意見情報獲得のためのクエリー関連のドメイン特徴語抽出

峠泰成　山本和英
長岡技術科学大学 電気系
E-mail:{touge, ykaz}@nlp.nagaokaut.ac.jp

1　はじめに

Web 文書の増大によって様々な情報を容易に取得することが可能となった. このことから, 大量の文書データから製品やサービスに対する意見・評判情報の獲得の研究が盛んになっている. 意見情報を収集するための重要な要素は, 評価対象の特定と評価結果である. 評価対象となる語は, 抽出対象となるドメインによっても異なるため, あらかじめ辞書として構築することは非常に手間となる. また, 抽出した意見情報の集約には, 評価対象項目に対して関連性を持たせる必要がある.

本研究では, 抽出対象となるドメインの違いを考慮し, 抽出しておくべきドメイン特徴語を自動獲得する手法を提案する. まず, 意見情報の抽出対象となる情報を明確にするため複合名詞を考慮する必要がある. そのため, 検索エンジンを用いた複合名詞同定処理を行う. 次に, 入力されたクエリーとの語の関連性に着目したドメインの特徴語抽出手法を行う.

抽出された語がクエリーとの関連性を持つドメインの特徴語抽出手法を提案する.

2　関連研究

意見情報を抽出する研究では, 評価対象となる項目の語を人手により抽出する手法が多く用いられる [1][2]. Kobayashi et al.[1] は評価表現との共起抽出パターンを用いて評価の対象表現と属性表現を半自動収集する手法を提案している. ドメインを限定して作成する場合には非常に有効であり, 精度の向上にもつながる. しかし, ドメインの違いにより人手ですべての辞書を構築するより自動構築する必要があると考える.

また, あるクエリーをもとにして関連用語 [3] や専門用語 [4] を抽出する研究や上位語, 下位語の関係抽出 [5] の研究も行われている. 佐々木ら [4] はウェブを用いて専門用語辞書の構築を行っている. 入力したクエリーに対する関連性に着目し語を抽出している. 関連語抽出の研究では, 複合語を考慮することや専門用語のみでなく一般的な語も抽出対象とする必要がある. 特に意見情報抽出の場合はこの傾向が強く, ドメイン共通の語（色, サイズ など）やドメインに特化した語 (テレコンバージョンレンズ, ドライブシャフトブーツ など) の両方を考慮した抽出手法が必要となる.

本研究では, 複合名詞を考慮し, ドメインにおいて評価対象となる特徴語を自動抽出する手法を提案する.

3　提案手法

3.1　ドメインの特徴語の定義

意見情報を抽出するドメインによって評価対象となる表現は異なる. 車のドメインにおいて評価情報として抽出される対象は, ハンドル, アクセル, シートといった表現であり, デジタルカメラのドメインにおいては, メモリー, シャッター, フラッシュといった表現である. 本研究では, あるドメインにおいて評価対象となる語をドメインの特徴語と定義する.

[例 1] { エスティマ }特徴語 の { 乗り心地 }特徴語 は良い.

対象となるドメインに対して, テキストデータからあらかじめ特徴語の辞書を自動作成することができれば有効性が向上する.

ドメイン依存の用語を網羅的に収集する際に, 用語としての妥当性を考慮した抽出をすべきである. 特に, 複合名詞をドメイン特徴語として抽出する場合, 個人により抽出する範囲が異なる恐れもある. 例えば, 車のドメインにおいて,「高速安定性」と「安定性」では, 評価される項目は車の安定性であるが, さらに具体的な高速安定性も辞書として作成するべきかは判断が揺れる. また,「エンジン」と「音」を元に「エンジン音」の複合名詞を作成する場合には, 元々の語が複合名詞の部分集合であるため, 元々の語を辞書として収集しておけば良いが, 自動収集の利点を活かし複合名詞の収集も行う.

本研究は大きく 2 つに分けられ, 1 つ目は, 複合名詞を考慮するための名詞連接から妥当性のある複合名詞を同定する手法, 2 つ目は, ドメインにおけるクエリー関連の特徴語を抽出する手法となる.

