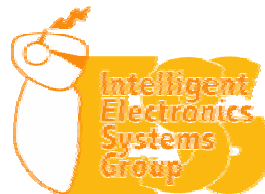


複数ブログ記事からの 関連事象を含む 概要記事の自動生成

井原健紘, 福富諭
電気通信大学

2007/03/06



まえおき – 個人的な興味

- 接続詞に興味あり
 - 目指す星は違う。が、仲良くしよう。
 - 文と文の繋がりを研究したい
 - とりあえず、文をそれっぽく羅列する
 - そんなアプリケーションを作った
-
- とりあえず羅列しただけで面白かった

今回の発表は接続詞とは関係ないです

もくじ

- 複数のブログを要約すると、ニュース記事っぽくなる
- それをウェブ公開してみた反応
 - 「情報処理は社会に何を与えるか？」
 - (今回の大会のスローガン)

背景 – 羅列された頻出単語

東電、横浜市、東京電力、江戸川区、日経平均、東西線、ピロウズ、リトルガッタス、警視庁、京葉線、福知山、献血、終戦記念日、天理、the pillows、東京メトロ、ゆりかもめ、千葉県

<http://d.hatena.ne.jp/hotkeyword>

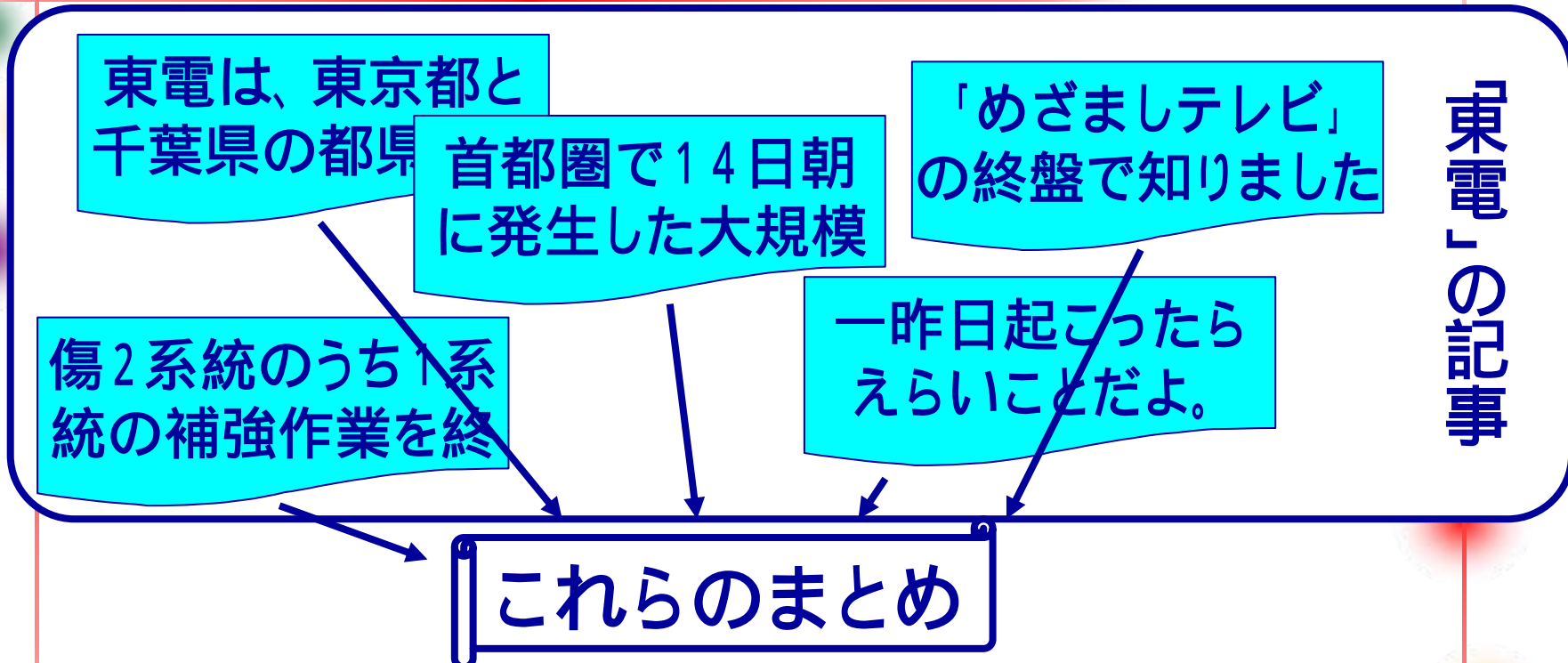
2006年8月15日(終戦記念日)の頻出単語
なぜその単語が頻出するのか
分からないものが多い

