

自然言語処理のための ウェブ→テキスト変換

河原 大輔
情報通信研究機構



NLP若手の会第1回シンポジウム (2006/09/15)

背景



- 情報爆発
<http://i-explosion.ex.nii.ac.jp/i-explosion/>
- 情報大航海

超大規模テキストが欲しい

- 新聞記事 (毎日, 読売, 朝日, 日本経済, ...)
 - 1年分10万円程度
 - ドメイン: 新聞
- **ウェブ: 日本語10億ページ**
 - サーチエンジンで検索: 単語の出現ページ数
 - **やっぱり言語処理したい**

ウェブから収集したコーパス

- Webコーパスの提案 (関口, 山本, 2003)
 - 220MB
- NW1000G-04 (NTCIR-5, Webタスク)
 - 1億ページ
- 大規模Webコーパス (河原, 黒橋, 2006)
 - 5億文
- ウェブアーカイブ (田村, 喜連川, 1999へ)
 - 9億ページ以上

Webコーパスの提案 (関口, 山本, 2003)

- HTMLタグや文字種の情報を用いて良質な日本語文を抽出
- 220MBのコーパスを構築 (新聞10年分ぐらい?)

NW1000G-04 (NTCIR-5 Webタスク)

- 1.4TB, 1億ページ
- 2004/01〜2005/01にかけて収集
- 主にjpドメインから収集: 50%がjpドメイン
 - **70〜80%が日本語ページ**

ウェブコーパス (河原, 黒橋, 2006)

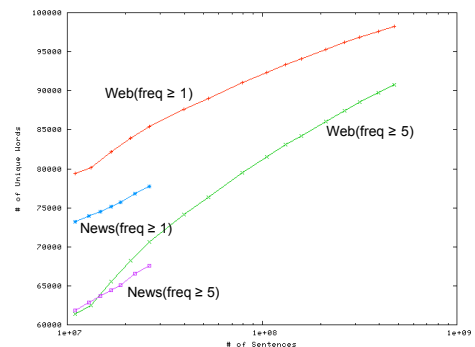
- 1.3TB(圧縮), 4億ページ(各種言語)
- 2004年春に収集に1カ月で収集
(by 東大田浦研)
- 日本語文を抽出
– 5億文

ウェブコーパスの作成手順

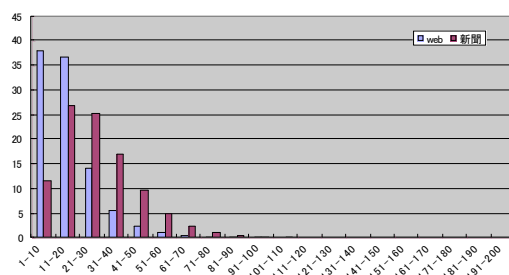
1. クローラによるページ収集(約4億ページ)
 2. エンコーディング情報による日本語ページ候補抽出
 - charset属性, perl Encode::guess_encoding()関数
 3. 言語情報による日本語ページ判定
 - 助詞の文字(が, を, に, は, の, で)を0.5%以上含む
 4. ページからの文抽出(HTMLタグと句点)
 5. 日本語文の抽出
 - ひらがな, カタカナ, 漢字を60%以上含む文
 6. 重複文の削除
- 約5億日本語文
(日本語として妥当な文: 995文/1,000文)

しょうがないので駅のレストランで食事をしようとした所、1日数本しかない山田線の存在に思い当たる。
もれなくプレゼント！
でも僕はTシャツの上に長袖のシャツ。
今回は某アイドルの高橋一也も参加したので客が若い。
団体Aが「まちづくり」をテーマにインターネット上で公開講座を開催しようとしている。
htaccessを置いたとたんそのディレクトリ以下で。
昨年の没後400年祭を機に復元した井戸を紹介する木下さん
恋は、真剣勝負。
ほめ言葉が多くなって嬉しいですね。
開校式並びに入学式を挙行、初代校長佐治勝弥、職員10名冲館小学校校舎一部を併用す。
佐治勝弥校長育一中学校長に任命される。
いまだに言うでしょう。
「買いいば」を見た伝えれば、お買い上げ合計金額より5%引きいたします。
政治も危機的状況ですし、物資も不足しています。
そういう長期的な存在理由とか、長期的なビジョンとか、何故ここが国のお金で、我々の税金でやらなければならないのか、その辺を評価する上で何かお考えになられていますでしょうか。
河北郡津幡町南中条・バリアフリー対応の学校案内や生徒会活動の紹介など
思いやりのある優しい子に育ってネ。
工学的諸問題に対処する際に必要な、線形代数・解析・確率・統計などの数学に關する知識を理解できます。

コーパスサイズと異なり単語数の関係



文長(単語数)の分布



ウェブ→テキスト変換

1. HTMLを入力
2. エンコーディング情報により日本語ページ判定
 - charset属性, perl Encode::guess_encoding()関数
3. 言語情報により日本語ページ判定
 - 助詞の文字(が, を, に, は, の, で)を0.5%以上含む
4. 文分割(HTMLタグと句点)
 -
, <P>, <PRE>中の改行, ...
5. 日本語文をマーク
 - ひらがな, カタカナ, 漢字を60%以上含む文

問題

- 特殊文字
 - 参照
 - e.g., «
- 文区切り
 - 「!」の扱い
 - e.g., Yahoo! Japan - ヘルプ
 - 謎なHTML
 - e.g. 地元活動は、安定的にミニ集会を連日とりおこなうなど、いわ

解析

- 形態素解析 (97～99%)
- 固有表現認識 (80～85%)
- 構文解析 (87～90%)
- 格解析、述語項構造解析
- 省略・照応解析
- …

最近のできごと

- 自動構築した格フレームの公開
 - 検索: <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/caseframe/>
- ウェブコーパスの配布
 - 生, 構文解析済み, 格解析済み(部分的)
 - ハードディスク送付→コピー後返送

今後

- 多言語展開
 - 英語は対応済み
- ツールの公開
- というか、みなさんどうしているのでしょうか?