるメンションの集合を作り，それらの集合にラベルを付与する．ここで得られたクラスタを**正解クラスタ**とし，自動構築された階層クラスタの比較評価に用いる．なお，正解クラスタから正解オントロジーの作成は 3.4 節と同様の手法により行う．正解クラスタには概念階層オントロジー構築に必要な情報が含まれていることから，自動構築された階層クラスタと正解クラスタを比較評価することで，正解オントロジーに対する評価とすることができる．対象データの統計量と作成した正解クラスタおよびオントロジーのサンプルをそれぞれ表 3，図 3 に示す．

5. 実験

3 節の手法で生成された階層クラスタリングが概念階層オントロジーを作成する上で有用な材料となることを検証する．

5.1 メンション抽出実験

5.1.1 実験のセットアップ

3.2 節で述べた LSTM-CRF モデルをメンション抽出に用いる．BiLSTM-CRF のハイパーパラメータを付録の A.1 節の表 A·1 に示す．モデル内で利用する単語分散表

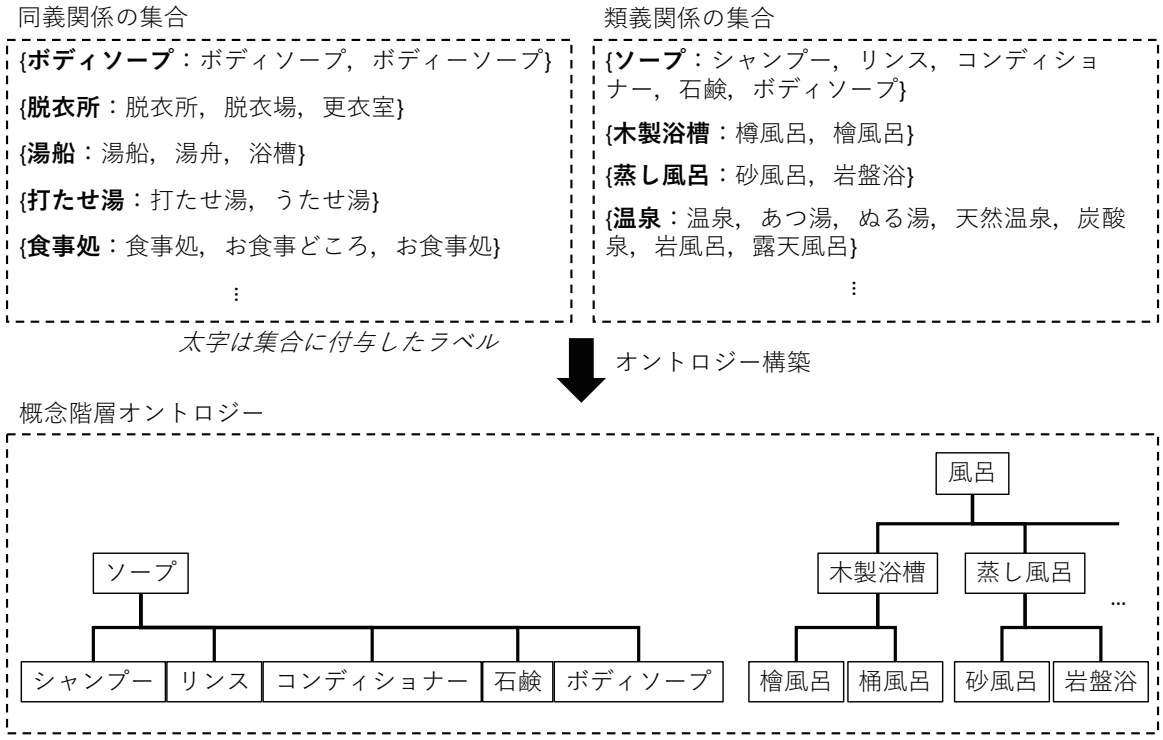

図 3 正解データのサンプル．

	学習	開発	評価
メンション数	9,802	3,338	3,344
文数	11,513	3,840	3,836
単語数	185,506	61,595	61,936

