

# 種々の重要表現強調表示ツールバーの開発

村田 真樹\* 金丸 敏幸\* 白土 保\* 馬 青\*\*, \* 井佐原 均\*

\* 独立行政法人 情報通信研究機構 ( {murata,kanamaru,shirado,qma,isahara} @nict.go.jp)

\*\* 龍谷大学 理工学部 (qma @math.ryukoku.ac.jp)

## 1 はじめに

本稿では、われわれが開発した以下の重要表現強調表示ツールバーについて紹介する。

- タイトル強調表示ツールバー
- 特許文書強調表示ツールバー
- 原子分子重要表現強調表示ツールバー
- 副詞表現強調表示ツールバー

これらのツールバーは、重要な表現を強調表示することで、文書の閲覧の効率化をはかることができる。これらのツールバーは Windows マシンにおいてクリックをするだけで簡単にインストールできる。Web ブラウザの Internet Explorer のツールバーとして機能する。開発環境も簡便にしており、上記ツールバー以外の種類の強調表示用のツールバーの作成も容易な構成にしている。

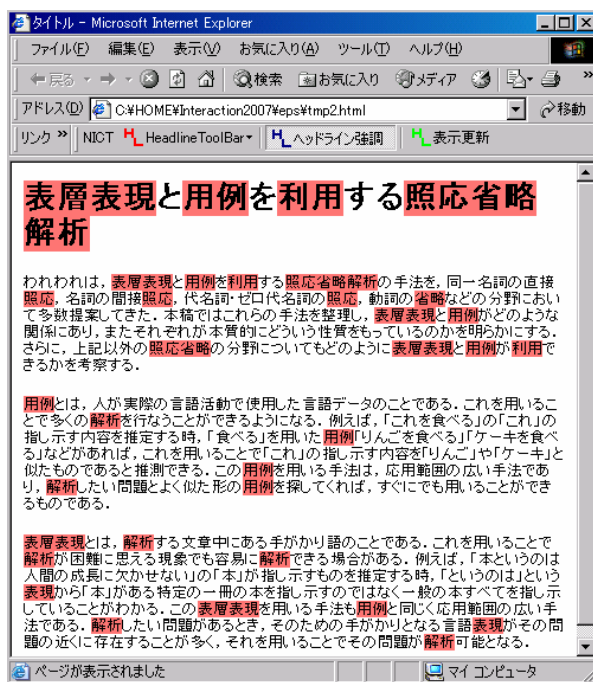


図 1: タイトル強調表示ツールバー

## 2 タイトル強調表示ツールバー

タイトル強調表示ツールバーは、Web 文書の H1,H2 タグなどを手がかりにタイトル部分を認識し、そのタイトル部分にあった単語を、タイトルと本文の両方で強調表示する [1]。ツールバーにボタンがあり、このボタンを押すことで、強調表示をしたり、強調表示をやめて通常の見出しにしたりできる。図 1 にタイトル強調表示ツールバーの例を示す。タイトルには重要な単語が使われるため、その単語を本文でも強調表示することで、本文において重要な単語を強調表示できる。強調表示された周辺の文書だけを読むことで、長い文章でも少しの範囲の重要な箇所だけを拾い読みすることができる。ここでは著作権の都合で見せていないが、Web 上の新聞記事を読覧する際にも本ツールバーを利用することができる。Web 文書に限らず、他の種類の文書でもタイトル部分を認識できれば、同様の強調表示をすることが可能である。

要約技術 [2] の一つの重要文抽出では、タイトルにあった単語を重要語として考えてそれらの単語をより多く含む文を重要文として出力する場合がある。本ツールバーでは、タイトルにあった単語を強調表示をするので、強調表示を多くしている文が、重要文抽出技術で取り出される文となる。すなわち、文章を短くするなどの要約は行っていないが、本ツールバーでは、強調表示を多くしている文を中心に読むことで要約と同じ効果が得られる。また、要約と異なり文章を短くするなど原文を変更する操作をしないので、場合によって、強調表示されている箇所の近辺の強調表示していない文も読み手の興味に応じて、手軽に読むことができる。元の文の情報を失わずになおかつ要約の効果も享受できる表示方法となっている。

