1B2-NFC3-6

書き手の意見理解促進のためのアノテーション付与推奨箇所抽出手法

Recommendation system of annotation for understanding writer's opinions

伊藤彩 Aya Ito　西原陽子 Yoko Nishihara　大澤幸生 Yukio Ohsawa

東京大学大学院工学系研究科システム創成学専攻
Department of Systems Innovation, School of Engineering, The University of Tokyo

When reading documents, people often find writer's opinions which are new information for them. Making annotations on documents can be useful as help for finding this information. But making annotations on documents is a very hard work for people because they have to read documents carefully thinking where to make annotations and what to write as thought of them. Then, they can be eager about making annotations on documents by showing where to make annotations on automatically. We made an experiment to know about sentences which people make annotations on when they want to understand writer's opinions deeply. This paper proposes a method to extract sentences people should make annotations on to understand writer's opinions deeply. In this method, sentences including writer's opinions are extracted by words included in sentences. Then, by making annotations on the extracted sentences and writing their thoughts on, users can understand writer's opinions deeply.

1. 文書におけるアノテーション

人が文書化されたデータに接する時, 新たな発見が生まれたり, 新たな価値が見出されたりすることがある. 例えば, 読み手が文書を読んで, それまでは無かった知識を得たり, 理解できなかった理論を理解したり, 読み手には無かった書き手の発想に気づいたりすることがある. 本研究では, こういった読み手にとっての新しい情報のうち, 書き手の意見に注目し, 書き手の意見理解支援を行うこと, さらに, この支援に, アノテーションを利用することを考えた. アノテーションとは, 動画, 画像, 文書と言った様々なデータに対して付与する注釈を指す. 本研究では文書におけるアノテーションに限定して, 研究を進める.

文書におけるアノテーションとは, 元の文書に追加された書き込みを指す. 本研究では, 読み手がアノテーションを付与しながら文書を読み進めることを想定している. これは, アノテーションを付与しながら文書を読むと理解が深まるという考えに基づいている. しかし, アノテーションを付与する作業は大変手間のかかるものである. その理由として, 文書のどこにアノテーションを付与すべきか考えながら読む手間がかかることが挙げられる. また, 重要と思われる箇所に印を付け, 更に思考内容を書き込むという作業自体の煩雑さも否定できない. つまり, アノテーション付与の煩雑さを軽減しながらも, 文書にアノテーションを付与しながら書き手の意見理解促進を支援するシステムが必要である.

そこで本研究では, 文書の書き手の意見の理解促進を支援するために, アノテーションを付与すべき文を自動抽出し, 計算機のインターフェース上で行えるシステムを提案する. これにより, ユーザーの負担軽減と文書の書き手の意見理解促進を図る.

連絡先: 伊藤彩, 東京大学大学院工学系研究科システム創成学専攻, 〒113-8656 東京都文京区本郷 7-3-1, a-ito@panda.sys.t.u-tokyo.ac.jp

2. 関連研究

2.1 アノテーションによる支援

アノテーションによる支援手法の研究は, 様々な分野において行われている. まず, 映像に関するアノテーションの研究例が挙げられる [1] [2] [11] [12] [16]. これらの研究では, 映像にアノテーションを付与することで, 意味情報を付帯させている. 映像におけるアノテーションは, 閲覧者のコメントや, 映像の内容を示す語などである. また, これらをユーザー間で共有することで, 欲しい画像の検索を支援できる.

アノテーションは遺伝子配列の研究にも用いられており, 遺伝子配列に関連したアノテーションシステムも多々開発されている [4] [5]. これらの研究では, 遺伝子にその特徴的な性質を意味情報として付帯させ, 研究者間で共有することに意義がある.

文書に対してのアノテーションに関する研究も多々なされている [3] [6] [13] [14] [10]. これらの研究では, 文書中の, ユーザーにとっての未知語や, 文書に含まれる情報の確実性, 修辞構造を表すフレーズにアノテーションを付与することで, 文書の持つ情報をユーザーに分かりやすく提示する役割をアノテーションが担っている. また, それにより, ユーザー間のコミュニケーションが促進される場合もある.

このような研究に共通しているのは, アノテーションが意味情報付帯の役割を果たしていることである. 一方, 本研究は, 単なる意味情報の付帯を超え, アノテーション付与により書き手の意見をより深く理解することを目指すものである.