3.2　名詞連接からの複合名詞同定

3.2.1　特徴語候補となる品詞と複合名詞

特徴語として表現することのできる語は評価の対象となる. 例として「フィット」,「IXY」などの対象表現や「アクセル」,「シャッター」などの属性表現となる. これらを構成している語の特徴は, 名詞, 未知語が候補となることが多い. さらに,「乗り心地」,「高速安定性」などの複合名詞[1]として構成されることも多い.

特徴語になりうる語は構成される長さも様々であり, 抽出すべき単位も複合名詞として意味をなす必要がある. これは, 意見情報を抽出し分析を行う場合, 最小の形態素単位で抽出するより情報を明確にすることができると考える.

[例 2] **エンジン音**の軽さには不満でした.

例 2 に示した語は 2 形態素からなる特徴語であるが, 評価されている対象の「エンジン」に対する意見ではなく,「エンジン音」自体が対象になっている. 従って, 意見情報解析を行う際には複合語の特定も重要な要素となる.

本節では, 抽出すべきドメインの特徴語の構成単位を特定する.

まず, 入力されたテキストデータに対し形態素解析を行い, 抽出候補を選定する. 本手法では, 茶筌[2]の品詞体系に準じ, 名詞-一般, 名詞-サ変接続, 名詞-固有名詞, 未知語, 記号列 (記号の連結) を抽出候補とした.

さらに,「助手席スライドドア」,「純正マフラー」などの複合名詞を取得するために, 候補となる品詞の結合を行う.

各形態素は名詞で表現されているが,「助手席スライドドア」の場合は 4 形態素を結合させた複合名詞で辞書に登録

[1] 本研究では特徴語候補の連接を複合名詞同定の対象とする.
[2] http://chasen.naist.jp/hiki/ChaSen/

しておく方が解析の際の情報が多い. 従って, 抽出候補となる品詞の連接をドメイン特徴語の候補として結合する. ただし, 結合を行う際, 複合語としての構成に誤りが多く生じるパターンは, あらかじめ例外規則を作成する.

3.2.2 検索エンジンによる複合名詞同定処理

結合して作成した複合名詞が必ずしも正しい結合でない場合も多く存在する.

[例 3] リモコンキー追加, カーステ取り付け, スポーツバンパー装着

ドメインの特徴語は評価される対象であるため, 語として意味をなさない場合や, 表現としての妥当性のないものについては削除すべきである. フィルタリングの対象となる候補は, 例 3 で示すような語と意味をなさない未知語 (ワーオなど) である.

複合語の作成の研究として, 中川ら [6] はコーパス中の語の連接頻度を用いた手法を提案している. 本研究では, 抽出対象となる候補表現の妥当性を判定するため, 同一ドメインのコーパス中の語では補完できない情報を考慮する,Web の検索エンジンを用いた手法で複合名詞の同定を行う.

デジタルカメラのドメインにおいて, 同一ドメインのテキストデータであれば, 「レンズ」や「シャッター」といった語は頻度情報が上位に出現する. しかし,「リモコンキー追加」のように複合名詞とした場合は, 頻度情報はほとんど 0 となる. 実際に「リモコンキー追加」という語は,「リモコンキー」と「追加」の 2 語から構成されていると考えるのが妥当である. そのため, 2 形態素目で同定するべきという情報を与える必要がある.

以下の STEP で複合名詞同定を行う.

STEP1 処理対象の語から同定候補を作成する

「リモコンキー追加」を「リモコン」「キー」「追加」の 3 つの形態素に分割し候補を作成する. 分割した候補から, 例 4 のように複合名詞を構成する組み合わせをすべて考える.

STEP2 検索エンジンによる同定候補ヒット件数の検索

作成した複合名詞候補の妥当性を判断するために, 一般的には大規模コーパスを用いるが, 本手法では Web 検索エンジンを使用する. 作成した候補に対して, 検索エンジンによる検索ヒット数を取得する.

