

Q&A サイトに投稿された質問の観点に基づくセグメンテーション

林 賢吾 飯田 龍 徳永 健伸
 東京工業大学 大学院情報理工学研究科
 {khayashi,ryu-i,take}@cl.cs.titech.ac.jp

1 はじめに

Yahoo!知恵袋¹や人力検索はてな²など、Web 上では Q&A サイトと呼ばれるユーザのトラブル解決のためのサービスが提供されている。ユーザはサイトに質問を書き込むことで、他の多数のユーザから回答を得ることができる。これらのサイト内の質問とその回答を効率的に検索することにより、類似した問題をかかえているユーザは問題を解決することができる。ただし、Q&A サイトの質問はユーザの問題が起こった状況や補足的な情報などが含まれているため、単純なクエリ検索ではユーザが意図しない検索結果を得ることになり、検索効率が悪い。例えば、例(1)は Ubuntu9.10 について質問をしているが、「9.40」という文字列も含むため、クエリに「9.40」と入力したユーザもこの記事を見ることになる。

(1) Ubuntu9.10 の起動時の音を消すにはどうしたらよいでしょうか？ 9.40 の時は、サウンドの設定で消すことができたのですが、9.10 ではサウンドの設定のところに適当なものが見当たりません。

この質問には「Ubuntu9.10 の起動時の音を消す方法が知りたい」という要求や「9.40 では起動音を消すことができた」という補足情報、「9.10 ではサウンドの設定で適当なものがない」という問題など、質問の意図を回答者に伝えるために質問がいくつかの観点で記述されているが、この観点を明示的に分け、観点によるファセット検索をおこなうことで検索が効率的になると考えられる。例えば、例(1)の場合は、質問者の要求部分である「Ubuntu9.10 の起動時の音を消す方法が知りたい」という記述だけに対してクエリ検索を行うことで、余剰な検索結果を抑制できる可能性がある。

このように、質問の観点を考慮した検索は Q&A サイト内の検索のために有益だと考えられるが、これを実現するためには少なくとも以下の2つの問題を解決する必要がある。

1. 観点の自動分類: Q&A サイトの質問を解析し、「要求」や「具体的な問題」など、質問の観点を記述した部分を自動的に同定する。
2. 観点間の情報の補完: 質問中の各観点を記述した部分を検索対象とする場合、代名詞や省略によって各観点部分に必要なキーワードが含まれていないと再現率の低下を招く。例えば、例(1)では、「9.10 ではサウンドの設定のところ...」を具体的な問題とし

て他の観点と分離してしまうと、Ubuntu に関して起こっている問題であるという情報が欠落する。これを省略補完や共参照解析、関係抽出などの技術を利用して復元する。

問題2については、既存の照応・省略解析などを利用することで解析が可能であると考え、本研究では、問題1に焦点を当てる。まず2節で Q&A サイトに出現する質問の観点を定義し、Web から収集した質問集合にその観点をアノテーションしたコーパスについて述べる。3節では、観点の特徴付ける表現がその観点の末尾に出現するという傾向を利用した観点分類の手法を提案し、4節で提案手法の有効性を評価した結果を示す。5節で関連研究との比較を行い、6節で結論と今後の課題について述べる。

2 観点の定義とアノテーション

Q&A サイトではユーザが遭遇したさまざまなトラブルや疑問について質問が書き込まれているが、本研究では分析を容易にし、問題の発散を避けるためにドメインを限定して質問を収集した。具体的には、ドメインをコンピュータに限定し、Ubuntu 日本語フォーラムで投稿数の多い初心者サポートフォーラムからランダムに2,000件の質問をその回答とともに抽出した。このうち1,000件を分析用(学習用)として利用し、残りの1,000件を評価用データとした。分析用の1,000件のうち200件を手で分析し質問中の記述をどう分割すべきか、またその分割した単位にどのような観点を付与すべきかを調査した。

表1: 質問の観点の定義

種類	概要	例
要求	直接の回答の要求	どうしたら良いでしょうか？
前提	問題の前提条件	Ubuntu10.4 を使用しています
操作	質問者の行った操作	動作確認を試みたところ
結果	問題の内容や操作の結果	画面が消えてしまいました
動機	質問者の意図や目的	Linux を触ってみようと思い
補足	条件が異なる場合の振舞い	Windows では問題ありません