2.2 文書内容理解支援

一方, 文書内容理解支援に関しても, 様々な研究が行われている. 意見文抽出や要約作成により, ユーザーに情報を提供するシステムがある [8] [9] [10]. これらのシステムでは, ある商品についてのユーザーの感想や, 文書中の意見文と文書のトピックに関連した文をキーワード検索することで, 抽出した意見を集約して提示している. また, 文書から, その情報の信頼性, 客観性を判断する研究も行われている [13] [7] [15]. これらの研究では文書内容解析や情報発信者解析により, 読み手が文書を容易に理解するための支援や, 何か文書を探す際に, 目的に沿った文書を効率良く検索するための支援を行っている. これらの研究に共

通する特徴は, キーワード検索を行っていることである. 抽出されるキーワードは, 文書中の単語の, 文書のトピックとの関連度や, 賛成, 反対といった書き手の二極的な思考や, 推量や伝聞など文書における書き手の態度を表す単語, 文書の修辞構造を示す語などである. これらに対し本研究では, 単なる読解やキーワード, 賛成, 反対といった単純な意見の判別だけではなく, 書き手の意見をより深く理解するために注目して読むべき文を自動分析し, さらにユーザーの思考を促そうとする点が異なっている.

3. 書き手の意見を理解する際に見られるアノテーション付与箇所に関する予備実験

人が文書を読んで書き手の考えを理解しようとし, アノテーションを付与する文の特徴を調べた.

3.1 実験手順

実験では, 被験者に文書を読んでもらい, 文書上にアノテーションを付与してもらった. 実験に使用した文書は, 原発はごめんだヒロシマ市民の会の会報 No. 258 [*1], 原発はごめんだヒロシマ市民の会の会報 No. 260 [*2] の 2 つであった. 以下に No. 258 より, 文の例を示す.

> 山口での「１０・２６」は前倒しで１０月２４日の日曜日に上関町で行われました。その前日に新潟の地震が起こっており、原発阻止の思いひときわです。とくに今年は、１０月５日、中国電力が四代八幡宮と神社地の売買契約をしたことと、それをふまえて中国電力が、予定地の詳細調査の申請を山口県に提出することが予測されることから、「神社地売却疑惑糾明」「詳細調査断固拒否」がテーマです。
> この「１０・２４」には原水禁や市民個人が中国各地や九州からも、また韓国から、核廃棄物処分場計画反対の闘いをしている全羅北道扶安の李さん（男性）と通訳の金さん（女性）の参加があり、山口の山本由紀子さんと李さんが「君のための行進曲」という韓国の闘いの歌を歌うなど、上関が中国、九州、韓国とつながる集会となりました。

被験者には, まず, 原子力発電に関する知識の有無について質問した. 質問では, 単語群 1 チェルノブイリ原子力発電所事故 東海村原子力発電所事故, 単語群 2 応力腐食割れ BWR 原発 シュラウド, といった単語について, 内容を知っているかどうかを尋ねた. 単語群 2 の三つの単語を知っている場合に, 被験者は原子力に関する専門家であると判断した. 次に, 原子力発電の運転・設置に関する賛否等, 様々な立場の人の考えを理解する上で気になった箇所をマークし, その箇所について考えたことを書き込んでもらった. 被験者は大学生と大学院生合計 13 名であった. 単語群 2 について「知っている」と回答した被験者はおらず, 全員が原子力に関しては素人であった. 被験者が付与したアノテーション箇所と, 書き込まれた被験者の考えを元に, アノテーションの種類を分類した. その後, アノテーションが付与された文の特徴を, アノテーションの種類毎に調べた.

3.2 実験結果と考察

付与されたアノテーションと書き込まれた被験者の考えに注目すると, アノテーションは 6 種類に分類された. 表 1 は各項目に対する被験者 13 名分のアノテーションの総数を示している. 表 1 の文章の構造を理解する効果があったアノテーションとは, 接続や係り受け, 被験者が個人的に知らなかった単語など単純に文章の構造, 日本語の意味を理解するために付与したものであった. 例えば, 「玉虫色」という箇所に対し, 「玉虫色とは 7 色のことでしょうか？」と書かれているものがこれに相当する. 文書に書かれている事実を理解する効果があったアノテーションと, 文書に書かれている書き手の意見を理解する効果があったアノテーションは, 文書に明確に書かれている事実や書き手の感情を確認するために付与されたものであった. 前者の例としては, 「三団体 (現地, 市民ネット, 原水禁)」という文書中の箇所に対し, 「反対派」と書かれているものがあった. 後者の例としては, 「これまでずっとすべての話し合いによって全員一致で物事を決めてきたのに, この契約だけが多数決で決められることは受け入れられない」といった文に「これだけ多数決とはひどい話ですね」と書かれているものあった.