[例 4](リモコンキー追加：15), (リモコンキー：44,700), (キー追加：575), (リモコン：2,150,000), (キー：8,480,000), (追加：19,000,000)
※ 数字は検索ヒット件数

STEP3 最長一致法による複合名詞同定

検索エンジンによる検索ヒット件数をもとに, 複合名詞の分割ポイントを検出する. 検索ヒット件数が閾値 m 以上を満たす候補語を形態素数が多い順に複合名詞として判定し, もとの複合名詞から同定していく. 形態素数が同じ場合は, 検索ヒット数の多い順に決定していく. すべての分割ポイントを検出したら処理は終了する.

「リモコンキー追加」は, 検索ヒット数の閾値 m を 1000 と設定した場合, 「リモコンキー」と「追加」の 2 語に同定されることとなる.

複合名詞同定を前処理として行い, ドメインの特徴語となる候補をすべて生成する.

3.3 ドメインの特徴語自動抽出手法

前節の複合名詞同定の前処理を行った後, 実際のドメインの特徴語抽出を行う.

3.3.1 候補絞り込み

入力されるドメインのコーパスは大規模なテキストデータとなり, ドメインの特徴語候補となりうる語の数も非常に多くなる. ドメイン特徴語は, メインクエリー（デジカメ, 車, 携帯電話 など）に対して関連の深い語となると考えることができるため, 候補語から関連性の高い語に絞り込みを行う. また, メインクエリーと特徴語との関連性をリンクすることで, 意見情報の検索向上に繋がると考える.

本研究では, メインクエリーの周辺には上位語や下位語が多く存在しているという仮説のもとに絞り込みを行う.

STEP1 入力した文書から候補語となるキーワードを抽出

ドメイン特徴語候補として扱う品詞は, 名詞-一般, 名詞-サ変接続, 名詞-固有名詞, 未知語, 記号列, 名詞-複合名詞 とする. また, 記号 (! , % など) の含まれる語, 数字列の語, ひらがな列の語は抽出対象外とする.

STEP2 1 文単位で絞り込みペアを作成

対象となる 1 文から隣接する抽出対象の候補からペアを作成する. 作成されるペアは 1 文単位で処理され, 抽出候補の前後の候補語それぞれを例 5 のようにペアとして抽出する.

[例 5] この**車**の**エンジン**にもう少し**トルク**があれば**運転**も楽しくなるのに.
→ { 車 , エンジン } , { エンジン , トルク } , { トルク , 運転 }

STEP3 メインクエリーを対象とした絞り込みペアの検索

ドメインを特定するため, 作成したペアからメインクエリーを検索する. 「A の B」のように, メインクエリーの周辺にはそのドメインに関連する語が現れやすいため, ペアの前後の出現位置を用いて特定する. メインクエリーはドメインを特定するキーワードであり, 作成したペアから前検索と後検索の両方を行い, メインクエリーとの隣接語を抽出する.

[例 6]
{ **車**, エンジン } → 「エンジン」を隣接語として抽出
{ 加速 , **車** } → 「加速」を隣接語として抽出

隣接語として抽出された語は, メインクエリーとの関連度が文書中の他の語に比べて高い.

STEP4 隣接語を対象とした絞り込みペアの検索

獲得した隣接語をもとに STEP3 と同様に前検索と後検索を行う. 隣接語から派生した候補語は, メインクエリーと隣接語の連想語となるキーワードとして抽出することができる.

[例 7]
{ **エンジン** , トルク } →「トルク」を連想語として抽出
{ アクセル , **エンジン**} →「アクセル」を連想語として抽出

連想語として抽出された語は, 一般的に出現する候補語に比べてメインクエリーとも関連があり, 隣接語との関連も強い.

STEP5 隣接語から抽出された連想語に対する絞り込み検索

STEP4 の手法により抽出された連想語において, ある隣接語の前検索, 後検索の両方に含まれていた場合, ドメインの特徴語として取得する.