表 4 メンション抽出実験に用いた学習・開発・評価セットの分割とその統計．

	事例数	P	R	F
すべて	3,344	88.31	85.65	86.96
既知メンション	2,825	93.20	91.26	92.22
未知メンション	519	59.96	55.11	57.43

表 5 メンション抽出の結果．

現として，*gensim*[*5]で提供されている word2vec (continuous bag-of-words) を利用して作成した．分散表現の学習データは，ウィキメディア財団が公開している日本語版 Wikipedia のダンプデータ[*6]の 2018-12-26 版を利用した．この Wikipedia テキストに対して，マークアップ記号の除去と単語分かち書きを前処理として行った後，分散表現の学習に用いた．単語分かち書きツールとして mecab 0.996[*7]を利用し，その辞書として Juman 辞書を利用した．[*8]分散表現の次元数は 512 に，window サイズは 15 に設定した．学習データに出現する頻度 10 以上の単語のみをボキャブラリとして登録した．

5.1.2 実験結果

評価は適合率 (Precision; P)，再現率 (Recall; R)，F 値を評価尺度として用いる．アノテーションされた正解メンションとモデルが予測したメンションを比較し，完全一致した場合のみを正解と判断する．本実験では，すべてのメンションに対する評価，学習データに出現するメンション (既知メンション) に対する評価，学習データに出現しないメンション (未知メンション) に対する評価をそれぞれ行った．結果を表 5 に示す．すべてのメンション (既知メンションと未知メンションの合計) に対して F 値 89.96 を記録した．この結果から多くの正解メンションを抽出できていることがわかる．その内訳として，学習データに出現している既知メンションに対しては，予想通りの高い F 値 (92,22) を記録している．一方，より困難であると予想されていた，学習データに出現しない未知のメンションに対しても，改善の余地はあるものの中程度の F 値 (57.43) で抽出できていることがわかった．今後，未知メンション抽出に対してさらなる改善を行う予定である．

5.2 階層クラスタ構築実験

4 節で述べたように対象のマイクロドメインは温泉施設とする．階層クラスタリングに用いる単語分散表現も温泉

[*5] https://radimrehurek.com/gensim/
[*6] https://dumps.wikimedia.org/jawiki/
[*7] http://taku910.github.io/mecab/
[*8] 処理対象のテキスト量が多く，実務上の理由から解析速度の早い mecab を利用した．

ⓒ 2019 Information Processing Society of Japan

	単語分散表現	説明	半自動	全自動	差
1	Wikipedia	Wikipedia のみを用いて学習	51.64	51.54	0.10
2	Wikipedia → 温泉	1 の学習済み分散表現を，温泉レビューを用いて再学習	68.79	64.97	3.82
3	宿・観光	宿・観光レビューのみを用いて作成.	83.61	81.82	1.79
4	宿・観光 → 温泉	3 の学習済み単語表現を，温泉レビューを用いて再学習	85.69	82.77	2.92
5	Wikipedia + 宿・観光	Wikipedia と宿・観光レビューのテキストを連結して学習	90.09	84.89	5.20
6	Wikipedia + 宿・観光 → 温泉	5 の学習済みの分散表現を，温泉レビューを用いて再学習	92.61	84.52	8.09

表 6 階層クラスタ構築に用いた分散表現の学習法と評価結果 (F 値).

施設ドメインに特化させた方が良い結果になることが期待できる．実験では，対象ドメインに特化した単語分散表現を用いる効果を示すため，Wikipedia の単語分散表現を用いた階層クラスタとの比較を示す.