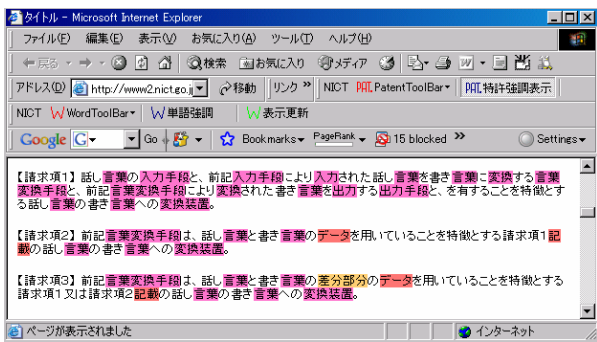


図 2: 請求項の強調表示

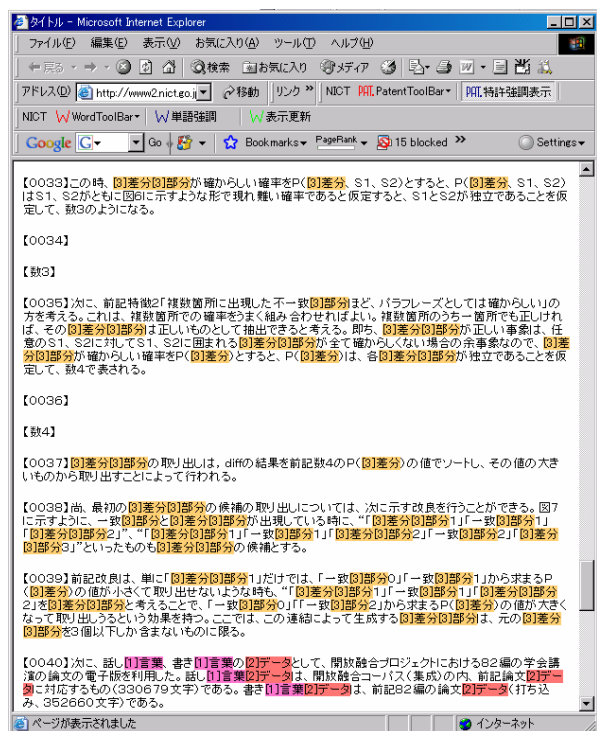


図 3: 実施例の強調表示

### 3 特許文書強調表示ツールバー

特許文書強調表示ツールバーは、特許の各請求項の特徴的な単語を、それぞれの請求項ごとの色で特許文書全体で強調表示するものである。例を図 2 と 3 に示す。各請求項で初めて出現した単語をその請求項の特徴的な単語とする<sup>1</sup>[3, 4]。図 2 では、請求項 2 ではじめて出現した「データ」が請求項 2 の特徴的な単語、

<sup>1</sup>各請求項だけで出現する単語をその請求項の特徴的な単語とするオプションも設けている。

請求項 3 ではじめて出現した「差分部分」が請求項 3 の特徴的な単語となる。他の単語は請求項 1 で出現したため、請求項 1 の特徴的な単語となる。これらの単語を特許本文でもそれぞれの色で強調表示する。図 3 に強調表示した特許本文の一部を示す。わかりやすいように取り出した請求項番号を [1][2][3] の形で強調表示した単語の頭に記している。0033 から 0039 の段落は請求項 3 の強調表示がなされ、これらの部分が請求項 3 の詳細な説明になっていることがわかる。このツールにより各請求項の詳細な説明の本文での箇所がすぐに特定でき便利である。

特許には複数のよく似た請求項を記述する場合がある。本ツールバーにより、それぞれの請求項ごとの特徴的な単語を色分けてして強調表示することで、よく似た記述の請求項であって各請求項の違いを容易に把握することができる。差分を抽出することはもどきの把握に役立つが [5]、本ツールバーでは、差分を抽出することを各請求項の違いの把握に役立てている。

また、各請求項の詳細な説明の本文での箇所を把握することで、特許明細書作成段階では、各請求項をサポートする実施例の記述が十分になされているかの確認を行うことができる。

また、特許明細書本文において、あまり強調表示がなされていない箇所を見ることで、そのあたりは請求項においてあまり記述していないことがわかる。特許明細書作成段階では、そのような箇所は、請求項に盛り込む内容（請求項に追加すべき内容）の候補として利用できる。そのような箇所の特定にも本ツールバーは役立つ。

本技術は、特許文書に限らず、簡条書きを含む文書であれば、同様の強調表示をすることが可能である。

### 4 原子分子重要表現強調表示ツールバー

原子分子重要表現強調表示ツールバーでは、原子分子分野の論文の重要な表現を、種類ごとに色分けして強調表示する（図 4）。原子記号や原子配置の式が強調表示され、その論文が扱っている原子などの重要情報 [6, 7] をすぐに把握でき便利である。論文を閲覧できる Web サイトがいくつかあるが、その Web ページを閲覧する際にも本ツールバーを利用することができる。