[例 8] 隣接語：「エンジン」
前検索：{ エンジン ： **ブレーキ** , **エンジンオイル** , 水 , **ハイブリッド** , ミラー, … }
後検索：{ **ブレーキ** , **エンジンオイル** , 角度 , 水平方向, **ハイブリッド**, 上がり, … ：エンジン }

→ “ブレーキ” “エンジンオイル” “ハイブリッド” を「エンジン」からのドメイン特徴語として抽出

STEP4 と STEP5 は STEP3 により抽出された候補すべてに対して行う. 以上の処理で抽出された結果をドメインの特徴語とする.

3.3.2 関連度の算出

抽出したドメインの特徴語とメインクエリーとの関連度の算出方法について述べる. 本研究では, 特徴語を抽出する手がかりとした 2 つのキーワード（メインクエリー, 隣接語）との関連度を算出し, 抽出した特徴語に関連性のスコアを付与する. 関連度の算出手法として, 本研究では検索エンジンの検索ヒット数を用いる手法を提案する. 3 つの語を用いて関連度を算出するため, 同じドメイン以外で出現する語の判別や, ノイズの削除に効果があると考えるが, 本研究では閾値によるフィルタリングは行っていない.

関連度 R の算出手法は式 (1) を用いる.

$$R(M,D) = \frac{H(M,D)}{H(D)} * \frac{H(N,M)}{H(N)} * log(S+1) \quad (1)$$

※ H(*): 検索ヒット数 , H(A , B): クエリー A と B の AND 検索ヒット数 , S: ドメイン特徴語を抽出した隣接語の頻度, M : メインクエリー , D : ドメイン特徴語 , N : 隣接語

式 (1) は, メインクエリー, 隣接語, ドメイン特徴語の 3 つ組の検索ヒット数を用いて算出する.

算出した関連度によりスコア順にソートし, メインクエリーとの関連度をもつドメインの特徴語として出力する.

4 評価実験

本研究での提案手法に対し評価実験を行った. 実験に使用するデータは,Web 掲示板[3]から取得した「携帯電話」,「車」,「デジタルカメラ」の 3 つのドメインの書き込みである. 実験データは, 携帯電話：約 85 万文, 車：約 106 万文, デジタルカメラ：約 116 万文のデータを使用した. また, それぞれのドメインについてのメインクエリーは,「携帯電話」,「車」,「デジタルカメラ」と設定した.

4.1 複合名詞同定結果

ドメイン特徴語を構成する複合名詞同定結果を示す.

検索エンジンには “Google”[4] を使用し, 検索エンジンによる候補抽出の閾値 m を 1000 件と設定し実験を行った.

各ドメイン別に複合名詞を分割した際の分割後形態素数（2〜5 形態素）に対し, 各形態素数ごとに 100 件を無作為に取得した時の同定精度を表 1 に, 実際の同定結果を例 9 に示す.

表 1: ドメイン別複合名詞同定精度

構成形態素数	携帯電話	車	デジタルカメラ
2	0.77	0.82	0.85
3	0.73	0.82	0.78
4	0.71	0.74	0.73
5	0.76	0.83	0.77
平均	0.74	0.80	0.78

[例 9] ハッピーボーナス/対象, 写メールモード/起動, ローパスフィルタ/ゴミ/付着, ミノルタ/VS/オリンパス
※ / は同定された分割ポイント

結果より, どの形態素数の語を同定するにも精度がそれほど変化せず複合名詞の同定を行うことができると言える. しかし, この同定結果によりその後の処理への影響が大きくなるため, さらに精度を向上させる必要がある.

4.2 ドメインの特徴語抽出結果

3 つのドメイン別に自動抽出したドメイン特徴語の抽出精度について述べる. まず, 隣接語から絞り込みを行ったドメイン特徴語の抽出語彙についての結果を示す. それぞれのドメインにおける自動獲得したドメイン特徴語抽出語彙数を表 2 に示す.

