

2010. 3. 8
報道発表資料

大規模ブログデータの研究を開始
— 「Yahoo!ブログデータ」の研究利用による
言語研究の新展開—

東倉洋一・大山敬三

国立情報学研究所

報道発表の主なメッセージ

- 「ブログコーパスの研究目的利用ガイドライン」を策定
- Yahoo!ブログデータを提供開始（2010.4予定）
- 新しい言語を対象とした新しい言語研究の開始

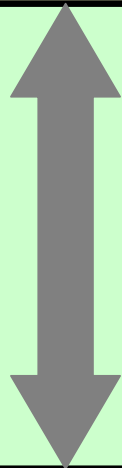
情報爆発の現状

- 5年で、10倍に
- 2011年に1800エクサバイト (EB)
【kB, MB, GB, TB, PB】

メディアや情報源の多様性

新しいWeb言語等を研究対象に

書き言葉



話し言葉

- ニュースサイト、ネット小説、
学術論文、書籍関連データ、
報告書、説明広報資料……
- ネット日記
- 質問・回答文（Q&Aサイト）
- **ブログ**
- 電子メール
- 携帯メール

研究用Webデータベースの必要性

【従来の問題点】

- プロバイダー等の情報サービス提供者の企業情報としてのみ利用され、非公開扱い

【問題解決への第一歩】

- ヤフー株式会社の「Yahoo!知恵袋」の提供

【研究コミュニティの強い要望】

- ヤフー株式会社の「Yahoo!ブログ」の提供による研究コミュニティへの貢献

「Yahoo!知恵袋データ」とは？ (2007年4月に提供開始)

- 質問したい人と回答したい人をむすび、
知恵と知識を参加者同士で共有
- 日本最大の知識検索サービス
- 2004年4月～2005年10月の質問総数
300万件、回答総数1300万件以上
- 投稿者数約24万人

ブログの言語研究への利用意義

- 社会・文化・行動現象の分析
- マーケティング
- 世論調査
- 知識の発見・獲得

ニーズ

新しい言語モデルと言語解析ツールが必要！

「Yahoo!ブログデータ」とは？ (2010年4月に提供開始)

- 2008年4月～2009年1月に投稿されたデータから、約500万語、2万4千記事のサンプルを抽出
- 個人情報，誹謗中傷等を人手で削除
- インターネット全体に公開されている記事のみを対象

【例1：音楽レビュー】

そのライブは、まるでサーカス。
Piano, Bass, Drumsの三人が、これでもかと思いた
こともないテクニックを繰り広げ続ける。
3人の緊張感と、恍惚とした表情、これ以上に気持ち
いい瞬間はないという笑顔。
何を見ても、何を聴かされても、「すごいすごいす
ごいすごい！」と口の中で小さくつぶやくしかな
かった。
これほどまでに驚かされたライブパフォーマンスは、
たぶん、はじめて。

【例2：日記（ケータイ）】

後輩と、一風堂経由のダーツ。

久々にラーメン食べて幸せ(´▽`)

一風堂以上に久しぶり(つ～か年単位で久しぶり)のダーツは、シャフトが折れるくらい白熱w
戦績は3勝4敗。

後輩がすんげ～ウマくていろいろ教えてくれて、超楽しかったです(´▽`)

ブログの言語解析の問題点

- 文の区切りが不明確
- 顔文字などの不要な文字列が混入
- くだけた文体による形態素解析、構文解析の誤り

新言語モデル
新言語解析ツール

自然言語処理

形態素



構文



意味



文脈

処理の深さ

ブログデータの特徴

- CGM→さまざまな情報を含む
 - 個人情報やプライバシー
 - 誹謗中傷, 卑猥表現 . . .

⇒ 利用には一定のルールが必要

☆「ブログコーパスの研究目的利用ガイドライン」の策定

(NII企画型共同研究「大規模テキストコーパス整備における個人情報等取り扱いの検討」)

ブロクの取り扱いガイドラインに関する共同研究

【目的】個人の特定に結びつく可能性のある表現や反社会的な表現などの情報の取り扱いを検討し、ガイドラインを策定する

国立情報学研究所	コンテンツ科学研究系・教授(代表者)	大山 敬三
	副所長・教授	東倉 洋一
	コンテンツ科学研究系・教授・学術基盤推進部長	安達 淳
	情報学研究データリポジトリ推進グループ	大須賀 智子
東京大学	生産技術研究所・教授	喜連川 優
東京大学大学院	情報理工学系研究科・教授	辻井 潤一
	情報理工学系研究科・教授	石塚 満
	情報理工学系研究科・社会連携担当	木戸 冬子
京都大学大学院	情報学研究科・教授	黒橋 禎夫
東京工業大学	精密工学研究所・教授	奥村 学
国立国語研究所	国立国語研究所・言語資源研究系・教授	前川 喜久雄
	国立国語研究所・言語資源研究系・准教授(H21年度)	山崎 誠
ヤフー株式会社	ソーシャルネット事業部・企画部・リーダー(H20年度)	堀下 剛司
	ソーシャルネット事業部・企画部(H20年度)	堀野 亜紀
	メディア事業統括本部・メディアサービス本部ソーシャル企画部・チームリーダー(H21年度)	寺岡 宏彰
	メディア事業統括本部・メディアサービス本部ソーシャル企画部(H21年度)	東保 知子

ガイドラインの概要（その1）

- 目的:
 - ブログコーパスを用いた研究における、個人情報や反社会的情報の取り扱いに係わる問題発生防止
 - コーパス提供者の理解と信頼の獲得
⇒ コーパス利用者の心得を定める
- 対象：
 - 言語学，自然言語処理，情報検査など
 - それ以外については別途に要検討

ガイドラインの概要（その2）

- 内容:
 - データの保管, 利用者の管理
 - 研究成果等としての公開の禁止など
 - 個人情報・プライバシー侵害情報（禁止）
 - 誹謗中傷, 脅迫, 人権侵害表現（禁止）
 - デマ, 詐欺, 有害情報等（読者に配慮）
 - 提供者等が削除した情報の復元禁止
 - 問題情報の追加削除
 - 対応するインターネットデータの取り扱い

データ提供と利用

- ガイドラインの公表
- ガイドラインに基づくYahoo!ブログデータの一部（500万語分）の処理（国語研）
 - 約6億語分のデータの機械処理について継続して共同研究
- 研究者へのブログデータ提供（NII）
 - 500万語、24000記事分（2010.4予定）
 - 約6億語、350万記事分（2010年度中予定）
- KOTONOHAへ搭載して公開（国語研）
- NIIの情報関連研究プロジェクトでの利用
 - NTCIR:情報検索・アクセス技術の比較と性能評価のための研究基盤関連, 他

まとめ

- ブログデータの取り扱いに関するガイドラインの検討・策定
- Yahoo!ブログデータの提供開始
- 新しい言語モデル・言語解析ツールの創出