Web ニュースの主題語に着目した深層情報抽出手法

11371116　見塚　圭一（灘本研究室）

あらまし:インターネット上には時々刻々と様々なニュースが流れている. しかしながらこれらのニュースの中には, 主題となる人物や組織の詳細や関係性がわからないなど, ニュースの背景を知らないためそのニュース自体が理解できないことがある. 本論文では, この主題となる人物や組織をニュースの主題語と呼び, 主題語に関する Twitter にのみ存在する情報を抽出する手法を提案する.

１．はじめに

近年, インターネット上から情報を得る機会が増加している. また, 日々様々なニュースがインターネット上には流れている. そのニュースの記事内には, 主題となる人物や組織があり, これらの関係がわからないとニュース自体が理解できないことがある. 本論文ではニュース記事の主題となる語を主題語と呼ぶ. ニュース記事の内容を理解するためには主題語を理解することが必要であると考えられる. しかしながら, ニュース記事内では, 主題語について詳しく説明されているとは限らない.

一方, ニュースの主題語には, 有名人や企業等, Wikipedia 上に記事が存在している語が多数ある. Wikipedia の記事は, 基本的な内容であり且つ中立的な立場から解説[1]が行われている. そこで, 本論文では, Wikipedia に掲載されている情報を基本情報として扱う. 一方 Twitter 上にも様々な人がそのニュースに対して様々な解説を行っている. その中には Wikipedia に掲載していないが重要な情報や, 信憑性の怪しい情報など様々な情報がある. このような Wikipedia に掲載されていないが重要な情報はニュースを理解するのに重要であると考える. そこで本研究では Wikipedia には含まれていないが重要な情報を深層情報と呼び, 深層情報の抽出手法を提案する.

具体的には Twitter から主題語に関する解説を行っていると思われるツイートを抽出し, その中から Wikipedia に存在していない情報を抽出する. その抽出した情報から重要度を求めることにより, 深層情報を抽出する. この深層情報を抽出し, 提示することにより, ニュース記事内容の理解支援をすることが可能となる.

2. 提案手法

深層情報抽出の流れを以下に示す.

1. ニュース記事の主題語の抽出
2. 解説ツイートの抽出
3. 深層情報の抽出

2. 1 ニュース記事の主題語の抽出

ニュース記事を象徴する単語はタイトルまたは本文の一文目に出現することが多いと言われている[2]. このことから, ニュースの主題語候補として, ニュース記事のタイトルと一段落目に出現する名詞, または名詞が連続している場合に連結させた複合名詞を選択する.

各主題語候補に対して以下の式を用いて重要度を計算する. その計算には, 大原ら[3]の提案する式 1 を用いる. S_iはニュース記事における主題語候補の重要度である.

$$S_i = \alpha \times tf_{il} + \beta \times tf_{im} + \gamma \times tf_{in} \qquad [1]$$

ここでは, iはある単語を示し,lはタイトル, mは第一段落の文章, nは第二段落以降を連結した文章を示す. tf_{il}, tf_{im}, tf_{in}は, それぞれ単語 i の各文章における出現頻度を示す. 実験より, $\alpha = 0.9$, $\beta = 0.075$, $\gamma = 0.025$とする. そして, S_iを各単語ごとに求め, その値が一番高いものをニュース記事の主題語と定義する.

表 1：解説ツイートの抽出条件

No.	条件
1	名詞＋「は」または、名詞＋格助詞＋「は」の形が含まれる。
2	文末に基本形、過去形の表現が含まれる。
3	文末に思考を表現する動詞を含まない。
4	！，？，（笑）、ｗｗを含まない
5	ひらがなの小文字を含まない。
6	連続のひらがなを含まない。
7	一人称の名詞を含まない。
8	顔文字を含まない。
9	○○ー、○○～を含まない

表 2：思考動詞一覧

思う　わかる　考える　知る　感ずる 感じる　願う　期待する　予想する

2. 2 解説ツイートの抽出

ツイートには, 意見や解説, 雑談などが混在している. 一方, Wikipedia には, 解説のみが存在している. したがって, ツイートと Wikipedia を直接比較することが不可能であると考えた. Wikipedia と同じ条件で比較を行うために, 解説ツイートを抽出する.

主題語を含むツイートから解説ツイートを抽出するために表 1 に示す 9 つの条件を用いる. ここで, 表 1 の 3 の思考を表現する動詞を思考動詞と呼ぶ. 表 2 に思考動詞の一部を示す. そして, 表 1 に示す条件を満たすものを本論文では解説ツイートと定義し, 解

説ツイートの抽出を行う.

