混成型別サンプリングを用いた名詞句分割

2011年3月9日

京都大学大学院情報学研究科
村脇 有吾　　黒橋 禎夫

言語解析における語彙知識

- 機能語、基本的な内容語は人手で整備できる
  - を, する, 走る, 駅, …
- 一般の内容語は人手で整備しきれない
  - ググる, ようつべ, 把える, とう痛, …

足りない内容語は自動整備

- 生テキストからの語彙獲得 [Murawaki+ 2008]

何となくググってみたらどう。
ググらずに答え・・・
・・・だけで、ググるためのリ・・・

ググ-る:動詞-ラ行

生テキストからの内容語獲得

手がかり: 文法的マーカ (後続付属要素) の規則性

動詞-ラ行

走	–らない	ググ	–らない
走	–るとき	ググ	–るとき
走	–った	ググ	–った
走	–って	ググ	–って

名詞

駅	–が	ようつべ	–が
駅	–を	ようつべ	–を
駅	–なら	ようつべ	–なら
駅	–って	ようつべ	–って

問題: 名詞句内に文法的マーカがない

- フェルミ -ϕ エネルギー
- 抗 -ϕ 甲状腺 -ϕ 剤

- ?? アンド -ϕ レイ
- ?? 常 -ϕ 山城

名詞句の分割誤り

1. 形態素解析の分割誤り
   - アンド + レイ
   - 常 + 山城
   - ちん + すこう (好こう/空こう/透こう)
2. 語彙獲得で名詞句を単形態素として獲得
   - フェルミエネルギー (正解: フェルミ + エネルギー)
   - 抗甲状腺剤 (正解: 抗 + 甲状腺 + 剤)

名詞句分割タスク

名詞句候補

常山城

関連
テキスト

常山城は日本の城。
・・・にまたがる常山にあった山城で
常山山頂からは児島湾・・・
最寄駅　　ＪＲ宇野線常山駅

名詞句分割

常山 + 城

統計的名詞句分割

名詞句候補

常山城

関連
テキスト

常山城は日本の城。
・・・にまたがる常山にあった山城で
常山山頂からは児島湾・・・
最寄駅　　ＪＲ宇野線常山駅

- 常 + 山城 ？
  - 「常」は単独では出現しない
  - 「山城」の出現もわずか

名詞句

構成要素 (候補)

統計的名詞句分割

名詞句候補

常山城

関連テキスト

常山城は日本の城。
・・・にまたがる常山にあった山城で
常山山頂からは児島湾・・・
最寄駅　　ＪＲ宇野線常山駅

- 常山 + 城 ？
  - 「常山」はテキストに頻出
    - 単独で
    - 他の名詞句の一部として

名詞句

構成要素 (候補)

注意: 名詞句同定も自明でない

ちんすこうとの・・・

形態素解析+名詞句抽出

ちん + すこう(好こう/空こう/透こう) + と + の + ・ ・ ・

名詞句?

- 名詞句の自動抽出は誤ることがある
- テキスト全体の分割を考える

統計的単語分割モデル [Goldwater+ 2009]

1. ユニグラムモデル
   - P(常山 + 城) = P(常山) P(城)
   - 事前分布をDirichlet過程とすると、任意の文字列を形態素候補として扱える
   - 高頻度のコロケーションを1形態素扱いする傾向
2. バイグラムモデル
   - P(常山 + 城) = P(常山|#) P(城|常山) P(#|城)
   - 階層Dirichlet過程によりユニグラムで平滑化
   - ユニグラムの問題を緩和

推論: Gibbsサンプリング

- 局所区間の確率的分割変更を繰り返すランダム探索

サンプリング手法	長所	短所
崩壊サンプリング		収束が遅い
型別サンプリング [Liang+ 2010]	収束が速い	バイグラムモデルに適用できない
混成型別サンプリング (提案手法)	収束が速い バイグラムモデルに適用できる	

形態素解析結果の微修正ですませたい

常山城は・・・にまたがる常山にあった

形態素解析による分割初期化

常 + 山城 + は + ・ ・ ・ + ・ ・ ・ + に + またがる + 常 + 山 + に + あった

統計的モデルによる分割修正

常山 + 城 + は + ・ ・ ・ + ・ ・ ・ + に + またがる + 常山 + に + あった

崩壊サンプリング(ユニグラム)