5.2.1 実験のセットアップ

階層クラスタ評価指標

人手で構築した正解クラスタと自動構築した階層クラスタを比較して評価する．階層クラスタの評価には，Zhao [3] らによって提案された，階層クラスタリング評価のための F 値を用いる．これは，正解クラスタに対し，生成された階層クラスタに含まれる全てのクラスタの中から，スコアが最大となる最適なアラインメントを探索し，そのスコアを求めるものである．正解クラスタと完全に一致するクラスタが階層クラスタに含まれる場合には，このスコアは 1 となる．このスコアが 1 に近いほど，正解クラスタに近く，階層クラスタを参照して正解クラスタを再現することが容易であると言える.

L_r を正解クラスタとし，n_r をその要素数とする．S_i を階層クラスタに含まれるクラスタとし，n_i をその要素数とする．S_i の要素のうち，L_r に含まれる要素の数を $n_r i$ とする．L_r と S_i の F 値は以下で定義される．ここで，$R(L_r, S_i)$ は $n_r i/n_r$，$P(L_r, S_i)$ は $n_r i/n_i$ で定義される.

$$F(L_r, S_i) = \frac{2 * R(L_r, S_i) * P(L_r, S_i)}{R(L_r, S_i) + P(L_r, S_i)} \quad (1)$$

クラスタ L_r の F 値は，階層クラスタ T の全要素に対して計算された F 値の最大値であり，以下の式で定義される.

$$F(L_r) = \max_{S_i \in T}(F(L_r, S_i)) \quad (2)$$

クラスタ全体の F 値は，式 3 のとおり，各クラスタの要素数の重み付き平均で計算される．c は正解クラスタに含まれるクラスタ数である.

$$FScore = \Sigma_{r=1}^{c} \frac{n_r}{n} F(L_r) \quad (3)$$

5.2.2 比較手法

メンション抽出のバリエーション

階層クラスタ自動生成手法の評価のため，以下に示す二つのプロセスで構築されたクラスタを評価する.

- 半自動：人手でアノテーションされた正解メンションをもとに構築された階層クラスタ
- 全自動：メンション抽出モデルを用いて自動抽出したメンションをもとに全自動で構築された階層クラスタ

両方のプロセスともにクラスタ生成は自動であるが，メンションの抽出方法が人手/自動である点が異なる．提案手法全体での評価に加えて，階層クラスタリング手法単独およびその際の単語分散表現の効果を評価するため，半自動のプロセスでは，入力とするメンションを正解クラスタと合わせている.

分散表現のバリエーション

階層クラスタリングは，各メンションに割り当てられた分散表現 (ベクトル) を用いて行われる．つまり，割り当てられた分散表現間の距離が近いメンション同士がクラスタを形成する．したがって，生成されるクラスタは，各メンションに割り当てられた分散表現に大きく依存する.

そこで本稿では，表 6 中の 6 種類の分散表現学習法をもとに階層クラスタを生成し，クラスタの質を比較・評価する．表 6 中では，以下に示す 3 種類のテキスト集合を用いて分散表現を学習している.

- Wikipedia: 日本語版 Wikipedia のダンプデータ
- 宿・観光: 「じゃらん net」に寄せられた宿および観光スポットのレビューの一部 [*9]
- 温泉:「じゃらん net」に寄せられた温泉施設のレビューの一部[*10].「宿・観光」レビュー文章集合に含まれる.

注意点として，「温泉」レビュー文章集合は「宿・観光」レビュー集合に含まれる (部分集合である).

分散表現を作成する際には，Juman 辞書に加えて，抽出されたメンションを単語登録した上で mecab0.996 で単語分かち書きを行なった．単語分散表現は gensim で提供される word2vec を利用して作成した．分散表現の次元数は 200 に，window サイズは 15 に設定した．データ中に出現する頻度 4 以上の単語のみをボキャブラリとして登録した.

5.2.3 実験結果

比較実験結果を表 6 に示す．正解メンションを用いて構築したクラスタ (半自動) において，「5 Wikipedia+宿・観光」と「6 Wikipedia+宿・観光→温泉」で学習した分散表現を用いたクラスタが高い F 値を記録している．また，

[*9] 実験に用いたレビューに含まれる Token 数はおよそ 570 百万個.
[*10] 実験に用いたレビューに含まれる Token 数はおよそ 750 千個. 温泉施設は観光スポットの一部であり，同レビューは前述の宿および観光スポットのレビューに包含されている.

