コーパスとテキストマイニング

11章　自然言語処理技術と機械学習手法を用いたテキストマイニング

第２回 異分野融合テキストマイニング研究会

中藤　哲也
九州大学 情報基盤研究開発センター

September 25, 2015

KYUSHU UNIVERSITY

Agenda

11.1 自然言語処理の諸技術
- 11.1.1 形態素・構文解析
- 11.1.2 固有表現の抽出
- 11.1.3 有用な情報を含む文の取り出し
- 11.1.4 教師あり機械学習手法

11.2 自然言語処理技術と機械学習手法を用いたテキストマイニングの事例
- 11.2.1 機械学習を用いた論文要約からの重要情報抽出
- 11.2.2 大規模データからの固有表現と数値表現の抽出と可視化
- 11.2.3 対訳コーパスからの日英対訳表現の抽出

11.3 むすび

Agenda

11.1 自然言語処理の諸技術
- 11.1.1 形態素・構文解析
- 11.1.2 固有表現の抽出
- 11.1.3 有用な情報を含む文の取り出し
- 11.1.4 教師あり機械学習手法

11.2 自然言語処理技術と機械学習手法を用いたテキストマイニングの事例
- 11.2.1 機械学習を用いた論文要約からの重要情報抽出
- 11.2.2 大規模データからの固有表現と数値表現の抽出と可視化
- 11.2.3 対訳コーパスからの日英対訳表現の抽出

