自然言語処理技術を応用したテキスト (会話) の話題特定

森田 大輝† 中島 陽子‡ 本間 宏利‡ 山本 和英*
†釧路工業高等専門学校 情報工学科
‡釧路工業高等専門学校 創造工学科
*長岡技術科学大学大学 電気電子情報工学専攻
{honma, yoko}@kushiro-ct.ac.jp

1 はじめに

昨今，インターネット技術の普及により，SNS やブログなどを通じて商品のレビューを発信する消費者が増えている．また，日々の生活で感じたことや不満な点も，そうしたツールを通じて気軽に発信できるため，企業にとって，そのような潜在需要を調査することは商品開発において非常に重要であるといえる．しかし，インターネット上のそれらのテキストには，基本的にノイズが大量に含まれるため，漠然とした調査を行なっても有益な結果を得ることは困難である．

本研究では，自然言語処理の技術である TF-IDF の考え方を応用し，テキスト中で展開されている「話題」を特定するシステムの構築を目的とする．これにより，テキストからのノイズの除去を行なった上で，効率的に有益な情報の収集が可能となる．

なお，類似研究として文書自動分類があげられるが，それはテキスト中の話題が 1 つしかないと仮定した上で，既存カテゴリに分類するものである [1]．本研究はテキスト中の会話群から複数の話題を時系列的に特定するものであり，文書自動分類研究とは趣旨が異なる．

2 研究概要

本研究で行う話題特定の手法は，最初にテキストをいくつかの文章群に分割する．分割された 1 つの文章群を 1 ドキュメントという単位で扱い，各ドキュメント内の各単語から連想される話題をドキュメントに付与する．その後，話題を付与したドキュメント内をさらに詳細に分析し，1 文ごとの話題の特定を行う．

話題の付与には長岡技術科学大学の自然言語処理研究室で作成された話題分類単語辞書[1]を用いる [2]．

[1]http://www.jnlp.org/SNOW/D11

なお，本研究の話題解析の対象は会話文のみに限定し，コーパスは国立国語研究所にて公開されている名大会話コーパス[2]を用いる [3]．

2.1 ドキュメントの定義

本研究では入力テキストを先頭から解析処理をおこなう．また，先述したように入力テキストを文章群に分割して解析を行うが，今回は 1 ドキュメントを 20 文とした．1 つのドキュメントに対して解析を行なった後，ドキュメントの先頭を 1 文後ろに移動する．ただし，この方法ではテキストの先頭と末尾部分の処理回数が，他のテキスト部分と比べて少なくなるため，あらかじめ入力テキストには前後に 19 文の空白を追加する．図 1 はテキストに対して実際にドキュメント分割を行った様子である．ただし，紙面の都合上，1 ドキュメント 10 文としている．

2.2 話題分類単語辞書について

単語への話題の付与に用いる話題分類単語辞書は，ある単語に対して，その単語から連想される話題を記録した辞書である．なお，1 つの単語に対して，連想される話題が 1 つとは限らない．掲載されている単語数は 12,623 語であり，話題数は 228 個である．入力テキストからドキュメントへの分割を行なった後，この辞書を用いて，各ドキュメント内の各単語に対して，連想される話題を付与していく．

なお，単語への分割には，オープンな形態素解析エンジンである MeCab[3]を用いる．

[2]https://mmsrv.ninjal.ac.jp/nucc/
[3]http://taku910.github.io/mecab/

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

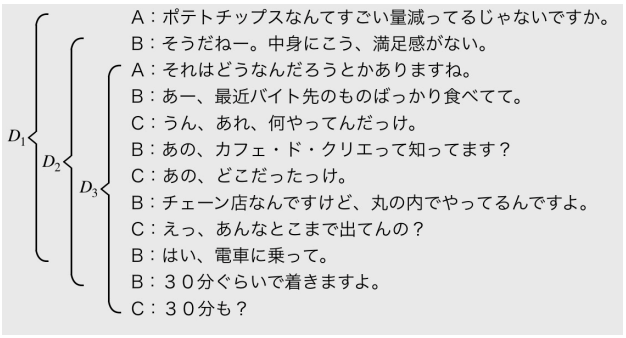

図 1: ドキュメント分割の例

2.3 TF-IDF による重み付け

テキストの解析には，TF-IDF を用いた重み付けを行う．TF-IDF のうち，TF とはある単語の文書内での出現頻度を表しており，式 (1) で求められる．式 (1) において，$\sum_{s \in d} n_{s,d}$ は文書 d 内のすべての単語の出現回数の和，$n_{t,d}$ はある単語 t の文書 d 内での出現回数を表している．