抽出した語にはまだノイズとなる表現が多く見られるため, メインクエリーとの関連性の高い語句を意見情報検索の情報として用いたい. ドメイン別に抽出された特徴語に対して関連度を算出しているため, 上位にはメインクエリーとの関連性の高い語が収集されているはずである. 抽出された語の上位 1000 語に対して, 人手によりメインクエリーの関連語であるか否かを判断した. 比較として入力文書での頻度上位と比較する. この結果を表 3 に示す. 関連度の上位 1000 語の抽出精度は 80% 程度であった. また, 実際に抽出されたドメイン特徴語と隣接語の関連性の例を表 4 に示す.

表 2: 本手法によるドメインの特徴語抽出語彙数

ドメイン名	獲得語彙数
携帯電話	3503
車	7122
デジタルカメラ	5803

表 3: 本手法によるドメインの特徴語抽出精度

ドメイン名	精度 (提案手法)	精度 (頻度による手法)
携帯電話	0.71	0.41
車	0.80	0.42
デジタルカメラ	0.76	0.39

関連度は複合名詞も考慮したスコアとなっているため, 出力結果も頻度情報では下位に出現する語を収集することも可能であった. 本手法で抽出された語はメインクエリー 1 語から派生した隣接語による関連性を重視している. 表 4 より, 隣接語とドメイン特徴語は関連性の高い語を多く抽出することができている. 実際に意見情報を抽出する際, 関連性を持つ語に対しての検索を行うための情報を保持するという点で有効性のある結果であると言える.

[3] http://www.kakaku.com/bbs/
[4] http://www.google.co.jp/

表 4: 抽出された隣接語と抽出されたドメイン特徴語

隣接語 (メインクエリー)	ドメイン特徴語
液晶画面 (携帯電話)	傷, 性能, 文字, 画像, デジカメ, 画面, ＱＶＧＡ モニター, 保護シート, カメラ, ケータイ, 画素 サイズ, 保護フィルム, バッテリー, 消費電力
キーレス (車)	ＯＰ, 電池, オプション, イモビ, ボタン, エンスタ 開錠, 鍵, ターボタイマー, セキュリティ, エンジン 鍵穴, エンジン始動, ドアロック, 赤外線 電波, スペア, 機能, 連動, エンジンスターター
ワイコン (デジタルカメラ)	テレコ, ズーム全域, ポケット, 一眼, テレコン マクロレンズ, オリンパス, 望遠, アダプター ワイコン, 保護フィルター, レンズアダプター, 径, フード, ＧＸ, ＰＬフィルター, レンズフード

5 考察

5.1 複合名詞同定処理について

本研究では, 複合名詞同定処理を検索エンジンを用いた手法で行った. 分割ポイントの同定に検索ヒット数を考慮する簡単な手法でも比較的良好な結果であった. しかし, いくつかの名詞を結合した場合の問題として, 動詞的な働きをするサ変名詞がある. 例えば,「リモコンキー追加」は「追加」が動詞的な働きをするため,「リモコンキー」で同定するべきである. 例の場合には,「キー」は「追加」よりも「リモコン」と結合する頻度が多く, 本手法での同定処理でうまくいく.

「新車購入条件」の名詞連接の場合,「新車/購入条件」とした同定が望ましい. しかし,「新車購入」と「購入条件」が同程度の検索ヒット数のため, 検索ヒットが多かった「新車購入」が正解と出力された. 誤りの多くはこのような状況であり, 判定条件に検索ヒット件数そのものを用いるのではなく, 語の結合度を付与した手法や文法的知識を用いる必要がある.

また, 本手法による同定間違いとして, 固有名詞の判定が挙げられる. 検索エンジンにおいても頻度の少ない固有名詞 (ヘルシーパーク裾野 など) については閾値により同定されない. これらに対応するためには, 閾値を一定にする設定するのではなく, 検索した結果に応じた閾値を検討する必要がある.

検索エンジンの検索ヒット数は直接的な値を返さないため, 検索結果に応じた処理をうまく活用することも改善策として考えられる.