+ 常 山 + 山 頂 + か ら + は +

+ 常 + 山 + に + あ っ た +

+ J + R + 宇 部 + 線 + 常 + 山 + 駅

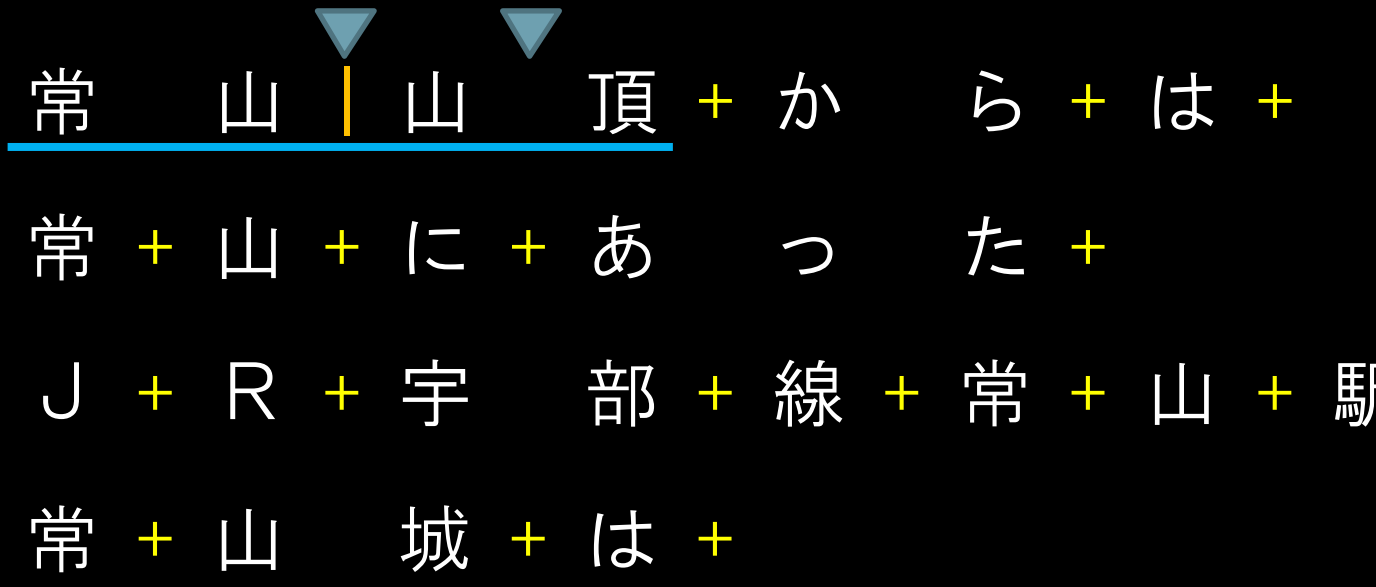
+ 常 + 山 城 + は +

z- : 注目している局所区間以外の現在の分割

境界あり: P(常|z-) P(山|z-, 常)