本ツールバーで強調表示する表現の種類は以下のと

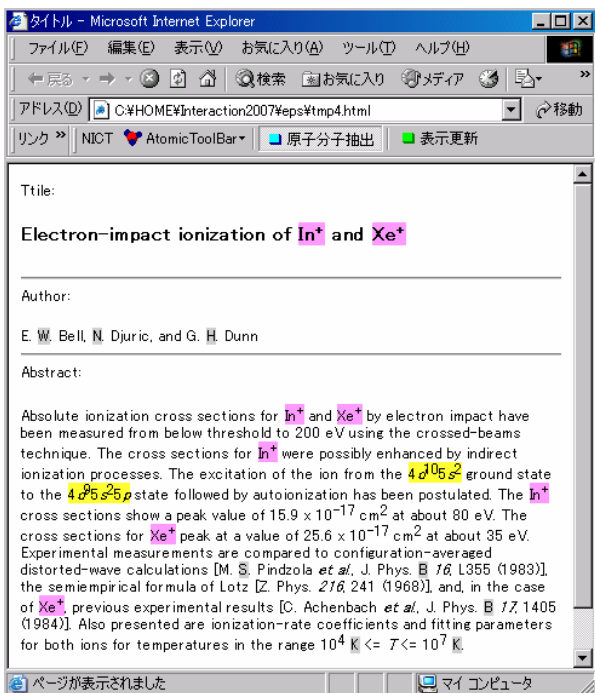


図 4: 原子分子重要表現強調表示ツールバー (表示論文の出典: E. W. Bell, N. Djuric, and G. H. Dunn, Phys. Rev. A48, 4286 (1983))

おりである.

1. 原子, 分子, イオンの表現 (桃色で強調表示)  
(例: helium, Li,  $O^{5+}$ ,  $Z=1$ )
2. 電子配置を示す表現 (黄色で強調表示)  
(例: 3p, 3s,  $3d^9 4s^2 4p 4f$ )
3. 微細構造を示す表現 (緑色で強調表示)  
(例:  $^1P$ ,  $^1S$ )
4.  $n=$  数字,  $l=$  数字 などの表現 (水色で強調表示)  
(例: "n=3, 4, and 5", "l=0,1,2")
5. 電子配置 + 数字 + l の表現 (橙色で強調表示)  
(例:  $1s2s2pnl$ )
6. 上記表現である可能性もあるが, 上記表現ではない可能性の高いもの. (灰色で強調表示)  
(例: "In the ..." の "In")

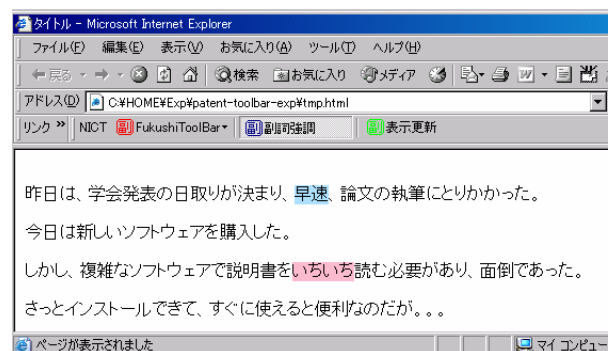


図 5: 副詞表現強調表示ツールバー

各表現の抽出には, perl の正規表現を利用した.

ここでは, 原子分子関係の重要表現を抽出して強調表示したが, 他の分野の重要表現であっても重要表現を定義すれば本ツールバーにおいて同様に強調表示をすることができる.

## 5 副詞表現強調表示ツールバー

副詞表現強調表示ツールバーは, 副詞に着目し, 肯定的な副詞を青、否定的な副詞を赤で強調表示する (図 5). 文章の書き手が持っているイメージが肯定的なのか, 否定的なのかをすぐに把握できて便利である. このツールバーは, 文献 [8, 9] に示す方法で作成した, 肯定否定に関する情報を含む副詞辞書を作成し, その辞書を利用して強調表示を行う.

本ツールバーは下記のような使い道がある.

- 他の人の書いた文書 (メールなどを含む) において, 本ツールバーに基づく強調表示を行うことで, その人が肯定的なイメージを持っているか否かを容易に把握できる.
- 自分の書いた文書 (メールなどを含む) において, 本ツールバーに基づく強調表示を行うことで, 自分が肯定的なイメージを持っているか否かを容易に把握できる.