宿・観光レビューを用いた場合 (3 6) がそれ以外 (1・2) と比べ大幅な性能向上が見られる．加えて，1 と 2，3 と 4，5 と 6 を比較すると，温泉レビューで再学習する効果が見られる．

自動抽出したメンションをもとに階層クラスタを構築したクラスタ (全自動) において，「半自動」に比べて若干スコアが低下したが，「宿・観光」レビューを用いたケース (3 6) では，F 値 80%以上を維持している．興味深いことに，「3 宿・観光」と「4 宿・観光→温泉」を比較すると温泉レビューでの再学習の効果が見られる一方，「5 Wikipedia+宿・観光」と「6 Wikipedia+宿・観光→温泉」を比較すると再学習の効果は見られなかった．正解メンションを用いた場合 (半自動) における「5 Wikipedia+宿・観光」と「6 Wikipedia+宿・観光→温泉」の比較では，温泉レビューでの再学習の効果が見られた点とも対照的な結果となった．この原因は，メンションの自動抽出精度とも関わる部分であると考えられるため，今後詳細に分析していく予定である．参考として，全自動で構築された「6 Wikipedia+宿・観光→温泉」の階層クラスタを付録の図 A·1 に示す．

6. 分析

節 5.2.3 の実験 1 において，スコアの違いが顕著であった「1 Wikipedia(半自動)」(図 A·2) と「6 Wikipedia+宿・観光→温泉 (半自動)」(図 A·3) の主な差異を次に示す．

「1 Wikipedia」の単語分散表現を用いて階層クラスタを構築した場合，表記揺れのメンションが適切にクラスタリングされないケースが多く見られた．例えば，{ ボディソープ，ボディーソープ }，{ 蕎麦，そば }，{ 食事処，お食事処，食堂 } は正解クラスタではそれぞれ同義クラスタとして定義したが，「1 Wikipedia」ではこれらのクラスタは再現されない．「6 Wikipedia+宿・観光→温泉」では，これらの表記揺れに関するメンションは適切にクラスタリングされている．このような表記揺れはレビューにおいては同様の文脈で多く出現するが，Wikipedia には多く存在しないためと考えられる．

また「1 Wikipedia」では，ドメインにより解釈が異なりうるメンションが，適切にクラスタリングされないケースも見られた．例えば，エアコンとコンディショナーが最近傍のペアとしてクラスタリングされた．エアコンと，エアコンの正式名称であるエアー・コンディショナーが同文脈で出現することが多いためと考えられる．しかし，温泉ドメインにおいては，コンディショナーはリンスの同義語として用いられるため，このクラスタリングは誤っている．

ドメイン固有の概念に関するクラスタリングも「1 Wikipedia」では再現されていないケースが見られた．例えば，正解クラスタでは家族風呂と貸切風呂を類義としており，「6 Wikipedia+宿・観光→温泉」の階層クラスタでは，この 2 つは最近傍のペアとしてクラスタリングされたが，「1 Wikipedia」では近くにクラスタリングされない．Wikipedia には，この 2 つの概念に関するコーパスが十分に含まれないためと考えられる．

上記は，最も顕著な違いが見られた「1 Wikipedia(半自動)」と「6 Wikipedia+宿・観光→温泉 (半自動)」を比較した考察であるが，1 と 2，3 と 4，5 と 6 の比較においても，上記と同様の傾向が見られた．ドメイン特化のコーパスを利用することで，マイクロドメインにおける同義・類義・上位下位のクラスタが多く含まれる階層クラスタを構築することが出来ると考えられる．

7. 関連研究

メンション抽出

Lample [2] らが固有表現抽出タスク向けに，BiLSTM（Bidirectional LSTM）と CRF を組み合わせた BiLSTM-CRF を提案し，高い性能を報告している．Web テキストから固有の名詞句を獲得する研究としては，池田 [5] らは Web ブログテキストから土産物の品名と店名を抽出するために BiLSTM-CRF を利用している．