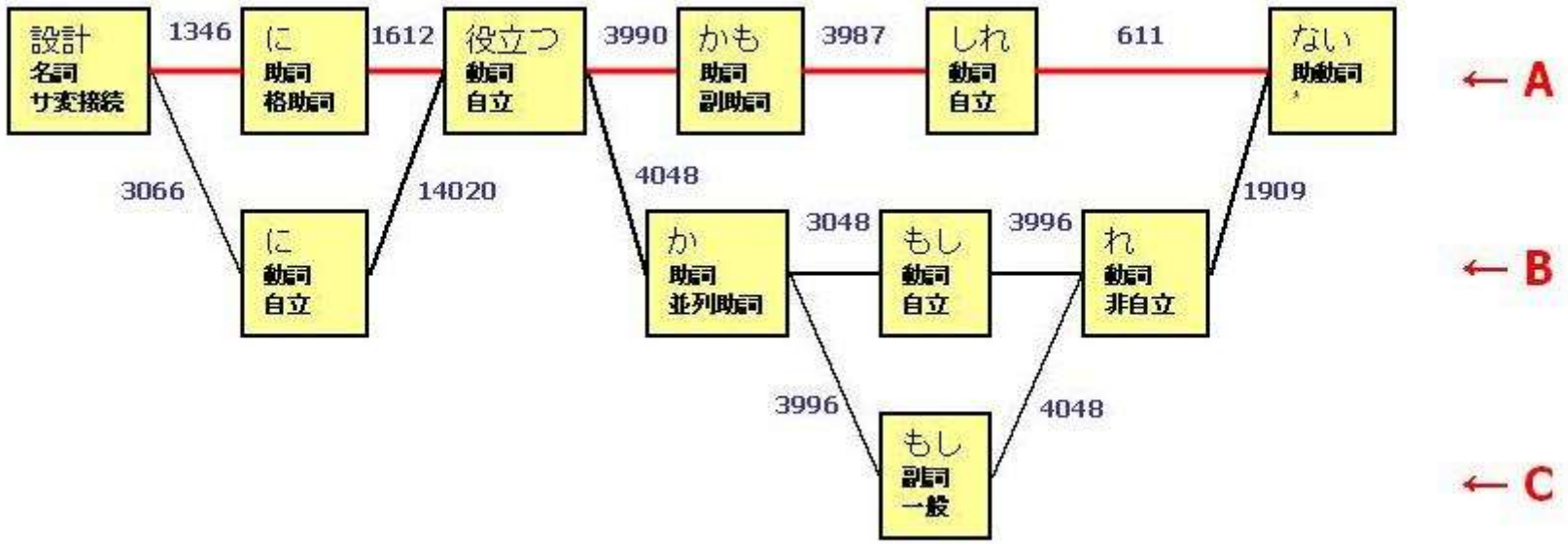
11.3 むすび

形態素解析

計算機で、文を単語に分割し各単語の品詞を決める

単語間連接コスト、単語コストなどを使う

例：「設計に役立つかもしれない」



ラティス構造を用いた形態素解析 ： 形態素解析エンジン Tofu の場合

形態素解析

フリー（無料）の形態素解析ツール

- JUMAN：
- KAKASI：分かち書き
- ChaSen(茶筅)：隠れマルコフモデル
- MeCab(和布蕪) :条件付き確率場

MeCabによる形態素解析の例

```
$ mecab
形態素解析にはMeCabが使われることが多いです。
形態素	名詞,一般,*,*,*,*,形態素,ケイタイソ,ケイタイソ,,
解析	名詞,サ変接続,*,*,*,*,解析,カイセキ,カイセキ,,
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ,,
は	助詞,係助詞,*,*,*,*,は,ハ,ワ,,
MeCab	名詞,固有名詞,組織,*,*,*,*
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ,,
使わ	動詞,自立,*,*,五段・ワ行促音便,未然形,使う,ツカワ,ツカワ,つかわ/使わ,
れる	動詞,接尾,*,*,一段,基本形,れる,レル,レル,,
こと	名詞,非自立,一般,*,*,*,こと,コト,コト,,
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ,,
多い	形容詞,自立,*,*,形容詞・アウオ段,基本形,多い,オオイ,オーイ,おおい/多い,
です	助動詞,*,*,*,特殊・デス,基本形,です,デス,デス,,
。	記号,句点,*,*,*,*,。,。,。,,,
EOS
```

Agenda

11.1 自然言語処理の諸技術
- 11.1.1 形態素・構文解析
- 11.1.2 固有表現の抽出
- 11.1.3 有用な情報を含む文の取り出し
- 11.1.4 教師あり機械学習手法

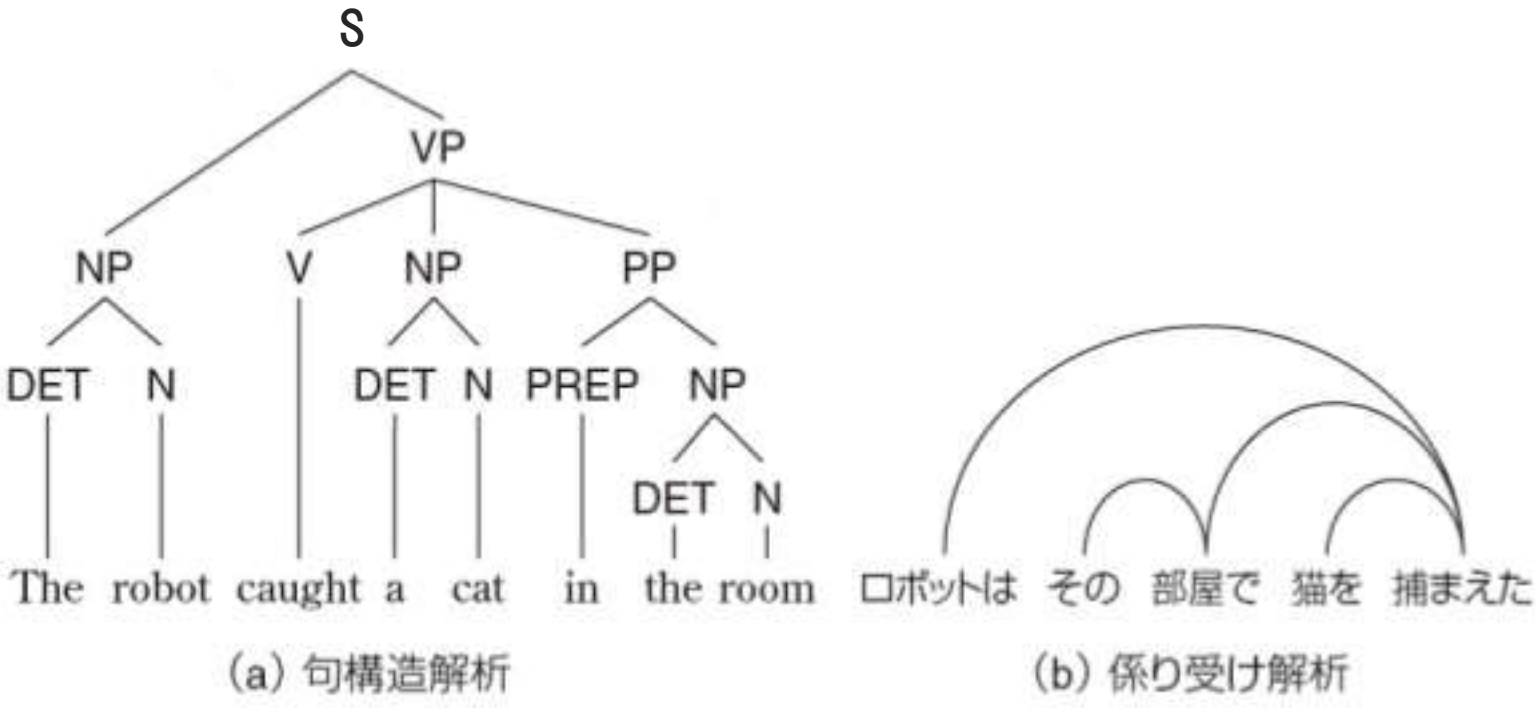
11.2 自然言語処理技術と機械学習手法を用いたテキストマイニングの事例
- 11.2.1 機械学習を用いた論文要約からの重要情報抽出
- 11.2.2 大規模データからの固有表現と数値表現の抽出と可視化
- 11.2.3 対訳コーパスからの日英対訳表現の抽出