近頃よく見かけるウェブサービス

目的 – それがなぜ頻出単語か

- 8月15日、「終戦記念日」を差し置いて一位に輝いた「東電」。
- なぜ、「東電」がこの日のトップの頻出単語なのかを知りたい。
- ほかの頻出単語についても同様。
- 「なぜ頻出単語か」の情報をまとめた記事を自動的に作る。

方策 – 人手でするとしたら



頻出単語(東電)のニュース記事やブログをたくさん読む
それらをまとめる(要約する)



計算機に代わってもらおう

- 記事を集めるのは人間より得意
- まとめるのは人間に劣る
- 複数ブログ記事から重要文抽出

次ページ、とりあえず結果

- 東電は、東京都と千葉県の都県境を流れる旧江戸川にかかる送電線にクレーン船が接触、電線が損傷を受けたのが原因とみて調べている。
- 埼玉県狭山市で1999年11月に起きた航空自衛隊練習機の墜落事故では、練習機が送電線を切断し、約80万世帯が停電した。
- クレーン船の接触に伴う当社特別高圧送電線損傷による停電事故について。
- 送電線は千葉県の火力発電所から首都圏に電力を供給する27万5000ボルトの大動脈。
- 首都圏での大規模停電で、東電は14日、クレーン船の接触事故で損傷した電線を含む通常の送電ルートを、15日にも再開させる見通しを明らかにした。
- 今回、このうち2系統計3本の電線の金属部分が欠けるなどの損傷を受けた。
- 一昨日起こったらえらいことだよ。

複数ブログの重要文羅列

- まとめの記事は**簡易版ニュース記事**。
- 自分で作ってみてわりと便利だった。
 - その日のことが斜め読みで分かる。
 - 要するに自己評価は高い。
- なので、ウェブで公開した。
 - このことについては後述する。

手法 – 第一文

- 頻出単語 (東電) を含む段落を抜き出す

.....東電は、都と千葉県の
.....停電で、東電は14日.....
.....クレーン船の接触に東電は.....

- 単語に分割し、名詞の出現頻度を計算

東:48、千葉:24、電線:35、クレーン:11、、、

- 文に含まれる名詞の頻度の和をその文のスコアとする

➡ スコア = 87

当社特別高圧送電線損傷による停電事故について。

スコアの最も高い文が第一文

手法 – 第二文

- 第一文の手法とほぼ同様
- 標題の単語(東電)の頻度を負に
 - 東電: 48 東電: -48
 - 標題の単語のない文をなるべく選ぶ

手法 – 第三文

関係のある文全部

- 第一文の計算で頻度の大きかった名詞

東:48、千葉:24、電線:35、クレーン:11、、、

- これらの名詞を含む段落の全ての文

.....東電は、
都と千葉県の

.....停電で、
東電は14日.....

.....クレーン船の
接触に東電は.....

- 第一文と第二文に似ている(使用されている単語が似ている)文が第三文
- アルゴリズムの奇妙さについては後述

手法 – 第四文

- 第一文の計算で頻度の大きかった名詞

東: 48、 千葉: 24、 電線: 35、 クレーン: 11、 ..

- これらの名詞を含む段落の全ての文

.....東電は、
都と千葉県の

.....停電で、
東電は14日.....

.....クレーン船の
接触に東電は.....

関係のある文全部

- 第二文と同じアルゴリズムを適用
- アルゴリズムの奇妙さについては後述

手法 – 第五文・第六文

- 第三文・第四文に同じ
 - すでに選ばれた文と似ている文は棄却
- ここまでで、精度の悪い重要文抽出
 - 「関連事象を含む」 = 「精度が悪い」

手法 – 第七文

- 感想のような文を抽出する
- 形容詞と判定詞を含むことが条件
- 第一文とほぼ同様の基準で選択
- 「一昨日起こったらえらいことだよ」

手法についてのまとめ

- 多少手法が異なっても生成される要約の質はあまり変わらない
- 第三文から第六文はコーディング担当の共著者の勘違いによって作られた
- その勘違いに筆頭著者は三ヶ月間気づかなかった
- アルゴリズムはかなりどうでもいい