オントロジー構築

鈴木 [6] は辞書の定義文を基にした上位語情報の抽出手法を提案した．隅田ら [7] は，Wikipedia の記事構造に含まれる節や箇条書きの見出しを用いて，上位下位関係を自動獲得する手法を提案した．松田ら [4] は，Wikidata を用いた遠距離教師あり学習により，Wikipedia アブストラクト内から大規模な関係知識を獲得する手法を提案している．これらの手法では辞書や Wikipedia，Wikidata を用い，それらに含まれる構造情報や概念間の関係を示唆している文章を活用することで，単語間の関係知識を大規模に獲得している．本手法では，Wikipedia や Wikidata には十分に含まれないマイクロドメインの知識を獲得するために，レビュー文章を用いた．レビュー文章には特定の文構造がなく，また概念間の関係を示唆する文も含まれないため，これらに依存しない手法を提案した．

8. おわりに

本研究では，レビュー文章集合からマイクロドメインのオントロジーを構築する手法を提案した．メンション抽出によりオントロジーに登録する概念を獲得し，対象マイクロドメインに特化した単語分散表現を利用することで，オントロジー構築に有用な階層クラスタを構築出来ることを示した．また，得られたオントロジーは，一般的な定義とは異なる，マイクロドメインに特化した概念と関係を獲得できていることを示せた．

今後は，複数のマイクロドメインで実験することで手法の有用性を検証し，実際の検索システムへ応用する予定である．検索システムへの応用にあたっては BERT による類似文検索との併用なども考えている．

ⓒ 2019 Information Processing Society of Japan

参考文献

[1] Kingma, D. and Ba, J.: Adam: A Method for Stochastic Optimization, *arXiv preprint arXiv: 1412.6980* (2014).
[2] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C.: Neural Architectures for Named Entity Recognition, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 260–270 (2016).
[3] Zhao, Y. and Karypis, G.: Evaluation of Hierarchical Clustering Algorithms for Document Datasets, *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, New York, NY, USA, ACM, pp. 515–524 (online), DOI: 10.1145/584792.584877 (2002).
[4] 松田耕史, 鈴木正敏 and 乾健太郎: Wikidata からの遠距離教師あり学習に基づく大規模関係知識獲得, **言語処理学会 第２５回年次大会 発表論文集**, pp. 660–662 (2019).
[5] 池田流弥 and 安藤一秋: 深層学習によるブログ記事からの土産の品名・店名抽出, **言語処理学会 第２５回年次大会 発表論文集**, pp. 526–529 (2019).
[6] 鈴木敏: 辞書からの上位語情報抽出とオントロジー自動生成, **自然言語処理**, Vol. 16, No. 1, pp. 101–116 (2009).
[7] 隅田飛鳥, 吉永直樹 and 鳥澤健太郎: Wikipedia の記事構造からの上位下位関係抽出, **自然言語処理**, Vol. 16, No. 3, pp. 3–24 (2009).

付　　録

A.1 LSTM-CRF のハイパーパラメータ

パラメータ名	値
単語分散表現の次元	500
LSTM の層の数	4
LSTM の隠れ層の次元	300
ミニバッチサイズ	32
最適化	Adam
L2 正則化. λ	0.0001
ドロップアウト率	0.1

表 A·1　実験に用いた LSTM-CRF のハイパーパラメータ.