2. 3 深層情報の抽出

深層情報は, 2.2節で抽出した解説ツイートと主題語に対するWikipediaの記事と比較することにより抽出する. まずはじめに, Wikipediaの文書と解説ツイート群を混合してクラスタリングを行う. そして, そのクラスタ内の解説ツイートとWikipediaの文の比率により深層情報を決定する. このとき, Wikipediaの文書は主題語の記事を句点で分割したものを1文書として扱う. ツイートは1ツイートを1文書として扱う. 解説ツイートの文書数がWikipediaの文書数に比べ, 多い場合がある. そこで, クラスタリングを行う際に, Wikipediaの文書群と解説ツイート群の要素数に差がある場合には, Wikipediaの文書群の要素数に合わせる.

クラスタリングには, 短文にも比較的対応可能な[4]クラスタリング手法である, Repeated Bisection[5]を用いる. クラスタリングを行う際の品詞は, 名詞と動詞を用いる. これは, 名詞のみを用いると, その名詞が起こした出来事による分類が不可能になると考えられるからである. 動詞を用いることにより, 名詞が起こした出来事による分類が可能になると考えた為である.

クラスタリングの結果, 各クラスタを以下の3つのパターンに分類する.

(1)ツイートの比率がある閾値以上のクラスタ
(2)Wikipedia の比率がある閾値以上のクラスタ
(3)ツイートとWikipedia がほぼ半数のクラスタ

この内の(1)のツイートの比率がある閾値以上のクラスタに含まれている情報を深層情報として抽出する. (2)(3)は基本情報とする.

また, 重要なクラスタの抽出手法として, 山本ら[6]の提案するクラスタの密度を示す凝集性A_iを用いる. (式2参照)

$$A_i = \sum_{t \in C_i} \left(\frac{t \cdot c_i}{|t||c_i|}\right)^2 \qquad [2]$$

ここでは, i 番目のクラスタC_iのセントロイドc_iとそのクラスタに含まれるツイートtのコサイン類似度をクラスタC_i内のツイートごとに求める. 求めたコサイン類似度の平方和をクラスタC_iの凝集性A_iと定義する.

式2を用いて, 抽出した深層情報のクラスタに対して凝集性の計算を行い, その値が閾値以上のクラスタを本論文では重要な情報として抽出する.

3. 評価実験

提案手法の有用性を示す実験を行った. 実験手順は, まず, 2 章で抽出したある話題に関する解説ツイートを用いて, 人手でその解説ツイートが深層情報であるかを判定した. そして, 提案手法を用いて深層情報を抽出し, 再現率, 適合率, F 値を求めた. 実験結果を表 3 に示す.

表 3　実験結果

再現率	適合率	F 値
46. 2%	88. 6%	60. 7%

F 値が 60%となっていることにより, 提案手法を用いて深層情報の抽出ができることがわかった. しかしながら, 再現率が低く, 適合率が非常に高い結果となった. これは, 解説ツイートの中に含まれてしまった解説をしていないツイートを深層情報に含めたため, 正解データの数が多くなってしまったからと考えられる. また, ニュースの主題語を含むツイートには最新情報が反映されるため, Wikipedia の基本情報がツイートに含まれることが少ないのではないかと考えられる.

4. まとめと今後の課題

本研究では, ニュース記事内容の理解支援として, ニュースの主題語に対する深層情報の抽出手法を提案した. 具体的には, 主題語に対する情報を Twitter から抽出し, その情報を Wikipedia と比較することにより, 深層情報を抽出する手法である. 今後の課題として, 実験より解説ツイートがうまく抽出されていないことが判明したため, 解説ツイートの抽出手法の改善が必要である. そこで, 機械学習を用いて抽出することを行う. これは, 解説という定義のあいまいなものを条件により抽出しているので, 解説を行っているが解説ツイートとして抽出されないものが存在するためである. そのようなツイートを抽出するためには機械学習を用いる必要があると考えた. また, 深層情報がデマである可能性もあるため、デマ情報であるかを判定し, 削除をする必要があると考える.

参考文献

[1]https://ja.wikipedia.org/wiki/Wikipedia:五本の柱
[2]北山大輔, 角谷和俊, “ニュースアーカイブのためのコンテンツ構成順序を用いた比較ニュース検索”. 電子情報通信学会第18回データ工学ワークショップ, A9-4, 2007.
[3] 大原 正章 , 真下 遼 ,灘本 明代, ``Webニュースからの観点抽出手法の提案''. 研究報告データベースシステム (DBS), 2015-DBS-162,27,1 - 6,2015-11-19.
[4]花井俊介, 灘本明代, ¥酷似レシピ抽出のためのクラスタリング手法の提案”. 第6 回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2014), F8-6, 2014.
[5] Ying Zhao and George Karypis.“Comparison of agglomerativeand partitional doc-ument clustering algorithms.” Technicalreport, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 2002.
[6] Yuki Yamamoto, Tadahiko Kumamoto, Akiyo Nadamoto.” Followee Recommendation Based on Topic Extraction and Sentiment Analysis from Tweets”, the 17th International Conference on Information Integration and Web-based Applications & Services Article No. 27、2015.