自然言語処理論I

13. テキスト処理(情報抽出)

この回のトピック
・情報抽出
・質問応答システム

情報抽出

- 情報抽出(information extraction, IE)
  - テキストから知りたい情報を直接取り出す

(参考)

- 情報検索(テキスト検索)
  - 知りたい情報が書いてある文書を取り出す

情報抽出

- 取り出したい情報の例
  - 製品名、メーカー、発売日、価格

NEXTSTEP 3.3J

キヤノンは、オブジェクト指向システム・ソフトウェア「NEXTSTEP 3.3J」(日本語版)の販売を開始する。主な特徴は以下の通り。...

価格はユーザ向けの製品版が98,000円、...

7月17日より出荷開始予定。発売目標は2000本/月。

フレーム型の情報抽出

- 取り出すべき情報をあらかじめフレームとして用意
- フレームの例

スロット	スロット値
メーカー	キヤノン
発売日	7月17日
製品名	NEXTSTEP 3.3J
価格	98,000円

情報抽出技術

- パターンマッチによる情報抽出
  - パターンの例

    <メーカー>は<製品名>の販売を開始する。

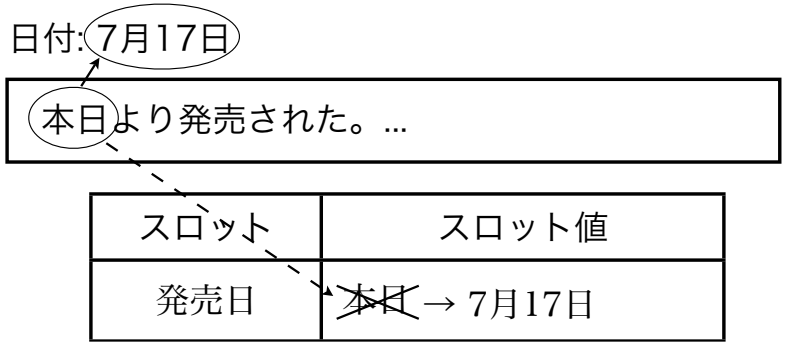
    価格は<詳細限定>が<価格>。

    <発売日>より出荷開始予定。

    - ✦ 対象となる文書の形式がある程度決まっていることが前提
  - パターンの作成
    - ✦ 人手で作成
    - ✦ コーパスから自動的に学習

情報抽出技術

- テキスト以外の情報を利用
  - 記事の発行日
    → 製品の発売日を特定する



質問応答システム

- 質問応答システム(QAシステム)とは?
  - 質問文の解答を文書集合の中から検索・抽出し、答えるシステム
  - 入力: 質問文
    - ✦ 東ティモールの首都はどこですか?
  - 出力: 解答
    - ✦ ディリ
  - 知識源: 文書集合
    - ✦ 新聞記事
    - ✦ WWW

質問応答システム

- 情報検索との違い
  - IR: 解答を含む文書を出力する
    - ✦ ユーザは文書の中から答えを探す必要がある
  - QA: 解答そのものを出力する

質問応答システム

- 情報抽出との違い
  - IE: 限られた情報のみを取り出すことを仮定
    - フレームの使用
  - QA: 取り出す情報に制限なし
    - ユーザは自由に質問してよい
    - 事実を解答とする質問が対象 (factoid型QA)
  - IE: 答えのある文書はあらかじめ用意されている
  - QA: 答えのある文書を探す必要がある

QAシステムの処理の流れ

- 質問文の解析
  - ユーザの質問の意図を理解する
- テキスト検索
  - 文書集合から解答を含む文書を取り出す
- 解答の抽出
  - 文書の中から解答を取り出す

質問文の解析

- 入力と出力
  - 入力: 質問文
  - 出力: 質問タイプ、検索質問
- 処理
  - 質問タイプ(質問意図)の判別 (何を尋ねているか?)
    - 「東ティモールの首都はどこですか」→ 場所を尋ねている
    - 「太平洋戦争は何年に始まりましたか」→ 時間
    - 「Play Stationを開発したのはどこですか」→ 組織
  - 質問文からのキーワード抽出
    - 「東ティモール」「首都」
    - 検索質問を作る

テキスト検索

- 入力と出力
  - 入力: 検索質問
  - 出力: 解答を含む(と思われる)文書
    - 一般に複数文書
- 処理
  - 情報検索技術の応用

解答の抽出

- 入力と出力
  - 入力: 文書、質問タイプ
  - 出力: 解答
- 処理
  - 情報抽出技術の応用
    - ✦ パターンマッチング
      - ○ パターンを事前に用意することは難しい
    - ✦ 質問文からのパターンの生成

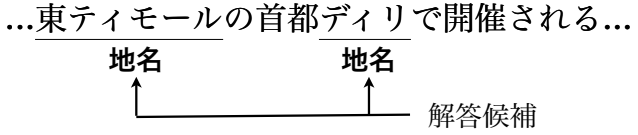
質問: 「東ティモールの首都はどこですか」

⇩

パターン: 東ティモールの首都は <解答> (だ|です)

解答の抽出

- 処理(続き)
  - 固有名詞抽出の利用
    - ✦ 解答の多くは固有名詞
    - ✦ 質問タイプで固有名詞の種類を限定できる
    - ✦ (例) 質問が場所を尋ねているとき
      → 地名を表わす固有名詞が解答の候補になる



固有名詞抽出

- Named Entity Extraction (NE)
- 固有名詞抽出とは?
  - テキスト中から固有名詞を検出する
  - 固有名詞の種類を当てる
    - ✦ 地名,人名,組織名,製品名 など
- どうやるか?
  - 固有名詞辞書は使わない(使えない)
  - 周辺の単語を手がかりとする
    - ✦ 人名の直後には「氏」がよく現われる
    - ✦ 地名の周辺には「開催」がよく現われる
  - 手がかりとなる単語をコーパスから学習する

まとめ

- (情報検索)
- 情報抽出
- 質問応答システム

- テキスト集合から情報を取り出す技術