表 1　予備実験で得られたアノテーションの効果の分類, および各効果に相当するアノテーションの数

効果の分類	数
文書の構造を理解する効果	19
文書に書かれている事実を理解する効果	74
文書に書かれている書き手の意見を理解する効果	62
文書に書かれていない内容に対して、読み手が自身の意見考える効果（派生・主観）	41
文書に書かれていない内容に対して, 読み手が書き手の意見を考える効果（派生・客観）	21
被験者の意図が不明で, 効果が分からないもの	21
13 名の被験者が付与したアノテーションの総数	238

一方, 文書に書かれていない, 書き手の主観的な考えを理解する効果があったアノテーション (以下「派生・主観」アノテーションと記す) と文書に書かれていない, 書き手の客観的な考えを理解する効果があったアノテーション (以下「派生・客観」アノテーションと記す) は, 文章に明確に書かれていない内容に関し, 読み手が思考を及ぼしたものであった. 前者は読み手が自分の考えを, 後者は読み手が書き手や書き手に関わる人達の心情を推察した結果を述べているものであった. 「派生・主観」アノテーションの例としては「電力事業の自由化」といった箇所に「競争原理を働かせるのは会社にとって良いことだけれども, 価格競争で欠陥のあるインフラになったら問題だ」と書かれているものがあった. 「派生・客観」アノテーションの例としては, 「広島では, 」といった文章に「広島や長崎では”原子力” という単語だけで毛嫌いする人も多いのではないか」と書かれているものがあった.

表 1 の文書の構造を理解する効果があったアノテーション, 文書に書かれている事実を理解する効果があったアノテーション, 文書に書かれている市民感情を理解する効果があったアノテーションは, 文書に書かれていることを再確認しただけであった. 一方で,「派生・主観」アノテーションや「派生・客観」アノテーションは, 文書に明確に書かれていない書き手の意見の理解に踏み込んでいた. したがって, 書き手の意見理解に有効な読み手の深い思考は, 文書に明確に書かれていない内容に考えが及んでいる「派生・主観」アノテーションと「派生・客観」アノテーションに含まれると言える. これら 2 種類のアノテーションが付与された箇所を含む文の特徴を調べた結果, 文の文末について, 次のような特徴があることが分かった. 被験者 13 名の「派生・主観」アノテーションと「派生・客観」アノテーションは 66 文あり, そのうち以下の特徴のいずれかに該当する文は 46 文であった.

*1 http://www.geocities.jp/no_nukes_hiroshima/news258.html
*2 http://www.geocities.jp/no_nukes_hiroshima/news260.html

1. 文末が現在形である
2. 文末が推量形・否定形・疑問形・勧誘形のいずれかである
3. 文末が省略・倒置・体言止めなど、通常の文とは異なる特殊文末である

4. 実験

予備実験の結果を踏まえ, 書き手の意見を理解する上で, アノテーションを付与すべき文の自動抽出の可能性を調べるために実験を行った.

4.1 実験手順

被験者には, 文書の書き手の意見を理解する上でより深く考えるべきだと思う文に下線を引き, そこから新しく余白に線を引いて, その文について考えたことを記述してもらった. 文とは, 句点, クエスチョンマークまたはエクスクラメーションマークで区切られているものとした.

実験で使用したデータは, 原子力関連の文書 4 本, 毎日新聞コラム 4 本, 遺伝子組み換え作物に関する文書 4 本の、合計 12 本の文書であった. いずれの文書も社会的に正解が不明である課題を扱っていることから, 書き手の意見を考える材料としてアノテーションを付与させるにふさわしいと考えて選んだ. 遺伝子組換え作物に関する文書の例を以下に示す.