ウェブ公開 (06/8/14-07/1/5) ブログ

K-1

有明ではK-1MAX。▽

須藤元気が負けちゃったので、肩を落としつつ退散。▽

須藤元気が負けたのが残念だったけど・・・▽

※松本山雅FCは天皇杯第77回（1997年）大会以来2回目の出場。▽

K-1MAX・・・やっぱり世の中うまくはいきません。▽

第1回は1984年なんですって。▽

穴戸気の毒だったけどこれにはワタ。▽

[Permalink](#) | [コメント \(0\)](#) | [トラックバック \(0\)](#) 

堀江貴文

証券取引法違反（偽計・風説の流布、有価証券報告書の虚偽記載）に問われたライブドア（LD）前社長、堀江貴文被告（33）は4日、東京地裁（小坂敏幸裁判長）の初公判で「起訴状のような犯罪を行ったこともないし、指示したこともない。▽

ホリエモンが興じたヒルズ族<高級闇カジノ>一部始終 堀江貴文被告、ライブドア、バカラ賭博。▽

堀江被告「起訴は心外」、検察冒険は主導的役割を指摘。▽

さらに検察側は堀江被告が株式交換による企業買収を積極的に進めたかのように主張するが、積極的だったのは宮内被告だ。▽

ライブドア事件後のニュースや新聞記事などを見ていると、堀江被告が犯罪を犯しているようにしか見えないのですが、堀江被告は無罪を主張しています。▽



ウェブ公開 (06/8/14-07/1/5)

- 結果として、話題にはならなかった。
- 意見はわりと集まった。
 1. ソーシャルブックマークコメント
 2. ブログへのコメント
 3. ブログへのトラックバック
 4. 利用者数

ソーシャルブックマークコメント(タグ)

- [nlp][形態素解析][日本語][人工無脳]
- [automation][プログラミング][自動生成][技術][自動解析][コンピュータ]
- [blog][サイト][web][hatena][キーワード]
- [news][まとめ読み][あとでRSS登録]
- [idea][hack][くだる][neta][ネット社
会][right][へえ][まったく新しいタイプ] [literacy]
- [*fab][コメントしてきた] [berryz][石村舞波]

概ね作成者の意図どおりに捉えている

ソーシャルブックマークコメント(日本語1)

- はてなキーワードの言及先リンクから記事を自動作成 × 7
- なかなか面白い。 × 3
- こういうの流行り…？
- これを次々配信していくとおもしろいことになると思ったが。
- 何コレ

ソーシャルブックマークコメント(日本語2)

- 自ら文を組み立てるのではなく一文引用により処理しているので、人工無能にしては尤もらしいものに仕上がる。
- うお。こんなのがすでにあったのか。他人のブログを切り貼りしてブログエントリを自動生成するサイト。引用元がちゃんと書いてあるからぎりぎりありかな。
- プログラムが自動的に色々なエントリから一行ずつ切り出してエントリを作っている。キーワードごとにエントリが作られそれなりに見えて面白い。Googlezonを想起させるなあ。

冷静、且つ、好意的

ブログへのコメント

- 否定的意見もわずかにある
- 言外の意を汲んで要約すると、
 - 「人の文を勝手に盗むな」
 - 「ちゃんと自分で文章を書け」
- のあたり
 - 「手法を教えてくれ」
- というコメントもいくつかある

トラックバックでの言及

- 人間が書いたブログの1エントリーには遥かに遠く及ばない出来
- 日々のお話とそれに対する複数の人の反応を眺めるにはかつてなく良くできた仕組み
- ざっと眺めていて「お、そんなことがあったのか！」と目をひかれたことも

利用者数

- 不明。
- はてなアンテナでは20名程度。
- そのほかのRSSリーダーなどは不明。
- ページビュー：300 / 日

ウェブ公開に関するまとめ

珍しい

便利

不道徳

Youtubeかよ。

反省

- エンターテインメント性が皆無
 - 閲覧者は何も入力できない
- やはり、著作権が微妙
- 要約の方向性をしっかり考えていたら？

まとめ

- 複数の記事を要約してニュースを作成
- アルゴリズムはどうでもいい
- ウェブ公開(8月から1月)
- 情報処理は社会に何を与えるか？
 - 珍しさ・便利さ・不快感

今後の予定

- 気が向いたらスクリプトの公開
- 接続詞について考える
- 文そのものをいじる
- インタラクティブなものを作る

個人的な雑感

- 自然言語の要約技術は、
- 思った以上に、
- 世間に認知されていないらしい。