自然言語処理入門
「言語か数学か：計算機のパワーによる統計的言語処理」

東京大学　情報基盤センター
（情報理工学系研究科、情報学府　兼担）
中川裕志
nakagawa@r.dl.itc.u-tokyo.ac.jp
http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/

大規模コーパスでの言語処理

- 大規模コーパスから何らかの有用な情報を抽出したい。
- 今までは人手でコーパスを処理して統計データを抽出していた。しかし、もはやコーパスが大きすぎて、人手では無理。計算機で処理する時代
- 100MBからGBくらいの大きなコーパスになると、形態素解析や構文解析のような重い処理を闇雲に行うことは、時間がかかり過ぎるし無駄が多い。
- もう少し軽い統計処理が検討されている。
- 単語など有用な言語単位を対象にする統計的処理が多く研究されている

統計的言語処理にはどんなものがあるのか

◆テキストからの情報抽出

- ●Nグラム統計
- ●単語、用語の抽出、頻度分布、インデクシング
- ●複数コーパスの対応つけ（Alignment）
- ●2言語コーパスの対応つけ、自動対訳抽出

◆テキストの変換処理

- ●文書分類
- ●自動段落分割
- ●自動要約、トピック抽出
- ●機械翻訳
- ●マルチモーダルのコーパスの対応つけ（exビデオインデクシング）

なぜ言語の統計？

- 統計量を計算する理由
  - １　我々の使っている言語について知る
  - ２　言語モデル（文法、言語運用（語用）、文書構成規則、など）を作る
  - ３　辞書や電子化辞書を作る
  - ４　言語処理に必要な計算機資源を見積もる

延べ数　と　異なり数

- 「便り　/の　/　ない　/　の　/　は　/　よい　/　便り」
- 形態素＝固有の意味を持ち、かつそれ以上分解できない言語の単位
  単語ってなに？
- 延べ形態素数＝7、　異なり形態素数＝5
- 頻度は → 便り＝2、の＝2、ない＝1、は＝1、よい＝1
- 単語とは？

1 「ので」は1単語か？

2 「日々」の「々」は？

3 「日銀」は1単語？

4 「日本銀行」は1単語か2単語か（あきらかに2形態素らしくはある）

5 実際、単語とは何か、というのは国語学者を悩ませてきた問題であった。

6 実用性の観点からすれば、辞書にどの単語を登録するかが問題なのだから、応用目的によって単語を決めればよい。

統計の対象となる言語単位

- 音素： 訓令式でローマ字表記したときのアルファベット1文字に対応する音
- 文字： 文字セットによる。しかし、アルファベットの大文字小文字、日本語漢字の新字 例えば「沢」と「澤」は同じ文字か違う文字か？
- 単語： 西欧の言語のように正書法（Orthography)によって単語間に空白があればよいけれど。 「虎の子」は1単語？→ Collocation の問題
- 複合語：

  1 複合名詞：「日本学術会議第5部会員選挙日程検討結果報告書別添資料補遺3ページ2行目」？？？？

  2 複合動詞：「追いかける」＝「追う」＋「かける」

  「老けこむ」＝「老ける」＋「こむ」

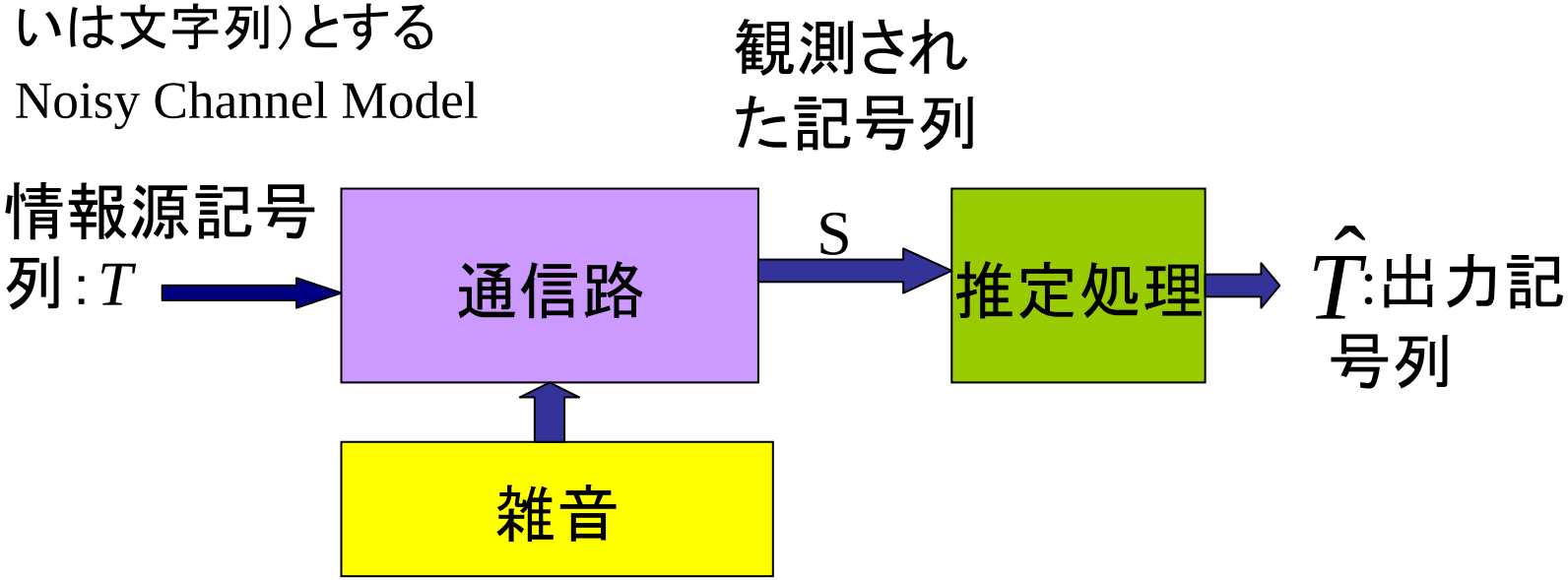
  「痛がる」？？
- 句： 名詞句、動詞句： 言語情報処理では文よりも句が目下のターゲット
- 文： 書き言葉では句点があるが、話し言葉では？
- 談話： 談話構造の認識、談話の範囲などが問題