> スーパーに並ぶ納豆や豆腐を手にとって見ると「原材料：大豆（遺伝子組換えではない)」という表示。これを見て安心して購入する消費者は多いだろう。「遺伝子組換えなんて怖くて食べたくない」「遺伝子組換え作物を食べたラットが死んだ」「環境破壊の原因となっている」。色々と言われてきた遺伝子組換え食品は、そのイメージからか、日本ではいまだに商用栽培はされておらず、消費者からは敬遠されたままだ。昨年にはハワイの遺伝子組換えパパイヤの安全性が日本でも問題ないとされ、今年中には流通するかもしれないということを知っている人はどれほどいるだろうか。

被験者は大学生・大学院生 16 名であった. 各被験者には 3 種類の文書を一つずつ読んでもらった. 一つの文書を異なる 4 名に読んでもらった.

本実験では, 予備実験で得られたアノテーションが付与される文の特徴を元に, 以下の 3 つの仮説を立て, 仮説の検証を行うことにより, アノテーションを付与すべき文の特徴を調べた.

- 仮説 1: 書き手の意見を理解する上で深く考えるべき文の文末は現在形である
- 仮説 2: 書き手の意見を理解する上で深く考えるべき文の文末は推量形・疑問形・否定形・勧誘形のいずれかである
- 仮説 3: 書き手の意見を理解する上で深く考えるべき文の文末は体言止め・省略などの特殊文末である
- 仮説 1 &仮説 2: 書き手の意見を理解する上で深く考えるべき文の文末は, 仮説 1 かつ仮説 2 である

仮説に該当する文の例を挙げると, 仮説 1 に該当する文は「〜だ」, 仮説 2 に該当する文は「〜何が起こったのか！？」, 仮説 3 に該当する文は「悪魔の廃棄物。」, 仮説 1 且つ仮説 2 に該当する文は「〜かもしれない」である.

4.2 仮説の検証方法

仮説の検証は, 各仮説に相当するアノテーションの適合率と再現率の算出により行った. 適合率と再現率を算出するために, 文書中で各仮説に該当する文数を数えた. 次に, 被験者がアノテーションを付与した文で各仮説に該当する文数を数えた. そして, 各文書において, その文書を読んだ 4 人の被験者のアノテーション付与文数の各仮説ごとの平均値をとった. s: ある文書中である仮説に該当する文数, $\bar{a}$: ある文書中で被験者がアノテーションを付与した文で, ある仮説に該当する文数, $\bar{m}$: ある文書に対し被験者がアノテーションを付与した総文数とすると, 仮説の適合率は $P = \bar{a}/\bar{m}$ となり, 再現率は $R = \bar{a}/s$ となる.

表 2 各仮説に相当するアノテーションの適合率と再現率の平均

仮説	適合率 (%)	再現率 (%)
仮説 1	49.06	20.48
仮説 2	4.20	20.79
仮説 3	9.65	16.20
仮説 1 且つ仮説 2	16.76	39.54
どれにも該当しない	20.34	15.66

4.3 実験結果と考察

各仮説の適合率と再現率の平均を表 2 に示す. 仮説に該当しないアノテーションが付与された文の適合率に対し, 95% の t 検定を行った. 仮説に該当しないアノテーションが付与された文の適合率が 15% となった. この結果から, ユーザーは, 書き手の意見を理解しようとするとき, いずれの仮説にも該当しない文よりも, 仮説のいずれかに該当する文に注目する可能性が高いことが分かった. 仮説が成立した理由としては, 日本語が文末意志決定型言語であるため, 文章を書く際に, その意見が表れる文末形態に一定の特徴があるからだと考えられる.

各仮説に該当する文の再現率は高くなかった. これは, 被験者のアノテーション付与数が少ないためと考えられる. ユーザがアノテーションを付与する数には上限があると考えられ, その上限を考慮しつつ, 書き手の意見をより深く理解させるようなアノテーション付与推奨方法を考えて行く必要がある.

5. アノテーション付与箇所抽出システム

現在開発中のアノテーション付与箇所抽出システムについて述べる. 現状のシステムを, 図 1 に示す. システムのウィンドウ上に三つのテキスト表示スペースがあり, 左から順に, アノテーションを付与する文書の入力エリア, アノテーションを付与すべき文を色付けして表示するエリア, アノテーションを記入するエリアが配置されている. 図 1 上部には SHOW1 ボタンから SHOW4 ボタンまでが設置されており, 番号順にアノテーション付与優先順位が対応し, 該当する条件を満たす文が異なる色でハイライト表示される. 優先順位は, 実験結果の再現率を考慮して, 仮説 1 且つ仮説 2, 仮説 2, 仮説 3, 仮説 1 としている. 優先順位に従い, 黄色, 青, ピンク, 緑にハイライトされる.