人手で質問の観点の分類を行った結果、質問は典型的には問題が起こった背景と具体的な要求に分けることができ、背景情報には解決したい問題の前提条件や質問者が行った具体的な操作やその結果起こった問題と直接関係する結果などが記述されていることがわかった。この分析に基づき、表1に示す6種類の観点を定義した。また、観点は文中に複数出現するが、基本的には行為や状態に関連する述語と対応づけられるため、節を単

¹<http://chiebukuro.yahoo.co.jp/>

²<http://q.hatena.ne.jp/>

位に観点を分類すればよいことがわかった。この際、文はヒューリスティックな規則に基づき認定し、またコンピュータドメイン特有の記述だと思われるエラー出力やプログラムなどの記述については例外的に1つの節として人手で判断して作業を行った。

この分析結果を基に、収集した2,000件の質問のうち、日本語以外の投稿や情報提供のみを目的とする投稿など16件を除いた、1,984件に観点的アノテーションを行った。節の判定は作業者が人手で行い、また表1に示した観点は互いに排他的であると仮定し、1つの節に1つの観点を付与することとした。ただし、6種類のいずれにも該当しない場合(形式的な挨拶など)には何も付与しないこととした。

表2: アノテーションされた観点の頻度

観点	総数	平均	分散
要求	2,340	1.179	0.397
前提	2,552	1.286	0.915
操作	2,747	1.385	1.904
結果	4,200	2.117	2.616
動機	822	0.414	0.402
補足	558	0.281	0.302
合計	13,219	6.663	13.364

表3: 質問あたりの観点数(異なり)

観点数	1	2	3	4	5	6
件数	39	135	371	792	551	96
割合 [%]	2	7	19	40	28	5

アノテーションされた観点の頻度を表2にまとめる。この結果から、Q&Aサイトの質問では具体的な要求と同程度もしくはそれ以上の頻度で前提、操作、結果に相当する記述が含まれていることがわかる。表3は質問あたりいくつの異なる観点が含まれるかを調査した結果である。表3より、98%以上の質問は観点を2種類以上含むため、観点の分割によって検索の効率化が期待できる。

次に、定義した観点の妥当性・客観性を調査するために、2人の作業者が50件の同一の質問に観点をアノテーションし、一致率の調査を行った。2人の作業者のうち1人は第一著者であり、もう1人はアノテーションの仕様書⁴と実際の作業例を参考にして予備的な作業を行うことで訓練された作業員である。観点は節単位でアノテーションするため、節の末尾文節の観点の一致についてCohenのカッパ統計量[2]を計算した。この結果、 $\kappa = 0.67$ であり、作業員2人の作業結果はかなり一致していると判断できる。

3 提案する観点分類の手法

2節で作成した観点付きコーパスのうち、分析用の1,000事例を人手で分析したところ、観点を特徴づける手がかりは節の末尾に出現することがわかった⁵。例えば、要求は「(どうしたら)よいでしょうか?」のように疑問形で出現しやすく、また操作は「(ブートさ)せ

³観点の頻度は付与された範囲を1つとして数えた。

⁴観点アノテーションの仕様書や細かい分析については紙面の都合上省略する。詳細については文献[1]を参照されたい。

⁵詳細については文献[1]を参照されたい。

表4: 分類に用いる素性一覧

種類	番号	概要
語彙	1	節末の文節に出現する単語 ($\{\text{uni,bi}\}$ -gram)
	2	節末以外に出現する単語 ($\{\text{uni,bi}\}$ -gram)
	3	節に係る文節内の単語
	4	節から係る文節内の単語
	5	動詞(形容詞)の直後の助詞、助動詞
	6	観点の末尾に現れる典型的な表現の有無
位置	7	質問中の位置
	8	質問の最初(最後)の節か否か
	9	文の末尾か否か
前後関係	10	直前(直後)の空行の有無
	11	節の末尾の品詞
	12	節の直後の品詞
	13	係り先の品詞
前後の素性	14	前後の節の素性 1~13
	15	素性 14 のさらに前後の節の素性 1~13
語彙	16	観点セグメント内の節の素性 1~6
位置	17	観点セグメント内の節の素性 7~9
前後関係	18	観点セグメント内の節の素性 10~13
前後の素性	19	前後の観点セグメントの素性 16~18

ると」のように「~すると」という表現で結果へとつながる。また、結果は「(~している様子は)ありません」「~できません」といった否定的な表現で記述されやすい。つまり、観点分類の問題は、観点の範囲(以下、観点セグメント)を適切に同定し、その観点セグメントに観点の種類(以下、観点タイプ)を付与することで、単純に節単位で観点タイプを付与するよりも精度を改善できる見込みがある。そこで、本研究では観点分類の問題を観点セグメントの範囲同定と観点タイプ分類に分割し、この2つの問題を段階的に解くことで最終的な観点分類を行う。

