Chapter 1 Section 3
EMPIRICAL EVALIATIONS

BOOK : Semi-Supervised Learning and Domain Adaptation in Natural Language Processing

Author : Anders Soegaard

茨城大学大学院理工学研究科情報工学専攻

菊池裕紀

この本では…

- 以下のアプリケーションに関するアルゴリズムの実証評価が示してある
  - document classification　(文書分類)
  - POS tagging　(品詞タグ付け)
  - dependency parsing　(係り受け解析)

document classification(1)

- 目的
  - 自動で事前定義されたクラスへ用例を分類する
  - 技術の適用例として、
    - authorship attribution　(著作権の帰属)
    - spam filtering　(スパムフィルタ)
    - topic classification　(トピック分類)
    - relevance filtering　(関連性フィルタリング)
    - adding MeSH terms to Medline abstracts　(データベース？)

document classification(2)

- 使用するデータセット
  - タグ付けされた用例

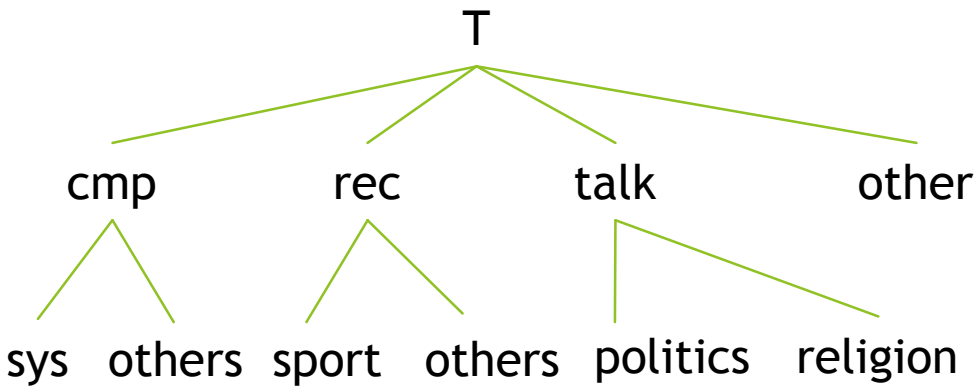
    $\langle y_1, x_1 \rangle, \ldots, \langle y_n, x_n \rangle$

    ※各データ点はバイナリまたは実数地
  - タスク

    用例を正しく分類できる関数fを求める

document classification(3)

- 実証評価
  - データ
    - Newsgroups dataset
    - HOCKEYとBASEBALLに関する投稿を見分けるときは、HOCKEY-BASEBALL problemと呼ぶこととする
    - Newsgroups datasetは階層構造になっている



document classification(4)

- データの抽出
  - 階層構造の最も高いレベルの領域適応の組み合わせを20個選出
    - COMPUTERS-RECREATIVE (cmp-rec)
  - 組み合わせにおいて、異なるデータセットを選出
    - IBM-BASEBALL
    - MAC-MOTORCYCLES ,etc.
  - 分類の際は
    - IBM-BASEBALLのデータセットでMAC-MOTORCYCLESを識別

POS tagging(1)

- 目的
  - ある系列 x の各要素に適切なラベル列 y を付与
    - This is a pen. → This(pronoun) is(verb) a(article) pen(noun).
  - 注意点
    - 単語は一つの語形で、異なる品詞を有する可能性がある

      Time files like an arrow.

      Time/N files/V like/ADV an/ART arrow/N.

      Time/N files/N like/V an/ART arrow/N.

POS tagging(2)

- 実証評価
  - OntoNotes 4.0コーパスを利用
  - 使用するドメイン
    - Yahoo! Answers
    - BBC Newsgroups
    - local business reviews
    - various weblogs
  - 新聞のデータからモデルを学習
  - the Enron email corpusをモデルのチューニング用に使用

dependency parsing

- 目的
  - 出力が文の内部構造を有する木となる構造推定タスク

    頂点：単語

    端　：文法機能

  - Chapter2にてarc-standard schemeを使用した例を示す