5 節で用いたハイパーパラメータを表**??**に示す．なお，Adam のハイパーパラメータ β_1 と β_2 は文献 [1] で推奨されている 0.9 と 0.999 にそれぞれ設定している．

A.2 階層クラスタ

ⓒ 2019 Information Processing Society of Japan

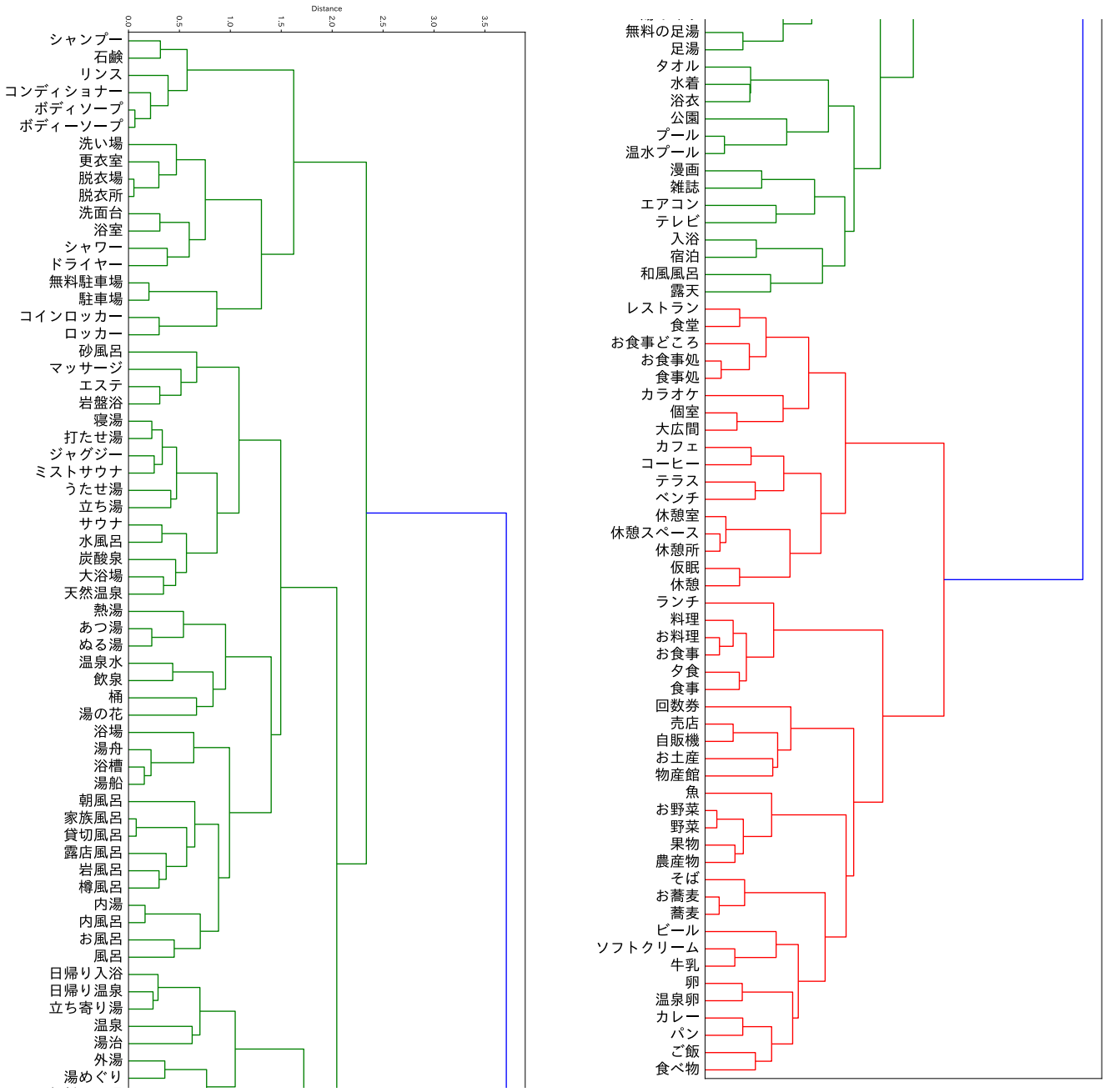

図 **A·1** Wikipedia+レビュー→温泉 (全自動) による階層クラスタ.