語彙範疇　ｖｓ　機能範疇　と　品詞

- 語彙範疇：　辞書に記載されている内容的意味のある単語。具体的には動詞，名詞。　内容語ともいう。
- 機能範疇：　固有の意味を持たないが、内容語の修飾や内容語同士の関係を記述するための言語要素。冠詞、助詞、屈折辞など。機能語ともいう。
- 情報検索のような文の意味内容を直接捉えようとする試みにおいては、内容語の抽出や統計的性質が重要で研究も進んでいる。一方で、言語学や機械翻訳などの言語を扱う正統派？？の領域では、言語現象としての機能語から、文法や意味に迫る方法をとる。現状では両者の間では乖離があることを認めざるをえない。
- 品詞：　内容語（もちろん機能語も）が文法的にどういう性質かを記述するのが品詞。統計的性質を調べる場合に、ここの単語（特に内容語の場合）は出現頻度も低く、むしろ同じ品詞のものを同一視して統計をとる方法が有力な場合あり。日本語の品詞体系は以下を参照されたし。
- 　　１　基礎日本語文法：　益岡隆志、田窪行則著、くろしお出版、１９９２（いわゆる益岡・田窪文法）など。

テキストの変換処理のモデル化

- テキストを記号列（単語列あるいは文字列）とする
- Noisy Channel Model



観測されたSから情報源記号列Tを推定し $\hat{T}$ を計算する

推定方法

- 通信路を条件付確率でモデル化：P(T|S)
- Sを知った上でのTの確率すなわち事後確率P（T|S)を最大化する $\hat{T}$ として求める。

$$\hat{T} = \arg\max_{T} P(T \mid S)$$ ここでベイズの定理により

$$= \arg\max_{T} P(S \mid T)P(T)$$

- P(T)は情報源記号列の既知の統計的性質が利用できる
- P(S|T)は情報源記号列Tがnoisy channelの雑音によってSに変化する確率。多数の（T,S)対のデータ（コーパス）のより計算する

Noisy Channel Modelの適用例

- 機械翻訳
  - P(S|T)　元言語のテキストTが翻訳先言語のテキストSに翻訳される確率
  - P(T)　元言語のテキストTの自然さ。例えば、N単語列のコーパスにおける　単語3-gram確率
  - $\hat{T}$　は機械翻訳の出力
- 文書要約
  - P(S|T)　要約文Tから長い元テキストSが、作られる確率
    - 元テキストとその要約文の集合が教師データとして与えられていれば計算できる。
    - とはいえ、どのような方法で要約されているかを統計的に同定するのは相当難しい。
  - P(T)　要約テキストの文としての自然さ
  - $\hat{T}$　が要約システムの出力する要約文

- 文書分類
  - P(T|S)においてSが与えられた文書、Tがカテゴリ

    推定されたカテゴリ $:\hat{T} = \arg\max_{T} P(S \mid T)P(T)$

  - P(T)はカテゴリTの文書の出現確率
  - P(S|T)はカテゴリTの文書において出現する文書S
  - Sのモデル化にはいろいろな方法があるが、簡単なのは、出現する単語$w_1,\ldots w_n$
  - P(S|T)=P($w_1,\ldots w_n$|T)だが、このままでは計算しにくいので$w_1,\ldots w_n$が独立とすると

    $$P(w_1,..., w_n \mid T) = \prod_{i=1}^{n} P(w_i \mid T)$$

  - これを naïve Bayse 分類とよぶ。