$$tf(t,d) = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}} \tag{1}$$

また，IDF はある単語の出現する文書数の割合の逆数であり，式 (2) で求められる．式 (2) において，$df(t)$ はある単語 t が出現する文書数，N は全文書数を表している．なお，$tf(t,d)$ と $idf(t)$ の積が TF-IDF 値となる．

$$idf(t) = \log \frac{N}{df(t)} + 1 \tag{2}$$

通常であれば，これらの値の算出に用いる文書は入力テキスト全体となるが，本研究ではテキスト内の各ドキュメントがこれに該当する．具体的な流れとしては，各単語の IDF 値を算出し，その後，各ドキュメント内の全ての単語の TF-IDF 値を算出する．

その後，後述の規則に従って，重みの高かった単語から連想される話題をドキュメントの話題とする．

2.4 ドキュメントの話題特定

各ドキュメント内の全ての単語に対して求めた TF-IDF 値を元に，各ドキュメントに対して大まかな話題を特定し，そこから詳細に解析処理を行うことによって，文単位での話題の特定を行う．

最初にドキュメント内の各単語の TF-IDF 値を算出する．話題分類単語辞書を用いて，各単語に対して，それから連想される話題を取得する．全ての単語の TF-IDF 値を話題ごとに合計した値を各話題の尤度とする．話題分類単語辞書において，1 つの単語から連想される話題が複数存在する場合，その単語の TF-IDF 値は，それぞれの話題の尤度として扱われる．各ドキュメントには，尤度が最も高かった話題が付与される．

図 1 の D_1 に対して上記の処理を行った様子を図 2 に示す．D_1 には，[食事・食品] の話題が連想される単語が 3 つ出現しており，それらの TF-IDF 値を合算した値 4.8 が D_1 における [食事・食品] の尤度となる．また，“バイト”という単語からは，[職業・労働] と [プログラミング] の 2 つの話題が連想されるため，それぞれの話題の尤度に“バイト”の TF-IDF が加算される．“電車”という単語は [鉄道・列車] 以外の話題は連想されず，他に [鉄道・列車] を連想させる単語も現れないため，“電車”の TF-IDF 値がそのまま [鉄道・列車] の尤度となる．これらの結果より，尤度が最も高い話題は [食事・食品] であるため，D_1 の話題は [食事・食品] となる．

D_1内での処理（ドキュメント毎にTF-IDF値が変化）

tfidf(“ポテト”) = 1.1
tfidf(“カフェ”) = 1.6
tfidf(“食べる”) = 2.1
} 尤度(“食事・食品”) = 4.8

tfidf(“バイト”) = 1.4 尤度(“職業・労働”) = 1.4

尤度(“プログラミング”) = 1.4

tfidf(“電車”) = 1.3 尤度(“鉄道・列車”) = 1.3

図 2: ドキュメント内の話題の尤度導出の例

同様に，D_2，D_3 についても，“カフェ”や“食べる”などの単語の作用により，ドキュメントの話題は [食事・食品] となる．

なお，ドキュメント内の単語だけでは話題分類辞書による話題特定が不可能な場合，そのドキュメントの話題は，1 つ前のドキュメントの話題が継続しているものとして扱う．

2.5 詳細な解析について

これまで，各ドキュメントに対しての話題付与を行ってきたが，最終的な出力は文単位で話題特定である．これを実現するために，尤度の低い話題に着目する必要がある．

まず，ドキュメント内で尤度の低い話題が出現した場合，その話題が以降のテキストで最も尤度が高く

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

なった場合は，最初にその話題の出現したドキュメントの末尾の文を話題の切り替わりとする．しかし，尤度の低い話題が，以降のドキュメントで話題として出現しなかった場合，即ち，尤度が 0 になった場合は，その話題及びそれを連想させる単語はテキスト内の話題の切り替わりに影響がないものとなる．図 3 に，具体的な例を示す．