11.3 むすび

構文解析（日本語係り受け解析）

構文解析の２つのタイプ

- 句構造解析（句構造文法）← 前から後ろに…英語など
- 係り受け解析（依存文法）← 語順が自由…日本語向き



(a) 句構造解析　　(b) 係り受け解析

構文解析（日本語係り受け解析）

係り受け解析ツール

- KNP：格フレームに基づく確率的モデル
- CaboCha(南瓜)：Support Vector Machines (SVMs) に基づく
- Yahoo!デベロッパーネットワーク：APIで呼び出し 50000リクエスト/24時間

CaboChaによる解析の例

```
$ cabocha
ロボットはその部屋で猫を捕まえた
ロボットは-------D
        その-D   |
        部屋で---D
            猫を-D
          捕まえた
EOS
```

Agenda

11.1 自然言語処理の諸技術
- 11.1.1 形態素・構文解析
- 11.1.2 固有表現の抽出
- 11.1.3 有用な情報を含む文の取り出し
- 11.1.4 教師あり機械学習手法

11.2 自然言語処理技術と機械学習手法を用いたテキストマイニングの事例
- 11.2.1 機械学習を用いた論文要約からの重要情報抽出
- 11.2.2 大規模データからの固有表現と数値表現の抽出と可視化
- 11.2.3 対訳コーパスからの日英対訳表現の抽出

11.3 むすび

固有表現抽出

固有表現（Named Entity）とは？

- 固有名詞(人名、地名), 人工名(組織名), 数値表現(日時)など
- 辞書に登録されてない場合が多い　→　形態素解析などで困る
- 全てを辞書に登録するのは困難
- テキストマイニングとしては重要な情報である可能性大

固有表現の抽出（方法の一例）

- 固有表現に関する情報（実例）を集める
  - 小渕首相が沖縄で会談した

- 教師あり機械学習(SVMなど)で分類器を作成
  - ○○首相→○○は人名
  - △△で会談した→△△は地名

Agenda

11.1 自然言語処理の諸技術
- 11.1.1 形態素・構文解析
- 11.1.2 固有表現の抽出
- 11.1.3 有用な情報を含む文の取り出し
- 11.1.4 教師あり機械学習手法

11.2 自然言語処理技術と機械学習手法を用いたテキストマイニングの事例
- 11.2.1 機械学習を用いた論文要約からの重要情報抽出
- 11.2.2 大規模データからの固有表現と数値表現の抽出と可視化
- 11.2.3 対訳コーパスからの日英対訳表現の抽出

