中国語形態素解析に対するSVMとコスト最小法の比較実験

吉田辰巳(※)　大竹清敬(※※)　山本和英(※※)

(※)豊橋技術科学大学 知識情報工学系

(※※)ATR 音声言語コミュニケーション研究所

研究の背景

- 言語解析器の重要性
  - 形態素解析器
  - 構文解析器
- 日本語や英語の場合は，多数開発
- 中国語の解析器は不充分

中国語解析の困難性（1）

・形態素分割が困難

英語:

Time	flies	like	an	arrow.

日本語:

象は鼻が長い。

中国語:

他是日本来的。

中国語解析の困難性(2)

・品詞付与が困難

例:「担心」

(1)心配 (名詞)
(2)心配する (動詞)
(3)心配な (形容詞)

研究の目的

中国語の形態素解析器を
なるべく手軽に構築したい

現在入手可能な資源やツールを使用

- どの程度の精度で解析できるか
- ツールごとの性質の違いは何か

コーパスと解析ツール

学習・解析コーパス

- PennChineseTreebank（CTB)

解析ツール

- YamCha（サポートベクトルマシン）
- MOZ（コスト最小法）

CTB

(pennsylvania Chinese TreeBank)

- 中国語の構文木コーパス
- 新華社通信の 325 記事が対象
- 単語分割，品詞付与，構造付与

YamCha

(Yet Another Multipurpose CHunk Annotator)

- SVMに基づく
- Chunkingによる言語処理
  英語の基本句同定や日本語の
  係り受け解析で，高い精度を示す
- 汎用的な解析器として利用可能
  対象言語や解析レベルに依らない

（※）奈良先端大の工藤らが開発

SVM

(Support Vector Machine)