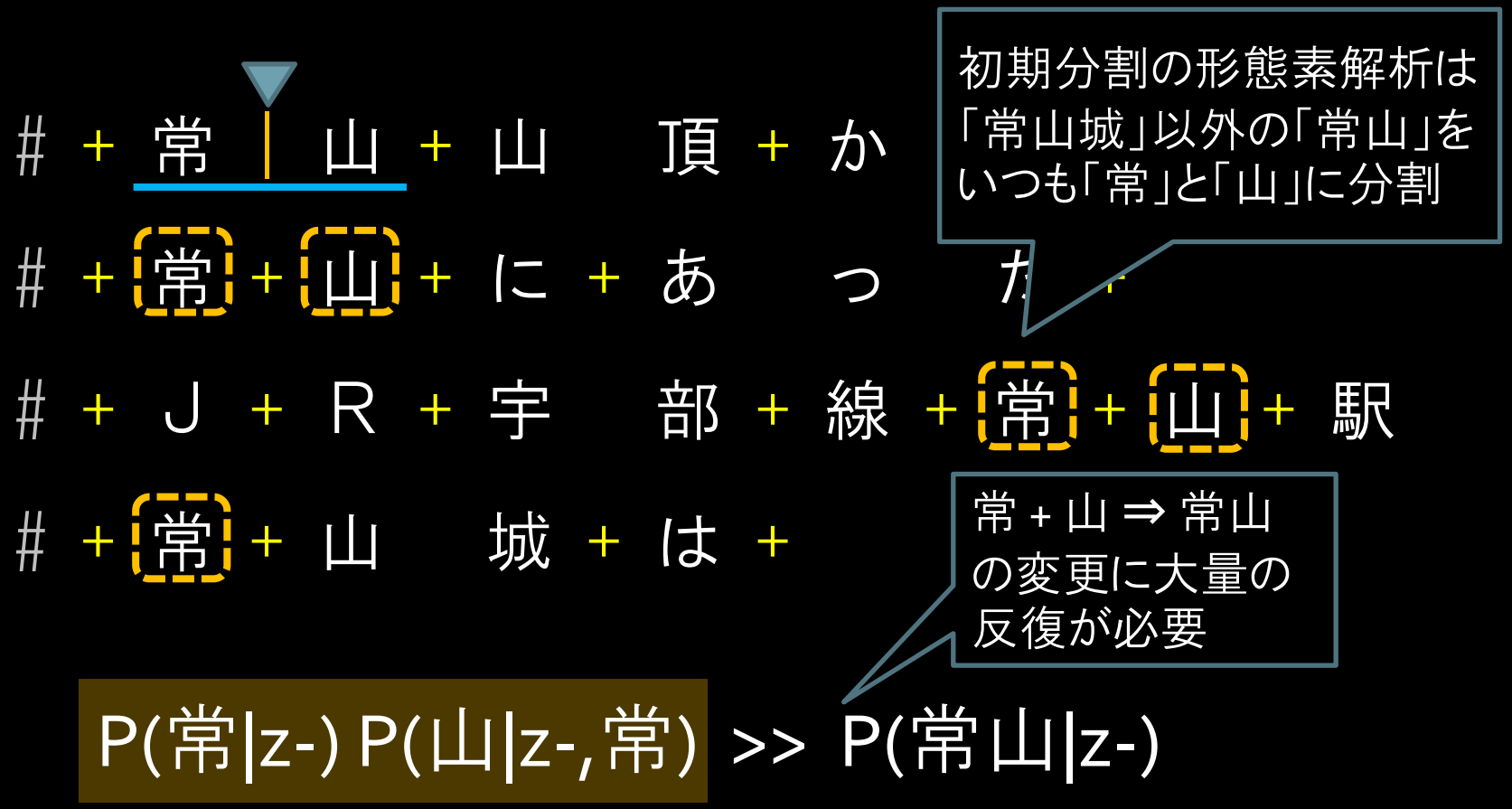
境界なし: P(常山|z-)

崩壊サンプリング(ユニグラム)