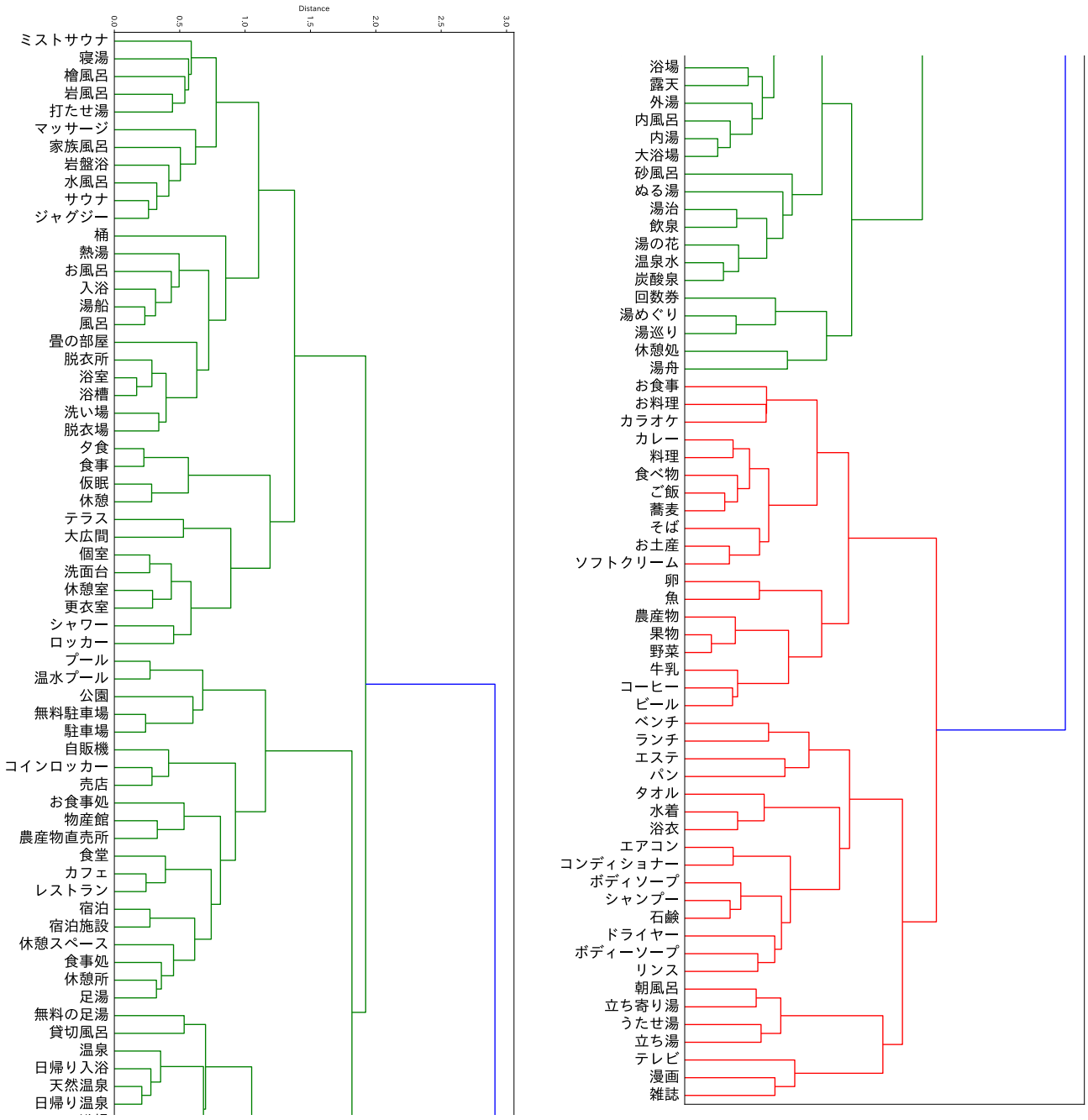

図 **A·2** Wikipedia(半自動) による階層クラスタ.