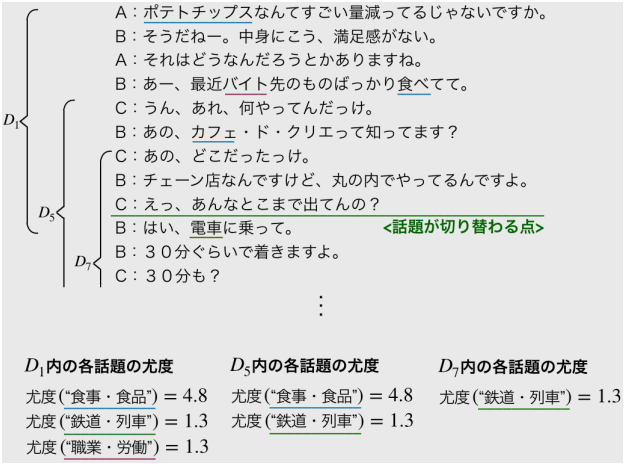

図 3: 詳細な解析の具体例

D_1 内で尤度が低い話題は，[鉄道・列車] と [職業・労働] の 2 つである．その後，D_5 では，[職業・労働] の話題を連想させる単語が出現せず，尤度が 0 になったため，“ バイト ”という単語と [職業・労働] という話題は，話題の切り替わりに影響がないものとなる．また，D_7 において，それまで尤度の低かった [鉄道・列車] が最も尤度が高くなったため，[鉄道・列車] の話題を連想させる“ 電車 ”という単語が初めて出現した D_1 の末尾の文が話題の切り替わりとなる．

3　実験結果・考察

話題付与の正確性評価には，手動で話題を割り当てた正解データとの比較をおこなう．辞書には話題分類単語辞書の他に改定版辞書を用いる．改訂版辞書は，本研究手法の精度向上を目的として，我々によって話題分類単語辞書の語彙不足や誤った話題の付与などの可能性を極力排除して構築したものである．改定版辞書は，話題分類単語辞書と同じ構造であるが，1 つの単語から連想される単語は 1 つの話題のみとなっている．例えば，“バス”という単語は，話題分類単語辞書では乗り物としての [バス] という話題の他に，[コンピュータ] や [音楽] などの話題が登録されている．このような単語に対しては，テキストに合わせて事前に人力で適切な話題を判断して辞書に登録した．登録されている単語及び話題は，評価用テキストに合わせ，91 個の単語と 32 個の話題のみとした．実際に名大会話コーパスのテキストファイルに対して評価を行った結果を表 1 に示す．

表 1: 話題特定の正解率

使用した辞書	正解率 [%]
話題分類単語辞書	27.80
改訂版辞書	62.85

結果は話題分類単語辞書を使用時の正解率が非常に低くなった．これは，1 つの単語から複数の話題が連想される場合，正確な話題を選択することが難しいためである．正解率の改善には，1 つの単語から連想される話題を一意に定めるか，辞書内の単語に優先順位を付与する等の対策を取る必要がある．

また，テキストの内容に準じた改訂版辞書を用いた場合，正解率が 6 割程度まで上昇した．

更なる精度改善には TF-IDF を用いた尤度指標判定手法と話題分類単語辞書の双方の改良が必要と思われる．文脈情報を活用した話題付与手法の実現，辞書の自動更新機能の実現，辞書形式の改良などを今後の課題とする．

参考文献

[1] 徳永健伸，情報検索と言語処理，東京大学出版会，1998.

[2] 桾澤優希，山本和英，“語の話題に基づく分類辞書の作成”，NLP 若手の会，第 11 回シンポジウム，2016.

[3] I. Fujimura et al. “Lexical and Grammatical Features of Spoken and Written Japanese in Contrast: Exploring a lexical profiling approach to comparing spoken and Written corpora”, Proc. the 7th GSCP International Conference. Speech and Corpora, p.393-398，2012.

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.