まず、観点セグメントの範囲同定については、人手作成した規則と機械学習を用いた手法を併用する。規則に基づき境界を認定する場合は、「~したら」など典型的にセグメントの末尾となりやすいパターンを分析用データをもとに列挙し、そのパターンに合致する場合にはセグメントの末尾として決定する。適用した規則を以下にまとめる。

観点セグメントの末尾認定規則: 末尾の表現が「たら/ところ/ので/ように/時に/ときに/ため/思っ/おもっ/出て(きて)/ても/と(接続助詞)」である場合にセグメントの末尾とする。

規則を列挙して適用するだけでは現象を網羅できないため、規則に該当しない節を対象にセグメント末か否かの2値分類問題を解く。学習・分類の際は対象となる節から表4に示す素性1~素性15を抽出して利用する。

次に、同定した観点セグメントに対して観点タイプを分類する。2節で定義した「要求」、「前提」、「操作」、「結果」、「動機」、「補足」の6種類の観点と、「それ以外」の7値を分類する問題を one vs. rest 法で解く。すなわち、観点セグメントに対してある観点タイプか否かを分類する分類器をそれぞれの観点について作成し、入力された観点セグメントに対して分類器のスコア(分離平面からの距離)が最大の観点タイプを割り振る。

観点タイプ分類の学習・分類には表4に示す素性16~素性19を利用する。表4の素性ではセグメントの末尾の特徴を捉えるために語彙的な素性は末尾の節に出現しているかそれ以外かで区別されており、これにより典

型的にある観点タイプに出現する語彙的な特徴を捉えることができる。また、質問中では、要求は質問の最後に出現しやすい、前提は質問の前半に書かれやすいといった傾向があるため、その観点セグメントが質問中のどの位置に出現するかという情報も利用する。さらに、「～すると、～できません」のように操作の後に結果が記述されやすいといった観点タイプ間の隣接関係も分類の際の手がかりとなる。前後文脈の利用法にはさまざまなものがあるが、ここでは前後の観点セグメントからも同様の素性を抽出し、それを対象のセグメントから抽出したものと区別して利用する。これにより、同じ観点タイプの連続や、典型的に遷移する観点タイプの特徴を利用できると考えられる。

提案手法では上述のように2段階で決定的に解析するため、最初の観点セグメントの範囲同定で、本来異なる2つのセグメントを誤って一つのセグメントとして認定してしまうと、観点タイプ分類でそのセグメントに一つの観点タイプを付与するため確実に間違ってしまう。逆に、正解よりも狭い範囲をセグメントとして認定しても、正解のセグメントとの境界さえ一致していれば、誤って分割したセグメントそれぞれに同じ観点タイプを付与すれば範囲同定の誤りは回復できる。したがって、セグメント範囲の同定では精度よりも再現率を優先する解き方が望ましい。そこで、学習時には学習データ内で交差検定を行い、最終的な出力の正解率が最大となるような観点セグメント分類器のパラメタ⁶を推定し、実際の解析にもそのパラメタを利用して解析を行う。

4 評価実験

3節で提案した手法の有効性を調査するために評価実験を行った。観点セグメントをあらかじめ分割して観点タイプを分類することの有効性を調査するために、ベースラインモデルでは節単位で観点タイプを分類するモデルを採用する。この観点分類の問題は節を単位とした系列ラベリング問題ともみなせるため、ベースラインモデルの一つとして条件付確率場 (CRF) [3]⁷を用いて解析する。もう一つのベースラインモデルとして節を単位とした7値分類問題を one vs. rest 法で解く。学習・分類には SVM[4]⁸を利用し、提案手法でも同様に SVM を利用する。カーネルはすべての問題で線形カーネルを使い、パラメタはデフォルトのまま利用した。素性の抽出には形態素解析器 MeCab⁹と係り受け解析器 CaboCha¹⁰を利用した。

観点分類の評価には3種類の評価指標を用いる。一つ目は節単位でどの程度観点タイプが正解と一致しているかという正解率であり、節それぞれに対してどの程度正しく観点タイプを分類できたかを見積る。二つ目は観点セグメント単位の評価であり、正解に付けられた観点セ

