

大規模ブログデータの研究を開始

ー「Yahoo!ブログデータ」の研究利用による 言語研究の新展開ー

東倉洋一・大山敬三
国立情報学研究所

1

報道発表の主なメッセージ

- Yahoo!ブログの研究コミュニティへのデータ提供契約の締結
- 国立情報学研究所より研究コミュニティに無償提供(2008.7予定)
- 新しい言語を対象とした新しい言語研究の開始

2

情報爆発時代



3

情報爆発の課題

- 欲しい情報が探せない
- 人気と重要性が異なる
- マイノリティ情報が埋もれる
- 情報が死蔵される
- 情報の信頼性、信憑性が不明
- 情報の保存が難しい
- 情報保有・提供による社会支配リスク

4

情報の多様性に着目

- テキストから多様な情報メディア（イメージ、サウンド）
- 画像から映像まで
- 音響、音楽から音声まで
- テキスト（言語情報）の多様性への対応

5

問題解決に向けた重点プロジェクト （国家施策としての取組み）

情報爆発時代に向けた新しいIT基盤技術の研究
（科研費特定領域）

- 非順序型実行原理に基づく超高性能データベースエンジンの開発（文科省）
- 情報大航海 プロジェクト（経産省）
- 情報信憑性検証技術（総務省）

- 
- 情報の巨大集積化と利活用（科学技術連携施策群）

6

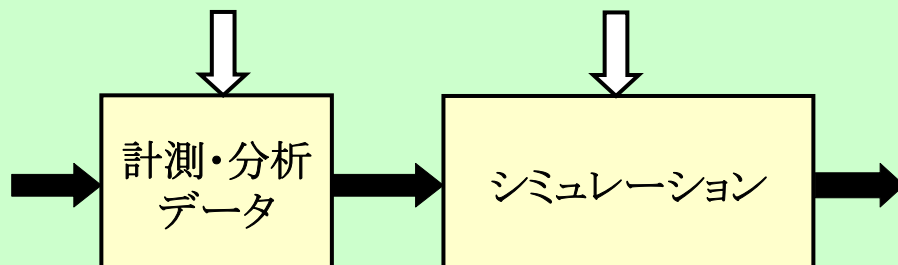
研究アプローチの変遷

- 理論研究
- 実験研究
- 計算論的研究
- データ中心科学研究 (Data-centric)
 - ◆ センサ(ネットワーク)と計測
 - ◆ スーパーコンピューティングとシミュレーション