- 特徴ベクトルの２値クラス識別器
- 文字認識，画像認識などに応用

YamChaは，複数のSVMの多数決
(pairwise classification) によって
２つ以上のクラスの識別を行う

Chunkingによる形態素解析

入力要素列（入力文）

YamChaによる解析

学習データ

……

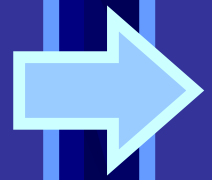
要素　区切り　タグ

要素　区切り　タグ

要素　区切り　タグ

要素　区切り　タグ

……



解析データ

……

要素　区切り　タグ

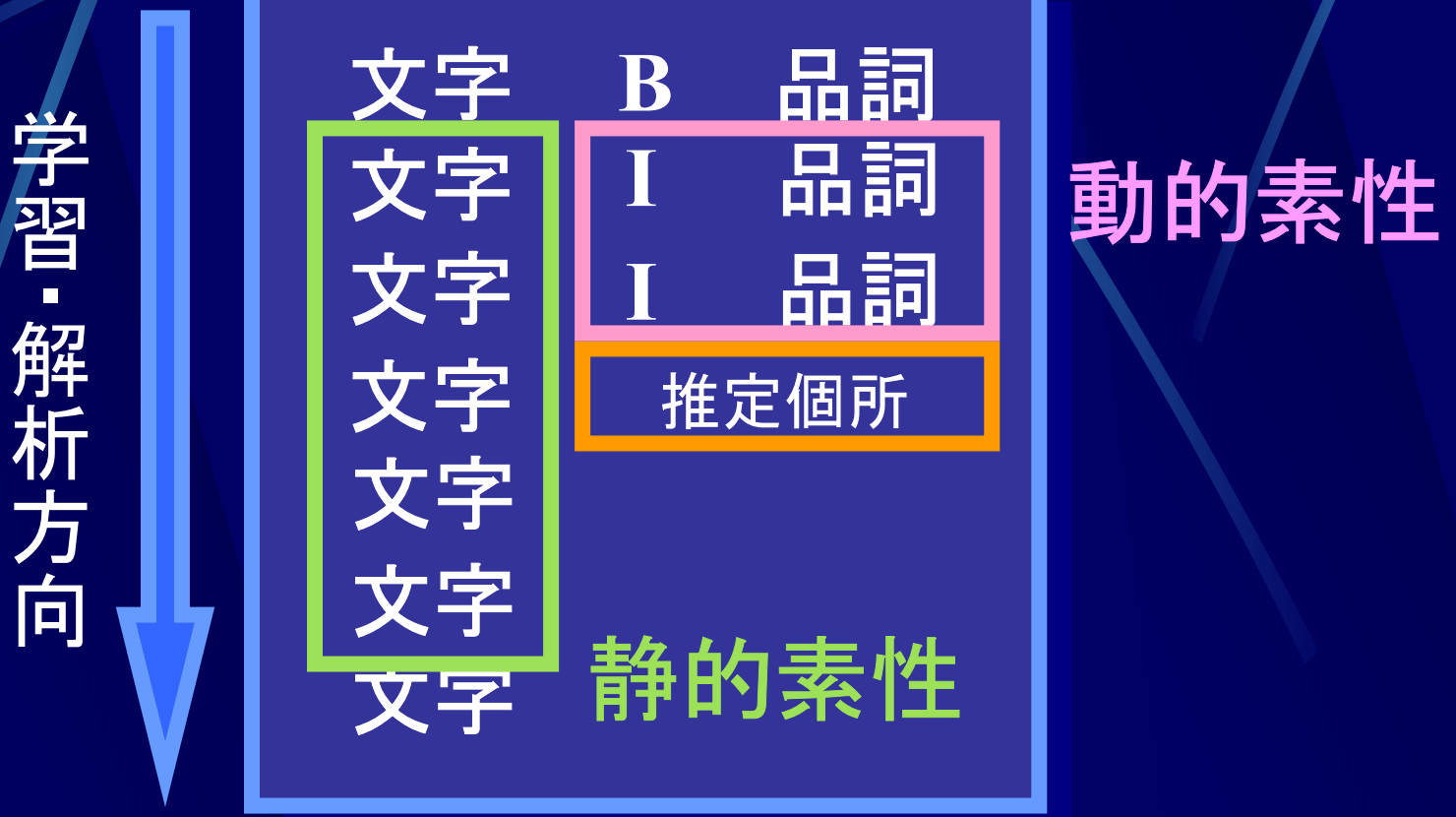
要素　区切り　タグ

要素　区切り　タグ

要素　区切り　タグ

……

学習に用いた素性



MOZ

- 言語に依らない 汎用的な形態素解析
- コスト最小法によって解析

形態素辞書 : 形態素の出現頻度
接続表 : 品詞 bi-gram

（※）奈良先端大の山下らが開発

形態素解析の比較実験

- closed test

  CTB全体（４１８１文）を学習データとする.
  無作為に抽出した1割（４１８文）を解析.

- open test

  CTB全体を母集団とする
  １０分割交差検定（cross validation）

closed testの正解率

	分割のみ		分割と品詞付与	
	再現率	適合率	再現率	適合率
YamCha	**99.91%**	**99.93%**	**99.58%**	**99.60%**
MOZ	**97.78%**	**98.82%**	**93.74%**	**94.73%**

YamChaの方が再現率･適合率が高い

open testの正解率

	分割のみ		分割と品詞付与	
	再現率	適合率	再現率	適合率
YamCha	**93.04%**	**93.71%**	**87.58%**	**88.20%**
MOZ	**92.19%**	**85.89%**	**86.32%**	**80.42%**