11.3 むすび

有用な情報を含む文の取り出し

教師あり学習で抽出

- 例：ユーザサポートのログ
  - トラブルか否かをログにタグ付け　←　学習用データ
  - SVMで学習させて、分類器を生成
  - 新しいログ
    - トラブルに関するものかを自動で判断
    - トラブル傾向の分析
    - サポート要員への関連情報の表示

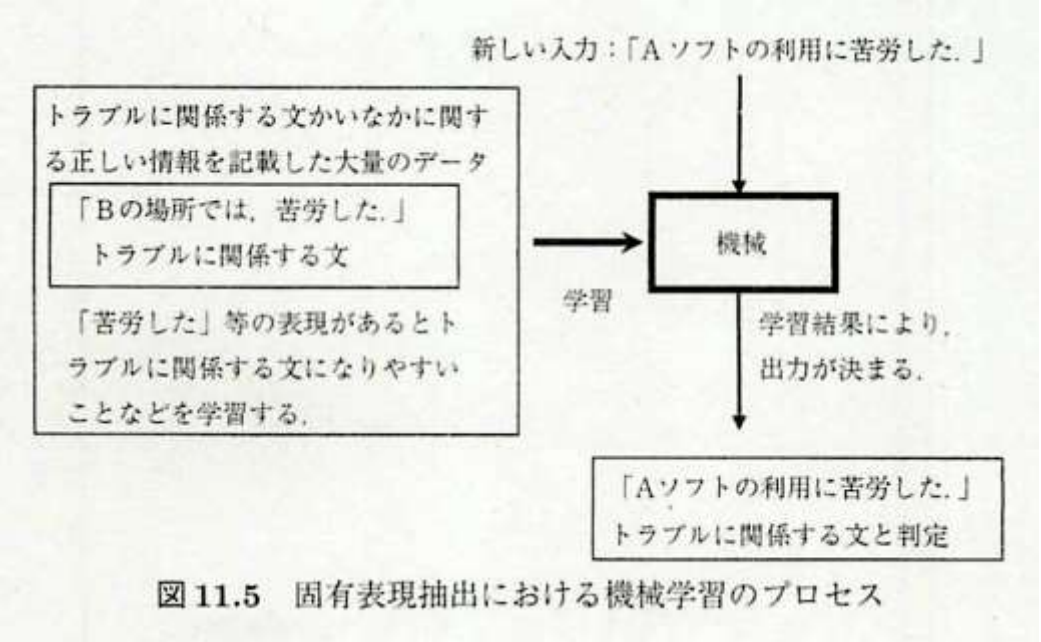


図 11.5　固有表現抽出における機械学習のプロセス

Agenda

11.1 自然言語処理の諸技術
- 11.1.1 形態素・構文解析
- 11.1.2 固有表現の抽出
- 11.1.3 有用な情報を含む文の取り出し
- 11.1.4 教師あり機械学習手法

11.2 自然言語処理技術と機械学習手法を用いたテキストマイニングの事例
- 11.2.1 機械学習を用いた論文要約からの重要情報抽出
- 11.2.2 大規模データからの固有表現と数値表現の抽出と可視化
- 11.2.3 対訳コーパスからの日英対訳表現の抽出