7

データの量と質の飛躍的增加

- 新センサー
- 先端計測装置
- 先端分析装置
- モデル
- スーパーコンピューティング



8

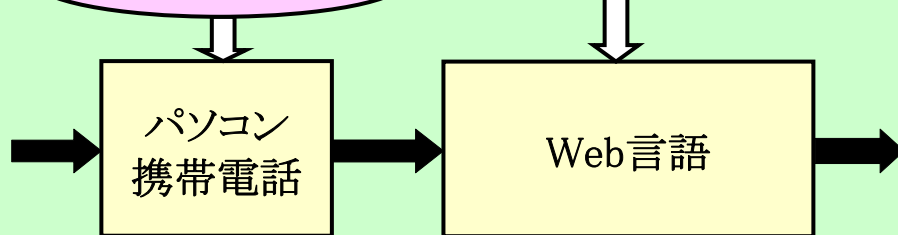
データ中心科学としての言語研究

- 視る・読む・聴く
- 考える・思う
- 創る・伝える

1億総クリエイター時代

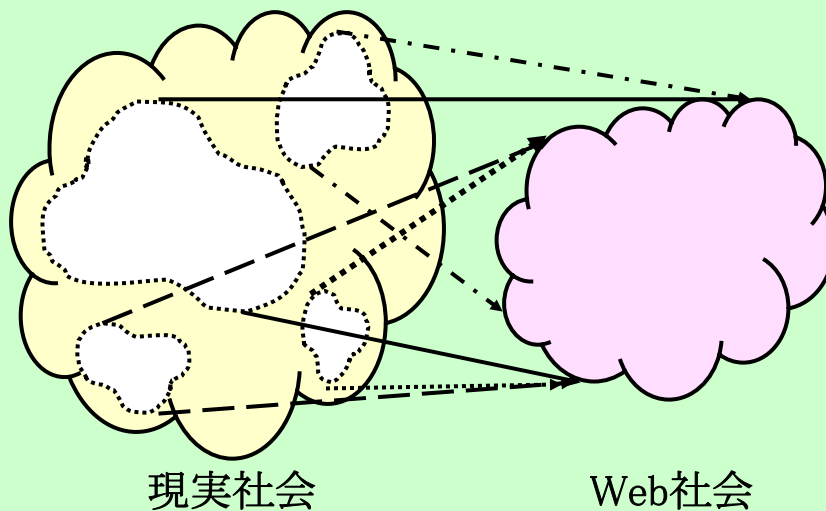
- 言語モデル
- 言語解析ツール

人間センサ



9

Web社会は現実社会の「鏡」



10

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

11

「Yahoo!知恵袋」とは？ (2004年4月にサービス)

- 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
- 質問総数300万件、回答総数1300万件以上
- 参加登録者約24万人

12

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！

13

新しいWeb言語等を研究対象に

書き言葉

- 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

話し言葉

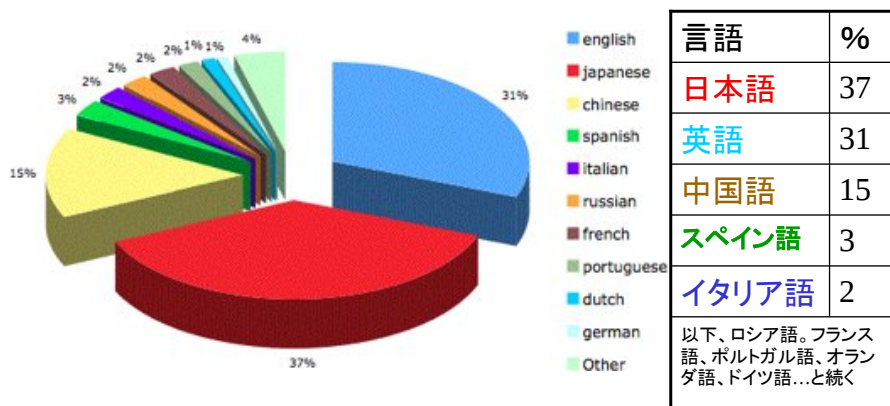
14

「Yahoo!ブログ」は？ (2008年7月に提供開始)

- 4月25以降に投稿されたデータから、最低500万語のサンプルを抽出
- インターネット全体に公開されている記事のみを対象
- 季節毎に数回、サンプル抽出

15

言語別ブログ記事数



Technorati調査(2007.4)

16

【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、これでもかと見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。

17

【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(´▽`)

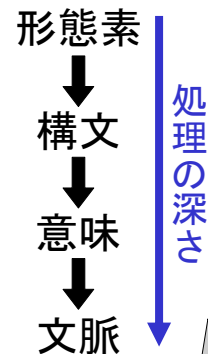
18

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



19

NIIからのデータ提供と利用

- 情報検索、情報分析、情報活用などの研究目的利用者へのデータ提供 (2008.7 予定)
- NIIでの情報関連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での利用

20

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
- 新しい言語モデル・言語解析ツールの創出

21

ブログ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などの情報の取り扱いを検討し、ガイドラインの策定を目指す

国立情報学研究所	コンテンツ科学研究系・教授	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
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・特任教授	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・准教授	奥村 学
国立国語研究所	研究開発部門・グループ長	前川 喜久雄
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	堀下 剛司
	ソーシャルネット事業部・企画部	堀野 亜紀

22