また, 否定的なイメージの単語を使うとよくない場面では, 否定的副詞として強調表示されている箇所を中心に, 文章の修正をすることができる. 文章を修正すべき箇所を見つけるのに役立つ.

本技術を用いて副詞だけでなく、形容詞など他の品詞の単語でも、同様の強調表示をすることが可能である。また、最近、重要視されている、プログ分析、評判分析にも利用できる。

## 6 おわりに

本稿では、われわれが開発した以下の重要表現強調表示ツールバーについて紹介した。

- タイトル強調表示ツールバー
- 特許文書強調表示ツールバー
- 原子分子重要表現強調表示ツールバー
- 副詞表現強調表示ツールバー

これらのツールバーは、重要な表現を強調表示することで、文書の閲覧の効率化をはかることができる。

今後は、さらに様々な強調表示ツールバーの開発を続ける予定である。また、種々のツールの効果を確認する心理実験を行う予定である。本稿で示した技術は特許登録/出願中であり、権利を保有している。共同開発、事業を進める企業を募集している。 <http://www.nict.go.jp/x/x161/member/murata/technique/technique.html> を参照のこと。

本ツールバーは、種々の自然言語処理技術の成果を見せるための枠組みとしても便利な可能性がある。例えば、形態素解析の成果を見せる際には、文章の各単語に品詞をつけるようなツールバーを作るとよい。また、言い換えの成果を見せる際には、文章の言い換えをするツールバー（例えば書き言葉を話し言葉に変える、または、英語文で発音のしづらい  $r, l$  をあまり含まない単語をなるべく使うように言い換えるツールバー [10, 11, 12]) を作るとよい。自然言語処理研究者にとって便利なデモツールの枠組みとして利用できると思う。

## 付録

本ツールバー群は、上田マイニング工房「UMIK」 (<http://www.umik.net/>) のオンラインショップで販売することを検討している。種々のツールバー構築用としてより汎用的システムとしての販売も検討してい

る。本ツールバーの購入・利用、また共同研究をしたい方は著者 ([murata@nict.go.jp](mailto:murata@nict.go.jp)) にご連絡ください。

## 参考文献

- [1] 村田真樹, 竹内和広, キーワード強調装置及びプログラム, 特開 2004-280176, (2004).
- [2] Masaki Murata, Solution data edit processing apparatus and method, and automatic summarization processing apparatus and method, *US7120613*, (2006).
- [3] Masaki Murata and Hitoshi Isahara, Using the diff command in patent documents, *Proceedings of the Third NTCIR Workshop (PATENT)*, (2002).
- [4] 村田真樹, 文書差分検出装置及びプログラム, 特許 3682535, (2005).
- [5] 村田真樹, 井佐原均, diff を用いた言語処理 — 便利な差分検出ツール mdiff の利用 —, *言語処理学会誌*, Vol. 9, No. 2, (2002).
- [6] Akira Sasaki, Masaki Murata, Toshiyuki Kanamaru, Tamotsu Shirado, Hitoshi Isahara, Yutaka Ueshima, and Mitsuru Yamagiwa, Methods of extraction information on the atomic and molecular states from scientific abstracts, *Journal of Plasma and Fusion Research*, Vol. 81, No. 9, (2005), pp. 717–722.
- [7] 佐々木明, 村田真樹, 柏木裕恵, 城和貴, Lukas Pichl, 加藤太治, 村上泉, 科学論文からの原子分子データの収集、データベース化の自動化支援に関する研究, *言語処理学会第 12 回年次大会*, (2006).
- [8] 金丸敏幸, 村田真樹, 井佐原均, 話者の意図に関わる副詞辞書の構築, *言語処理学会第 12 回年次大会 併設ワークショップ (W1) 「感情・評価・態度と言語」*, (2006).
- [9] 金丸敏幸, 村田真樹, 井佐原均, 事態の捉え方に関する副詞辞書の構築, *日本認知科学会第 23 回大会*, (2006).
- [10] 村田真樹, 井佐原均, 話し言葉と書き言葉の diff, *ワークショップ「話し言葉の科学と工学」*, (2001).
- [11] 村田真樹, 井佐原均, 自動言い換え技術を利用した三つの英語学習支援システム, *情報科学技術レターズ*, Vol. 3, (2004), pp. 85–88.
- [12] 村田真樹, 井佐原均, 「言い換え」言い換えの統一モデル — 尺度に基づく変形の利用 —, *言語処理学会誌*, Vol. 11, No. 5, (2004).