11.3 むすび

教師あり機械学習手法

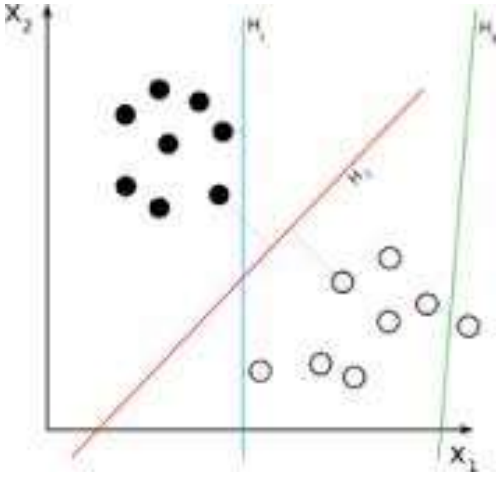
機械学習：サンプルからルールを見つけ出す

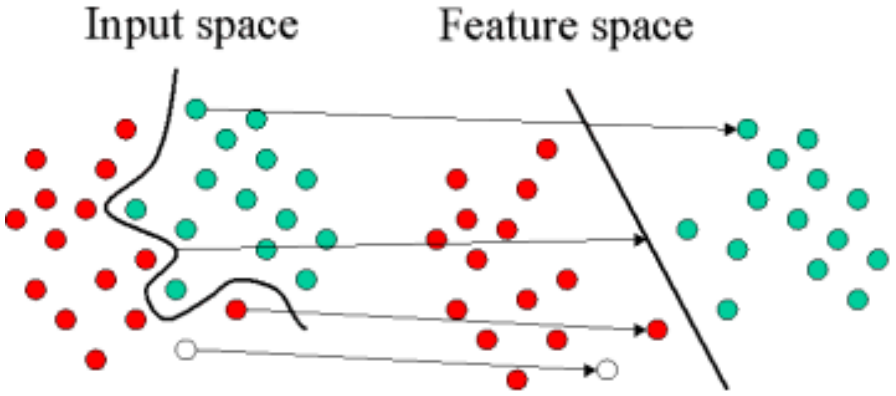
- 教師あり学習
  - 大量のデータと、その一部についての正解例
  - SVMなどで分類器を作る
  - 最初から解のパターンが分かっている
  - 実は、余りマイニングっぽくない…
- 教師無し学習
  - 大量のデータのみ
  - 主成分分析などを用いて、データの持つ特徴（or 構造）を抽出
  - 何を特徴量とするか…が難しい
  - Googleのディープラーニングは、教師無しで猫を認識した

教師あり機械学習手法

SVM(Support Vector Machine)

- 教師あり学習によるパターン認識モデル
- 正解付きデータ（学習データ）で分類器を作成　（通常数千～数万）
- 未知データを　正例、負例の２種に分離する
- 非線形分離も可能（カーネル関数）





Agenda

11.1 自然言語処理の諸技術
- 11.1.1 形態素・構文解析
- 11.1.2 固有表現の抽出
- 11.1.3 有用な情報を含む文の取り出し
- 11.1.4 教師あり機械学習手法

11.2 自然言語処理技術と機械学習手法を用いたテキストマイニングの事例
- 11.2.1 機械学習を用いた論文要約からの重要情報抽出
- 11.2.2 大規模データからの固有表現と数値表現の抽出と可視化
- 11.2.3 対訳コーパスからの日英対訳表現の抽出

11.3 むすび

機械学習を用いた論文要約からの重要情報抽出

論文のアブストラクトから重要な情報を取り出す（菊井2006）

- 重要な情報として
  - 精度表現：精度を表す表現（97%）
  - 主要な分野：自然言語処理研究の分野名（機械翻訳、構文解析、抽出）
  - 言語名：日本語、英語など
  - 組織・人名：大阪大学、坂口志文、木俣肇
- 使用ツールはYamCha：複数のSVMを組み合わせて、複数クラス分類
- データ
  - 主要な分野と言語名　：　約500個の論文アブストラクトを学習データ
  - 精度表現と組織・人名　：　約2000個の論文アブストラクトを学習データ
  - 素性　：　前方３単語、後方３単語
    - 例「提案　手法　の　精度　は　97%　で　あった」
- 結果
  - 重要な表現の抽出性能８割

Agenda

11.1 自然言語処理の諸技術
- 11.1.1 形態素・構文解析
- 11.1.2 固有表現の抽出
- 11.1.3 有用な情報を含む文の取り出し
- 11.1.4 教師あり機械学習手法

11.2 自然言語処理技術と機械学習手法を用いたテキストマイニングの事例
- 11.2.1 機械学習を用いた論文要約からの重要情報抽出
- 11.2.2 大規模データからの固有表現と数値表現の抽出と可視化
- 11.2.3 対訳コーパスからの日英対訳表現の抽出

