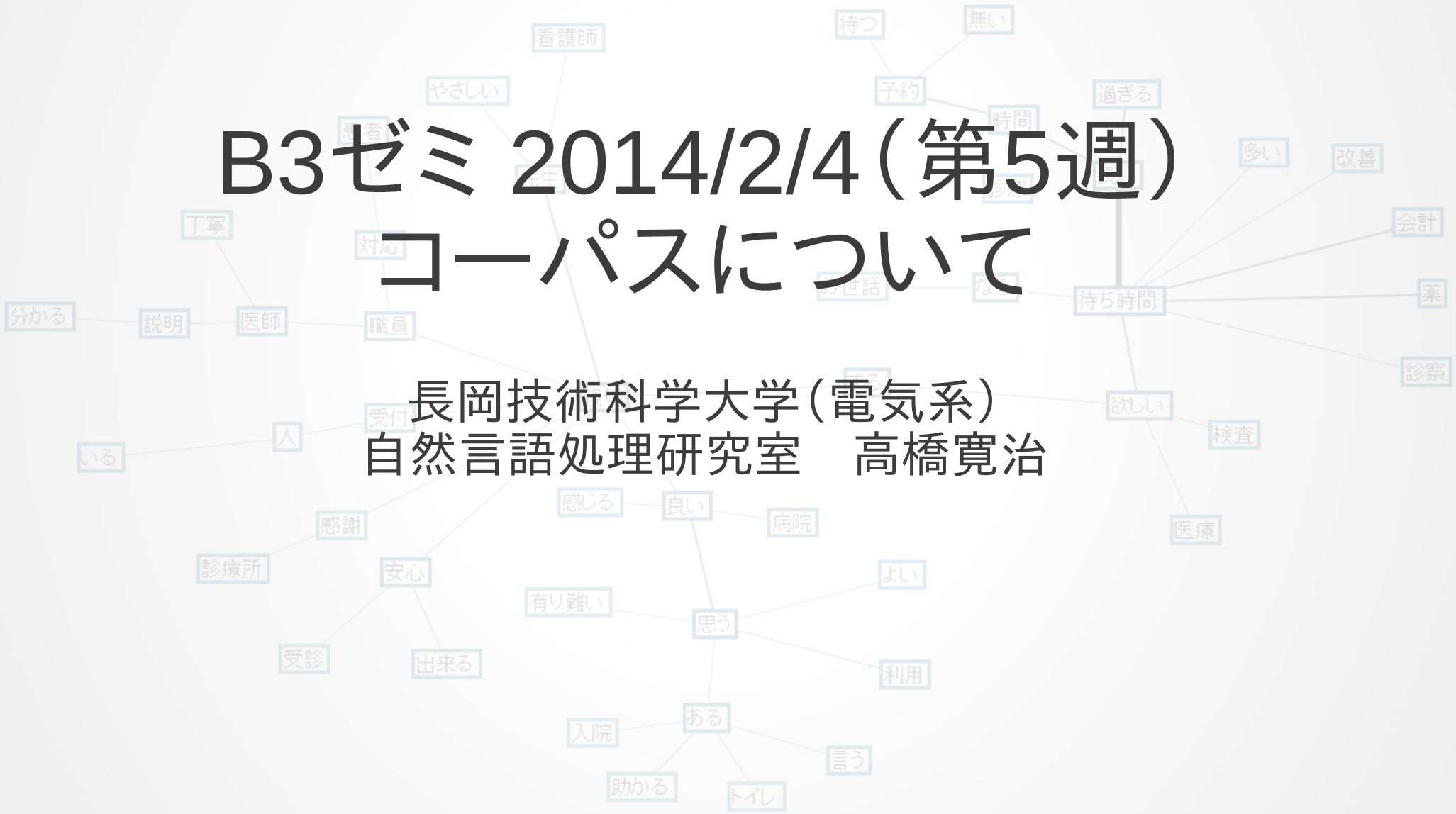
B3ゼミ 2014/2/4（第5週）
コーパスについて

長岡技術科学大学（電気系）
自然言語処理研究室　高橋寛治



はじめに

- 「ベーシックコーパス言語学」の本にそって勉強資料を作成
  - 石川慎一郎　著
  - 応用言語学

- 2回のゼミで、終了予定

- 自分が自然言語処理をする側ということを忘れない

第3章　さまざまなコーパス

- 研究の目的やスタイルに応じて様々なタイプのコーパスが開発されている
  - データの収集法、収集データ内容、研究対象
- ただし分類は研究者独自のもの
  - 一般コーパス、標本コーパス
- 通例のコーパスは、個別言語のスナップショットの縮尺図
  - 母集団のデータを均衡的に取捨選択して作られた言語標本
- コーパス
  - Brown Corpus, British National Corpus, Bank of English, 現代日本書き言葉コーパス、青空文庫

第4章　コーパスの作成

- 研究目的によって、自分で言語データを集めてコーパスを作る
- 自作コーパスによる調査を予備研究として位置づけ、大型コーパスを用いた研究につなげる
- 流れ
  - データの収集
  - データの電子化
  - データのアノテーション

4.1 データの収集

- コーパス設計
  - 母集団を定め、どのようなデータをどう集めるか
  - 収集データの選定
    - 母集団の定義
    - 層化観点の決定
    - 層間比率の決定
    - 無作為抽出の実施
    - 著作権への配慮
    - 標本の収集
    - サンプルサイズの決定
    - 属性情報の記録