6. まとめ

本研究では, アノテーション付与の煩雑さを軽減しつつ, 書き手の意見を理解するために, アノテーションを付与すべき箇所を抽出するシステムを提案した. 実験により, 人が付与するアノテーションの種類と, 読み手が書き手の意見を理解するために深く考える文の文末の特徴についての知見を得た. 今後は仮説を基にしたアノテーション付与推奨箇所の条件の絞込みなどシステムの改良を行い, ユーザの文書理解の支援効果を確認する評価実験を行いたい.

図 1　提案するアノテーションシステム．アノテーションを付与すべき箇所がハイライト表示される．

参考文献

[1] 山本大介, 増田智樹, 大平茂輝, 長尾確: 映像を話題としたコミュニティ活動支援に基づくアノテーションシステム. 情報処理学会論文誌, Vol. 48, No. 12, pp. 3624-3636, 2007.

[2] 山本大介, 長尾確: 閲覧者によるオンラインビデオコンテンツへのアノテーションとその応用. 人工知能学会論文誌, Vol. 20, No. 1, pp. 67-75, 2005.

[3] 藤井薫和, 重信智宏, 吉野孝: 機械翻訳を用いた異文化間チャットコミュニケーションにおけるアノテーションの評価. 情報処理学会論文誌, Vol. 48, pp. 63–71, 2007.

[4] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez and R. Apweiler: The Gene Ontology Annotation (GOA) Database: sharing knowledge in Unipot with Gene Ontology. *Nucleic Acids Research*, Vol.32, pp. 262–266, 2004.

[5] V. Ambros, B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun and T. Tuschl: A uniform system for microRNA annotation. RNA (Cambridge), Vol. 9, pp. 277–279, 2003.

[6] 東中竜一郎, 長尾確: アノテーションを用いて Web ドキュメントを分かりやすく提示する方法. 第 3 回インターネットテクノロジーワークショップ (WIT2000) 論文集, 2000.

[7] 松本章代, 小西達裕, 高木朗, 小山照夫, 三宅芳雄, 伊東幸宏: 文末表現を利用したウェブページの主観・客観度の判定. 第 1 回データ工学と情報マネジメントに関するフォーラム, 2009.

[8] B. Liu, M. Hu and J. Cheng: Opinion Observer: Analysing and Comparing Opinions on the Web. *WWW '05 Proceedings of the 14th international conference on World Wide Web*, pp. 342–351, 2005.

[9] L. Ku, Y. Liang and H. Chen: Opinion Extraction, Summarization and Tracking in News and Blog Corpora. *Proceedings of AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 100–107, March 2006.

[10] 綾聡平, 松尾豊, 岡崎直観, 橋田浩一, 石塚満: 修辞構造のアノテーションに基づく要約生成. 人工知能学会論文誌, Vol. 20, pp. 149–158, 2005.

[11] B. C. Russell, A. Torralba, K. P. Murphy and W. T. Freeman: LabalMe: A Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision, Vol. 77, No. 1-3, pp. 157–173.

[12] J. Jeon, V. Lavrenko and R. Manmatha: Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 119–126, 2003.

[13] 川添愛, 齊藤学, 片岡喜代子, 戸次大介: 確実性判断に関わる意味的文脈アノテーションの試み. 情報処理学会研究報告, No. 2, pp. 77–84, 2009.

[14] 松吉俊, 江口萌, 佐尾ちとせ, 村上浩司, 乾健太郎, 松本裕治: テキスト情報分析のための判断情報アノテーション. 電子情報通信学会論文誌, J93-D(6), pp. 705–713, 2010.

[15] 加藤義清, 河原大輔, 乾健太郎, 黒橋禎夫, 柴田知秀: Web ページの情報発信者の同定. 人工知能学会論文誌, Vol. 25, pp. 90–103, 2010.

[16] 桑原教彰, 桑原和宏, 安部伸治, 須佐見憲史, 安田清: 写真のアノテーションを活用した思い出ビデオ作成支援–認知症者への適用と評価–. 人工知能学会論文誌, Vol. 20, pp.396–405, 2005.