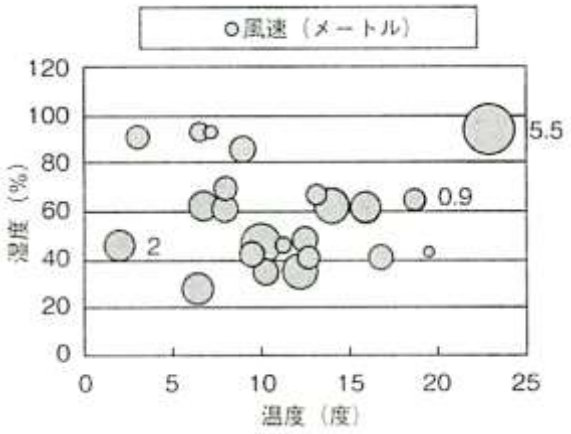
11.3 むすび

大規模データからの固有表現と数値表現の抽出と可視化

村田真樹ほか,「大規模記事群からの数値固有表現情報に関わるテキストマイニング可視化」第6回情報科学技術フォーラム, pp.157-160, 2007.

- 2年間の新聞から取り出したデータ
  - 固有表現：組織名, 人名, 地名, 日付表現, 時間表現, 金額表現, 割合表現, 固有物名
  - 項目表現：形態素解析を利用して名詞連続…？

項目表現：マラソン



項目表現：弾道ミサイル

固有物名（ミサイル名）	国名
アロー	イスラエル
シャヒーン	パキスタン
ノドン１号	北朝鮮
テポドン	北朝鮮
テポドン２	北朝鮮

Agenda

11.1 自然言語処理の諸技術
- 11.1.1 形態素・構文解析
- 11.1.2 固有表現の抽出
- 11.1.3 有用な情報を含む文の取り出し
- 11.1.4 教師あり機械学習手法

11.2 自然言語処理技術と機械学習手法を用いたテキストマイニングの事例
- 11.2.1 機械学習を用いた論文要約からの重要情報抽出
- 11.2.2 大規模データからの固有表現と数値表現の抽出と可視化
- 11.2.3 対訳コーパスからの日英対訳表現の抽出

11.3 むすび

対訳コーパスからの日英対訳表現の抽出

Qing Ma et al. "High-Precision Japanese-English Parallel Translation Expressions Using Lexical Information and Rules, Paclic25, 2011.

- 使用データ：対訳コーパス
  - NICTコーパス：日本の新聞記事とその英訳　約40,000対
  - JENNADコーパス：読売新聞とDaily Yomiuriの記事を自動対応付け約180,000対
  - ロイターコーパス：ロイター通信の日米ニュース記事を自動対応付け約70,000対
- 手法
  - 単語n-gramをそれぞれから抽出　nは3以下
  - 類似度計算　単語n-gram x_j と x_e

$$\mathrm{sim}(x_j, x_e) = \frac{2 f(x_j, x_e)}{f(x_j) + f(x_e)}.$$

  - しかしそのままでは精度が低い→辞書を使用

J1:暗殺されたラビン首相
E1: prime minister rabin
　　who was assassinated
J2:不均衡を縮小するため
E2: to reduce imbalance
J3:拉致疑惑問題
E3: the kidnapping issue
J4:揶揄された
E4: have been ridiculed

Agenda

11.1 自然言語処理の諸技術
- 11.1.1 形態素・構文解析
- 11.1.2 固有表現の抽出
- 11.1.3 有用な情報を含む文の取り出し
- 11.1.4 教師あり機械学習手法

11.2 自然言語処理技術と機械学習手法を用いたテキストマイニングの事例
- 11.2.1 機械学習を用いた論文要約からの重要情報抽出
- 11.2.2 大規模データからの固有表現と数値表現の抽出と可視化
- 11.2.3 対訳コーパスからの日英対訳表現の抽出

11.3 むすび

むすび

- テキストマイニングの基礎となる諸技術
  - 形態素解析
  - 構文解析
  - 固有表現抽出
  - 有用な情報を含む文の取り出し
  - 教師あり機械学習手法

- 研究事例の紹介
  - 論文要約からの重要情報抽出
  - 固有表現と数値表現の抽出と可視化
  - 対訳コーパスからの日英対訳表現の抽出