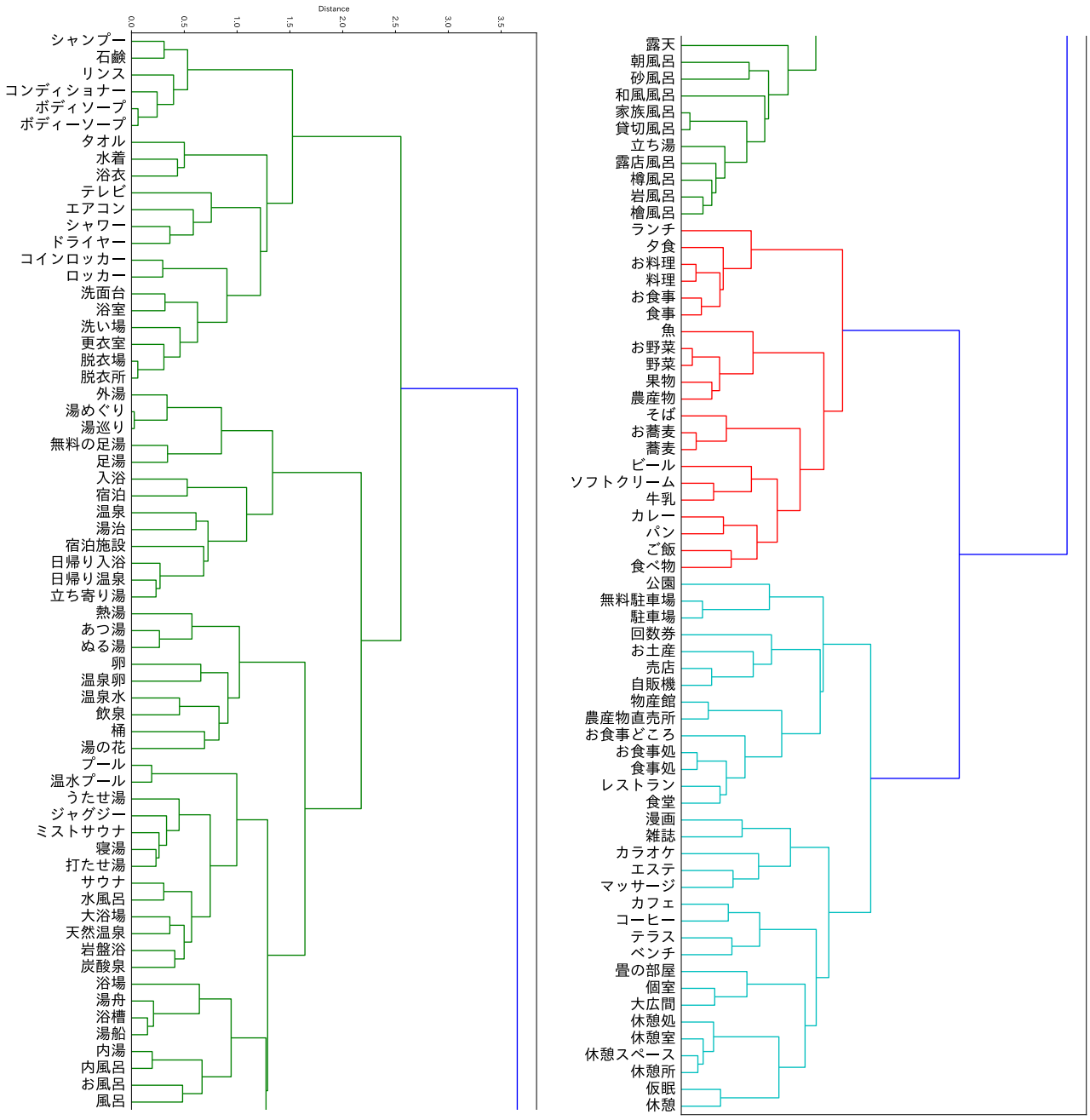

図 **A·3** Wikipedia+レビュー→温泉 (半自動) による階層クラスタ.