4.2 データの電子化

- 汎用的環境で分析可能にするために
  - ファイルタイプの統一
    - テキストファイルが原則
      - 文字以外の情報が含まれていない
  - 文字コードの統一
    - 文字化けを防ぐ
      - SJIS, EUC-JP, Unicode
      - 改行コード　CR, LF, CR+LF

4.3 データのアノテーション

- 詳細な機械検索を行うためにアノテーション
- 品詞タグ付け
  - 言語学的な情報や解説を書き込む、例　品詞タグ
- 英語テキストの処理
  - 品詞タグ付け
- 日本語テキストの処理
  - 形態素解析

- 本の内容は以上でおわり
  - 5章以降は、コーパスの扱い・研究についてなので

- ここからの内容は論文から、自然言語処理におけるコーパスに関して

ちょっと休憩

＊論文内容とは関係ない

- 読める？

  これは　しんぜんげごしりょで　りよう　される　コパースに　ついて　まめとた　プレンゼテーション　です

- 単語の中の文字の順序を変えても読めてしまう？

このスライド以降の説明に参考・引用した論文について

- 自然言語処理の再挑戦～統計的言語処理を越えて～
- 乾　健太郎、　浅原　正幸
- 日本知能情報ファジィ学会誌 2006年10月 No.5　特集：テキストの可視化と要約より

4　入手可能な資源とツール

- 言語処理はラベル付け問題
  - 自分で設計したラベルを付与したり言語解析ツールの出力を修正したりするコーパス管理ツール

（冷凍食品）

名　　称	冷凍　餃子
原材料	野菜（キャベツ、白菜、たまねぎ、ねぎ、にら、ニンニク、しょうが）、いりごま、植物油（ごま油）、調味料（アミノ酸等）、食肉（豚肉）、香辛料、食塩、たれ（醤油、醸造酢、みりん、調味料（アミノ酸等）、ラー油（ごま油、香辛料、植物油）、皮(小麦粉、塩) (原材料の一部に小麦、大豆を含む)
内容量	３００ｇ（１５個入り）　賞味期限　2005.07.07
保存方法	－１８℃以下で保存のこと。
使用方法	冷凍保存の上、解凍せずに、焼餃子、蒸し餃子、揚げ餃子として調理して下さい。
冷凍前加熱有無	加熱してありません。　加熱調理の必要性　加熱して下さい。
製造者	○○○株式会社　ＡＢＣ 東京都○○区○○町○○　TEL 000-000-0000
使用上の注　　意	賞味期限内にお召しがり下さい。解凍してしまった場合はなるべくお召し上がり下さい。 使用原料に伴うアレルギー体質の方は、ご遠慮下さい。

4.1 辞書とコーパス

- 辞書
  - ある目的のために集められた単語の集合
  - 品詞や意味などの情報が付与されている
- コーパス
  - ある目的のために集められたテキストデータ
  - 新聞記事や小説、話し言葉の書き起こしなど

4.1 辞書とコーパス

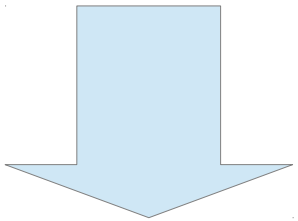
- 大規模な形態素解析用辞書
  - JUMAN辞書　品詞、活用形、読み、代表表記
  - IPADIC　品詞、活用形、かな表記、発音
- シソーラス
  - 分類語彙表　シソーラス
  - 日本語語彙大系　シソーラス、構文パターンなど
- 表層各情報が付与された辞書
  - IPAL辞書　意味情報、表層格情報
  - EDR辞書　品詞、かな表記、発音、活用情報、表層格情報、シソーラス、共起情報

ラベルが付与されていないデータ

- 大規模テキストデータ
  - 新聞記事
    - 毎日新聞社、朝日新聞社など、研究目的に販売
  - 著作権が切れた文学作品など
    - 青空文庫、プロジェクト杉田玄白

4.2 言語解析ツール

- 全ての分野のテキストについて言語情報が付与されたデータが収集できるわけではない



- 形態素情報や係り受け関係を付与する

4.2 言語解析ツール

- ChaSen
  - IPADIC品詞体系の品詞を付与する形態素解析器
- JUMAN
  - 益岡･田窪品詞体系の品詞を付与する形態素解析器
- MeCab
  - 条件付き確率場に基づく形態素解析器
- CaboCha
  - SVMに基づく係り受け解析器

4.3 コーパスの作成・管理ツール

- 単純なパターンマッチングによる検索では時間がかかる
- テキストから高速に文字列を検索するライブラリ
  - 接尾辞配列
    - テキスト中の接尾辞を辞書順に並べ、それに対するポインタを配列に格納したもの

接尾辞配列

例　Abracadabra

1	2	3	4	5	6	7	8	9	10	11
a	b	r	a	c	a	d	a	b	r	a

開始位置	接尾辞	LCP
11	a	0
8	abra	1
1	abracadabra	4
4	acadabra	1
6	adabra	1
9	bra	0
2	bracadabra	3
5	cadabra	0
7	dabra	0
10	ra	0
3	racadabra	2

4.3 コーパスの作成･管理ツール

- ChaKi
  - 形態素解析結果および係り受け解析結果に特化したコーパスを検索するツール
- ラベル付与には･･･
  - XMLで保持されることが多い

おわり

- 今回はこれで終わりです

おつかれさまでした