5.2 ドメイン特徴語抽出結果について

入力文書から抽出されたドメイン特徴語は, 複合名詞も考慮した結果を得ることができた.

評価実験より, 抽出した結果の関連度スコア上位 1000 語には, 一般的な文書頻度の手法に比べ, ドメインに特徴的な表現が多く収集された. 特に, 複合名詞は文書全体から見れば頻度が低くなり, メインクエリーとの関連性が高い語であっても下位に出現する場合が多い. 検索エンジンを用いたスコアリングを行うことによって, 擬似的な大規模コーパスとして活用することができ, 対象ドメインとそれ以外のドメイン情報との差を判別することができる

ドメインの特徴語を抽出するために絞り込みのペアを作成したが, 表 4 に示すように連想関係を把握するための手法として有効である. しかし, メインクエリーから隣接語を獲得した際, 隣接語に制限を加えていない. これによって, 隣接語にノイズ表現が入ってしまうことが避けられない. 隣接語を取得する段階で隣接語とメインクエリーとの関連度を算出し, ある閾値でノイズとなるデータを削除する手法で解決することができる. しかし, 閾値の設定は容易に決定することができないという問題を含んでおり, ノイズとなるデータを削除することにより, 必要である語を削除してしまう可能性もある. これらに対応した処理が必要である.

アプリケーションとして意見情報検索を行う場合, 検索効率を考慮すると関連度の上位にくるキーワードを用いて行うことが有効であると考えることができる. 精度の向上には, 必要である隣接語をさらに絞り込みを行ってからドメインの特徴語抽出を行うべきである.

さらに, 本手法では同一ドメインからの抽出について議論したが, 検索エンジンと語の組み合わせを用いることで, 様々な分野の文書からクエリーに関連する語を抽出することも可能であると考える.

また, 実際に関連性に基づいた意見情報の検索効率について実験を通して検討すべきである.

6 おわりに

本研究では, 複合名詞を考慮したクエリーに関連するドメインの特徴語を獲得する手法を提案した.

まず, 対象文書の名詞連接から的確な複合名詞を同定する処理として検索エンジンを用いた. また, ドメインにより評価対象となる語が異なるため, クエリーをもとに関連表現を収集する手法を提案した. クエリーに関連する語をドメインごとに抽出でき, 語と語の関連性を持ったドメインの特徴語を獲得することができる. 語と語の関連性には検索エンジンを擬似的な大規模コーパスとして扱い, 関連度を算出した.

問題点として, 高頻度語は関連表現を大量に取得してしまう傾向にあるため, 扱う関連語の強さによりスレッショルドを作成する必要がある. また, 語と語の関連性を用いた意見情報の検索効率が改善されるか否かを検討すべきである.

謝辞

本研究の一部は、平成 17-19 年度 総務省 戦略的情報通信研究開発推進制度 (SCOPE) の支援によって実施した.

参考文献

[1] Nozomi Kobayashi, Kentaro Inui and Yuji Matsumoto, Collectiong Evaluative Expressions for Opinion Extraction, Proc. of IJCNLP2004, pp.584-589, 2004.

[2] Bing Liu, Minging Hu and Junsheng Cheng, Opinion Observer : Analyzing and Comparing Opinions on the Web, Proc. of WWW2005, pp.342-351, 2005.

[3] 山本英子, 梅村恭司, 辞書を用いない関連語リストの構築方法, 情報処理学会研究報告,NL-148-12, pp.81-88, 2002.

[4] 佐々木靖弘, 佐藤理史, 宇津呂武仁, ウェブを利用した専門用語集の自動編集, 言語処理学会第 11 回年次大会発表論文集, pp.895-898, 2005.

[5] Kosuke Tokunaga, Jun'ichi Kazama and Kentaro Torisawa, Automatic Discovery of Attribute Words from Web Documents, Proc. of IJCNLP2005, pp.106-118, 2005.

[6] 中川裕志, 湯本紘彰, 森辰則, 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理,Vol.10,No.1,pp.27-45,2003.