YamChaの方が再現率･適合率が高い

品詞付与の誤りの傾向

解析器による傾向の差異はない

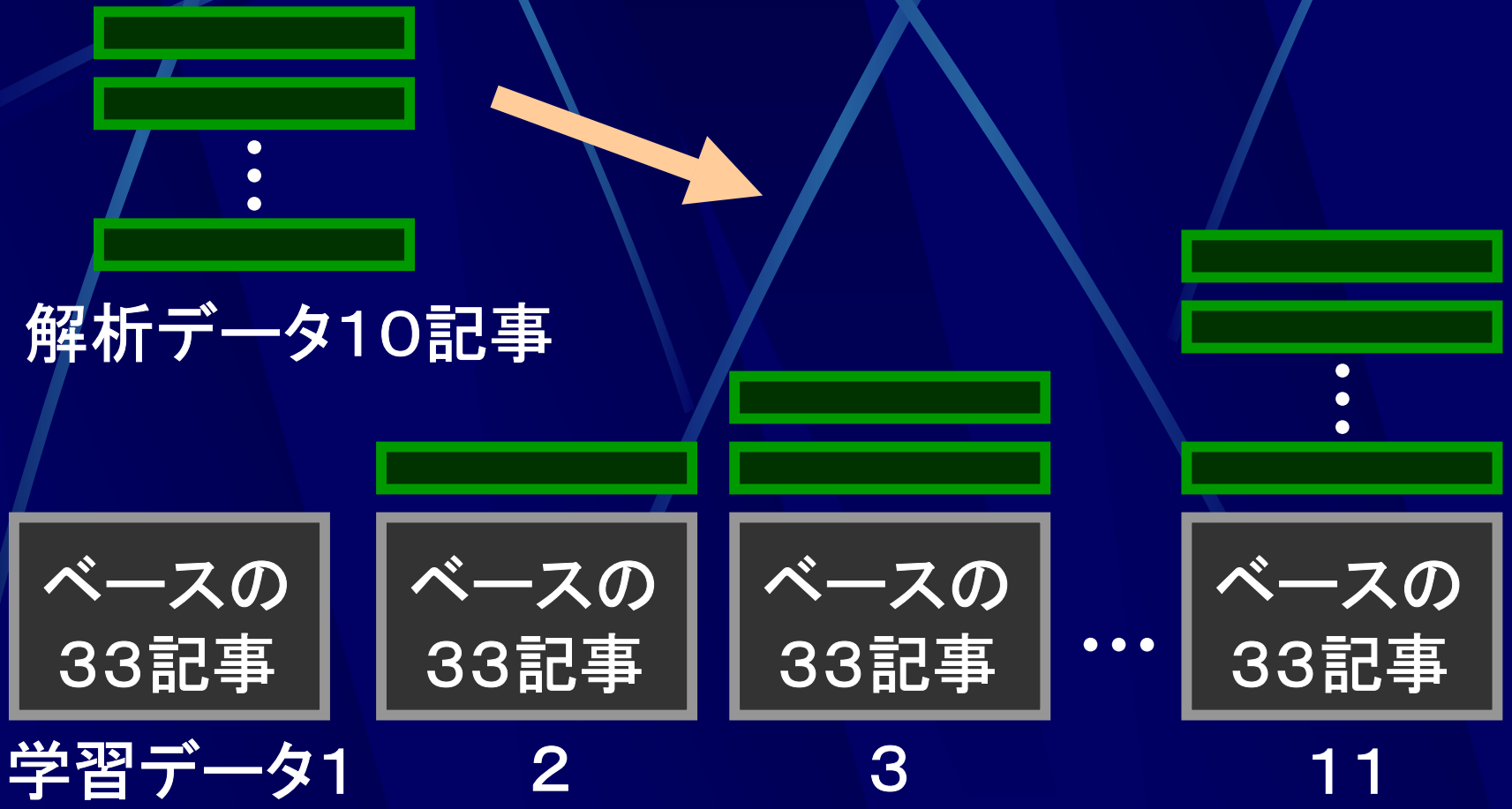
- 動詞-名詞間の誤りが多い
  - 複数品詞を持つ語が多い
- 名詞-固有名詞間の誤りが多い
  - 文法上，機能の違いはなく，
    正解データ中にも不統一がある