表 5: 観点分類の正解率

手法	節単位	セグメント単位	部分正解
SVM	0.668	0.462	0.673
CRF	0.656	0.431	0.694
提案手法	0.682	0.523	0.721

グメントそれぞれに対して、観点の境界が正しく判定され、付与された観点タイプも正しい場合に正解、それ以外を不正解とした正解率を用いる。検索を目的とした評価では後者のセグメント単位でどの程度正しいかが重要となると考えられるが、この指標ではセグメントの範囲が少しでも違う場合には不正解となるため部分的に正解している場合を考慮できない。そこで、部分正解を評価に反映させるため、以下のように再現率、精度、F値に相当する値を見積る。

再現率: 正解データのある観点セグメント内の節に対して、節単位でどのくらい正解の観点タイプのラベルを付与できているかという割合を、評価データのすべての観点セグメントについて求め平均した値。

精度: システムが出力した観点セグメント内の節に対して、節単位でどのくらい正解データの観点タイプのラベルが一致しているかという割合を、システムが出力した観点セグメントすべてについて求め平均した値。

F 値: 再現率と精度の調和平均。 $(2 * \text{再現率} * \text{精度}) / (\text{再現率} + \text{精度})$ で求める。

このF値を以下、部分正解率と呼び、最初に述べた節とセグメントに関する正解率をそれぞれ節単位の正解率とセグメント単位の正解率と呼ぶ。提案手法では観点セグメントのパラメタを決める必要があるが、節単位の正解率が最大となるように学習データを5分割交差検定して観点セグメント分類器のパラメタを求めた。素性 19 (前後の観点セグメントの情報)の窓幅については、ベースラインの SVM と CRF では前後2つの観点セグメントを参照した場合が最も正解率が高く、一方提案手法では素性 19 を利用しない場合が最も正解率が高くなった。紙面の都合上、それぞれの手法で最も正解率が高い場合の結果について報告する。

実験結果を表5に示す。表5より、各評価指標でベースラインである SVM と CRF の結果が同程度であるのに対し、提案手法は2つのベースラインの正解率の高い方よりも節単位で1.4ポイント、セグメント単位で6.1ポイント、部分正解で2.7ポイント、正解率が向上しており、この結果から観点セグメントをあらかじめ同定し、セグメントの末尾の情報を利用して観点の分類を行う2段階の処理が有効であることがわかる。節単位、セグメント単位の正解率については有意水準 0.001% でマクネマー検定を用いて検定を行い、提案手法がベースラインと比べ有意に性能が良いことがわかった。

次に、提案手法の結果について、観点ごとの再現率、精度、F値を節単位、セグメント単位、部分正解のそれぞれについて求めた結果を表6に示す。表5では観点タイプすべてについての評価を行ったが、必ずしもすべての観点が質問の検索に必要なとは限らない。分析用データを人手分析した結果、要求や前提、結果など質問を構成する上で必須もしくは重要な観点を正しく分類できていることが重要であり、動機、補足など質問の補足的な記述に関しては分類を失敗しても比較的検索には影響を及

⁶具体的には SVM のモデルが出力したスコア (分離平面からの距離) に対し、どのスコアまでセグメントの末尾とするかという値をパラメタとし、これ動かして、正解率が最大となるパラメタを推定する。

⁷<http://mallet.cs.umass.edu/>

⁸<http://svmlight.joachims.org/>

⁹<http://mecab.sourceforge.net/>

¹⁰<http://code.google.com/p/cabocha/>

表 6: 各観点分類の実験結果

観点	節単位			セグメント単位			部分正解		
	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
要求	0.862	0.899	0.880	0.716	0.755	0.735	0.861	0.915	0.887
前提	0.640	0.583	0.610	0.363	0.382	0.373	0.638	0.633	0.635
操作	0.526	0.770	0.625	0.473	0.565	0.515	0.587	0.801	0.677
結果	0.669	0.834	0.742	0.523	0.574	0.547	0.722	0.852	0.782
動機	0.723	0.712	0.718	0.561	0.570	0.566	0.745	0.736	0.741
補足	0.445	0.200	0.276	0.289	0.144	0.192	0.475	0.209	0.290
観点なし	0.757	0.493	0.597	0.572	0.445	0.501	0.800	0.574	0.668
マイクロ平均	0.682	0.682	0.682	0.520	0.523	0.522	0.713	0.729	0.721
マクロ平均	0.660	0.642	0.635	0.500	0.491	0.490	0.690	0.674	0.669