計算時間

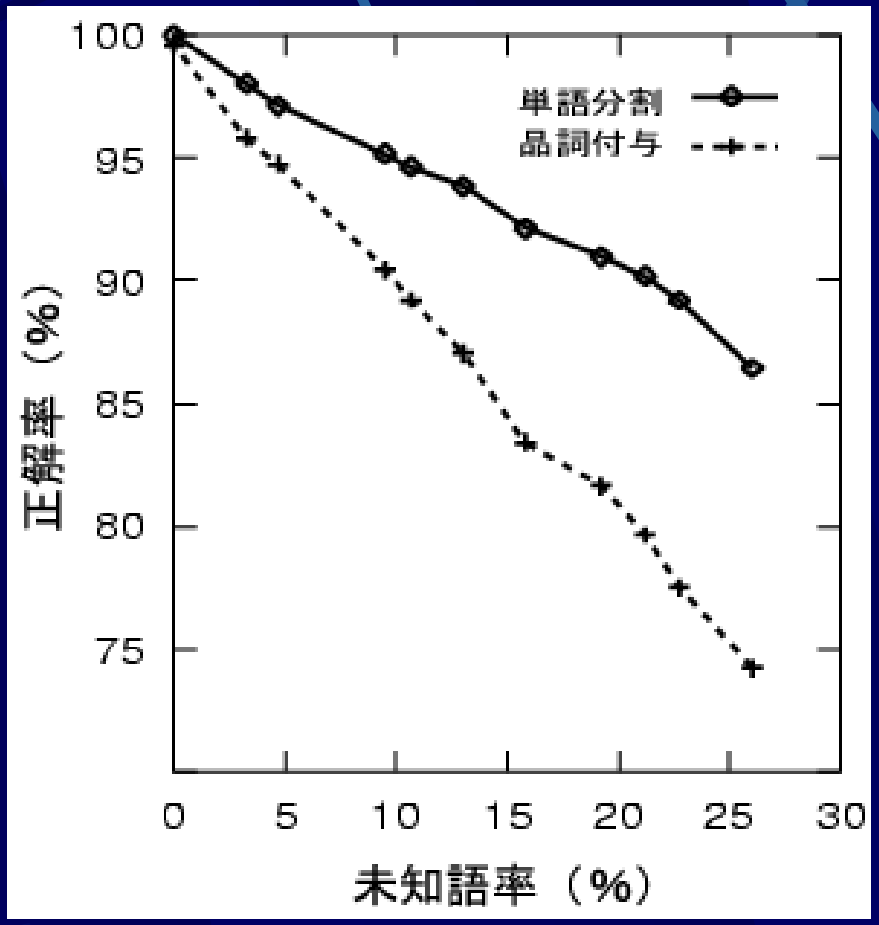
計算機:
PentiumIII 600MHz メモリ:256MB

	学習時間	解析時間
YamCha	約6時間	約35分
MOZ	約3秒	約1秒

未知語と解析精度


解析データ10記事
ベースの33記事
ベースの33記事
ベースの33記事
…
ベースの33記事
学習データ1
2
3
11

未知語による影響



YamCha

未知語の解析正解率

コーパス全体を占める未知語の割合が変化しても，未知語の正解率はほぼ一定

YamCha ：約**40**%の分割と品詞付与
MOZ ：約**10**%の分割のみ

⇒ YamCha の方が頑健である

再現率・適合率低下の原因

報告されている日本語の解析結果よりも，再現率・適合率が低い

その原因は？

- 学習コーパスの量？
- 日本語と中国語の性質の違い？

コーパスの大きさ

CTB：約10万語から成る構文木コーパス

PKU：人民日報タグ付きコーパス1ヶ月分
約112万語から成る

PKUの方が解析が困難

- 未知語率が高い
- 1文ごとの平均形態素数が多い
- タグセット数（品詞の種類）が多い

PKUでの実験結果

	PKU約10万語		PKU約100万語	
	再現率	適合率	再現率	適合率
YamCha	**80.19%**	**80.99%**	**91.72%**	**91.72%**
MOZ	**84.57%**	**75.67%**	**89.87%**	**87.75%**

言語依存性(1)

京都大学テキストコーパスから
CTBと同じ10万語に相当する文を使用

	分割のみ		分割と品詞付与	
	再現率	適合率	再現率	適合率
YamCha	**92.02%**	**93.23%**	**88.17%**	**89.33%**

日本語でもほぼ同じ結果

言語依存性(2)

中国語と日本語で，解析結果の
再現率･適合率に差は見られない

ただし，CTBよりも京大コーパスの方が
未知語率やタグセット数が大きいため，
解析が困難であると考えられる.

⇒中国語解析の困難性

考察(1)

- SVMは既存のコスト最小法と比べて
  再現率・適合率が高いが，計算量が多く，
  より多くの計算時間を必要とする
- 中国語の形態素解析は困難であるが，
  大量のタグ付きコーパスを用いることで
  再現率や適合率をより高めることができる
  （特にYamChaの方が大きい）

考察(2)

- コスト最小法では未知語を正しく解析できないが，SVMでは約半数を解析
- SVMは，学習コーパスを拡張すると，初めから学習し直す必要があるが，コスト最小法では，コーパスや辞書によって拡張を容易に行える

まとめ

既存のツールによる中国語形態素解析の性能比較評価を行った

YamChaの利点：

- 再現率や適合率が高い
- 未知語に対する頑健性が高い
- 学習量増加で適合率・再現率増加

MOZの利点：

- 学習・解析時間が短い
- 学習コーパスと別に辞書を扱える