ばさないと考えられる。このような考えから結果を考察すると、提案手法は要求、結果については比較的高い正解率を得ていることがわかる。前提については、前提となる記述に行為や否定的な表現が含まれるために操作や結果などと間違える場合が多く、相対的に悪い結果となっている。この問題についてはより質問の構造、特に前提が述べられた後に操作、結果と続くといった、質問の流れをより反映した特徴を捉える必要がある。

5 関連研究

Q&A の質問検索にはすでにいくつかの先行研究があり、質問の構造を反映した検索や不要箇所の除去などが提案されている。例えば、Lin ら [5] は中国語の Q&A サイトの質問中の挨拶やお礼などの検索に必要な無い表現を自動的に削除し、また 1 つの投稿の中で複数の事柄を質問している場合に質問を分割し、その分割結果を使った質問検索の実現について議論している。また、Wang ら [6] は Q&A サイトに投稿された質問に対し、複数の回答が必要な場合に限定し、質問中でそれぞれの回答を求める要求とその背景の対応付けを行っており、この処理を適用することで Yahoo! Answers の質問集合を用いた類似質問の検索において検索精度が向上することを報告している。

本研究では、Q&A サイト中の質問を観点単位にセグメンテーションする問題を扱っているが、類似する問題として科学論文を対象とした観点のセグメンテーションの研究がある [7, 8, 9]。例えば、Teufel ら [7] は論文中で他の論文を引用している文脈を文単位で目的、背景など 7 つの観点に分類し、その結果を用いて観点を考慮した論文要約を行っている。山崎ら [9] は医療文献の概要について背景や手法、結果などの観点を文単位で付与する問題に取り組んでいる。既存のセグメンテーションの問題設定では文を単位に観点を捉えているのに対し、Q&A サイトでは 1 文中に複数の観点が頻繁に出現するため、より細かい節単位で分類している点が異なる。このため、本研究で扱う問題では文の中の構造と文間の関係を適切に捉える必要がある。

6 おわりに

本研究では、Q&A サイト内の質問検索を効率化させるために、質問内の記述を前提や結果など 6 種類の観点に分類する問題を設計し、観点のセグメントの末尾にそのセグメントを特徴付ける表現が出現するという点に着目し、観点の自動分類を行う手法を提案した。Ubuntu

日本語フォーラムの質問に実際に観点をアノテーションし、そのデータを評価用データとして観点の自動分類の評価を行った。この結果、節を単位に分類を行うベースラインと比較して、提案手法が有意に分類性能が良いことを示した。

今後の課題として、まず他のドメインへの適用が考えられる。観点の人手アノテーションに関して、コンピュータのドメインではかなり高い一致率でアノテーションできたが、他のドメインで同じ結果になるかを調査する必要がある。また、提案手法で利用した観点セグメント末尾の特徴を捉えるという手法がどの程度一般的なのかについても調査を行う必要がある。

さらに、提案手法を利用した実際の検索の効率化についても調査する必要がある。文献 [1] で予備的に調査を行い、検索件数を減らしながらもある程度のユーザが意図した質問を検索できることがわかったが、これを大規模に評価する必要がある。また、1 節で述べたように本手法は質問検索に必要なだと考えられる部分問題のうち観点分類しか解いておらず、もう一方の観点間の情報補完の問題を解いた結果どのように検索効率に変化するのかという点についても調査を行う必要がある。

参考文献

- [1] 林賢吾. Q&A サイトに投稿された質問の観点に基づくセグメンテーション. Master's thesis, 東京工業大学, 2012.
- [2] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46, 1960.
- [3] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, 2001.
- [4] V. N. Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing Communications, and control. John Wiley & Sons, 1998.
- [5] Chuan-Jie Lin and Chun-Hung Cho. Question pre-processing in a QA system on internet discussion groups. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pp. 16–23, 2006.
- [6] Kai Wang, Zhao-Yan Ming, Xia Hu, and Tat-Seng Chua. Segmentation of multi-sentence questions: towards effective question retrieval in cQA services. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 387–394, 2010.
- [7] Simone Teufel and Marc Moens. Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, Vol. 28, pp. 409–445, 2002.
- [8] Yufan Guo, Anna Korhonen, and Thierry Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 273–283, 2011.
- [9] 山崎貴宏, 新保仁, 松本裕治. 系列パターンを素性とした論文概要文の自動分類. 情報処理学会研究報告 知能と複雑系, pp. 129–134, 2003.