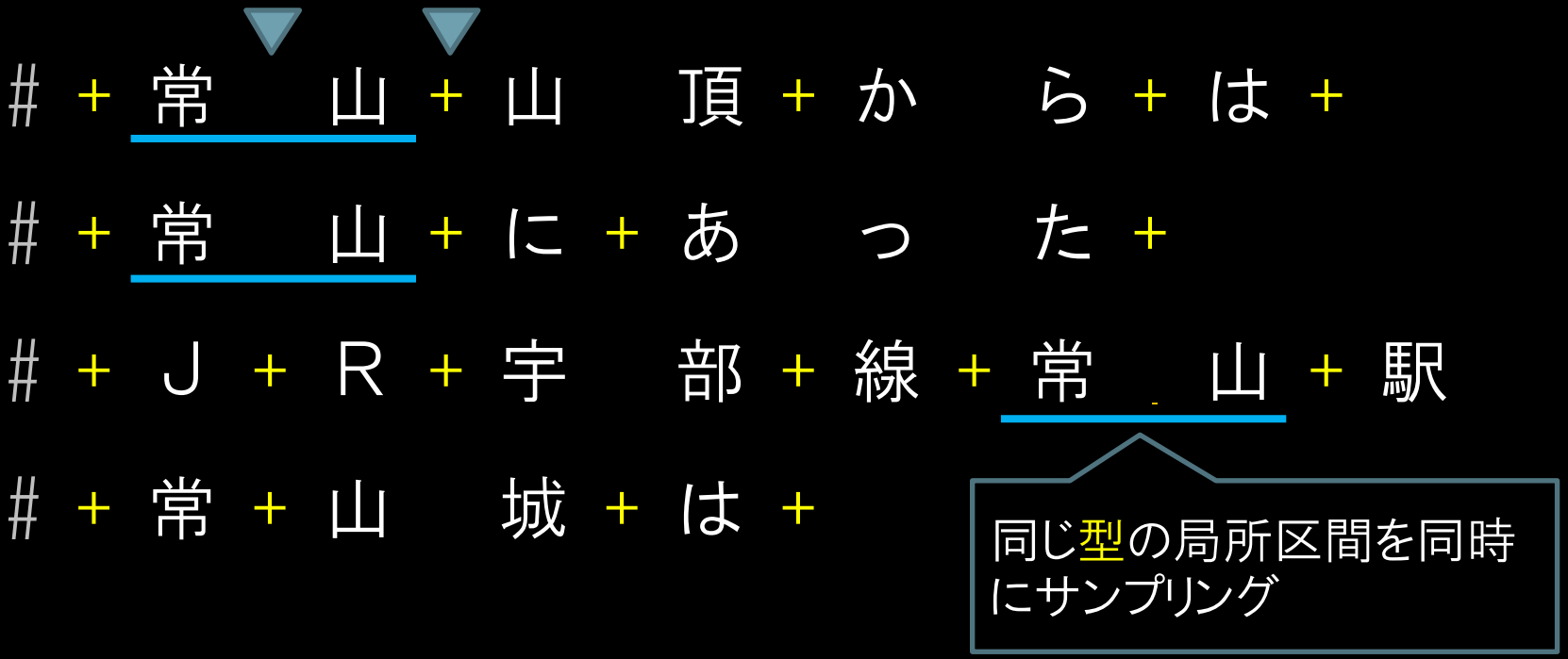


境界あり: P(常山|z-) P(山頂|z-,常山)
境界なし: P(常山山頂|z-)

崩壊サンプリングは収束が遅い



型別サンプリング(ユニグラム)



P(3か所とも「常 + 山」|z-’)
<< P(3か所とも「常山」|z-’)

バイグラムモデルの型

+ 常 | 山 + 山　頂 + か　ら + は +

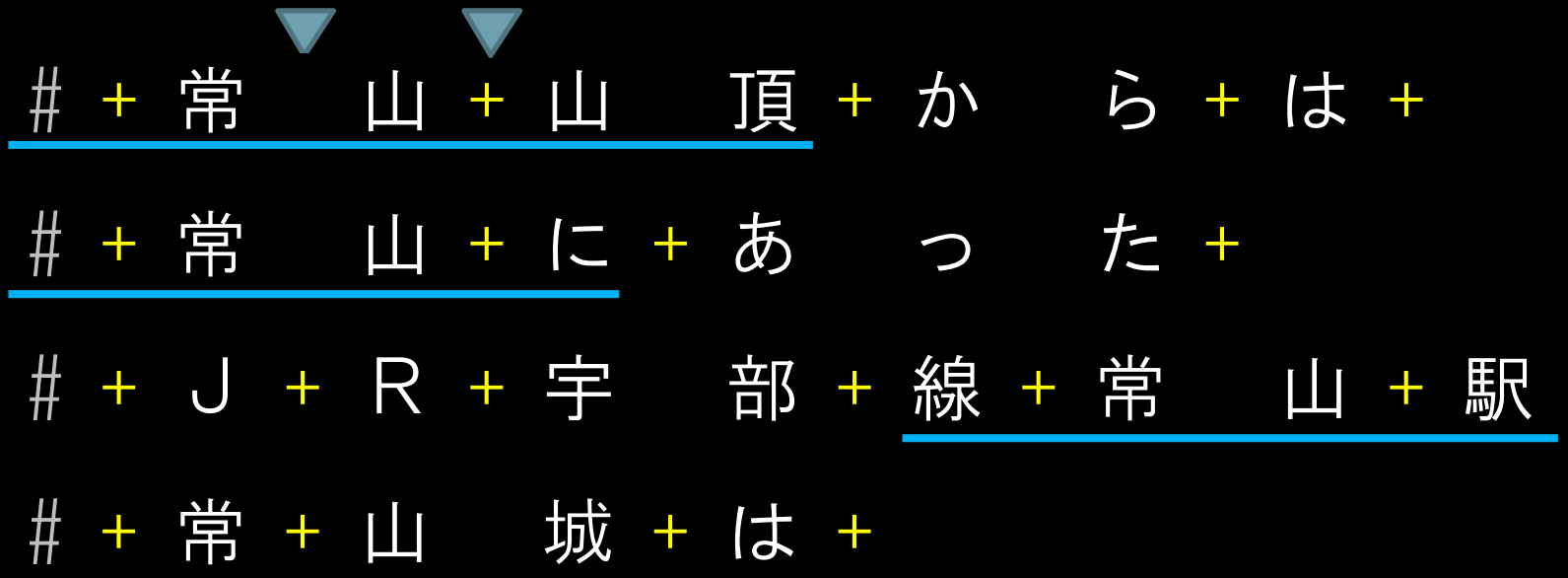
+ 常 | 山 + に + あ　っ　た +

+ J + R + 宇　部 + 線 + 常 | 山 + 駅

+ 常 + 山　城 + は +

ユニグラムモデルと違って
同じ型に属さない

提案手法: 混成型別サンプリング



1. バイグラムレベルでは別の型だけど、同時確率を計算
2. Metropolis-Hastingsで補正

実験: 設定

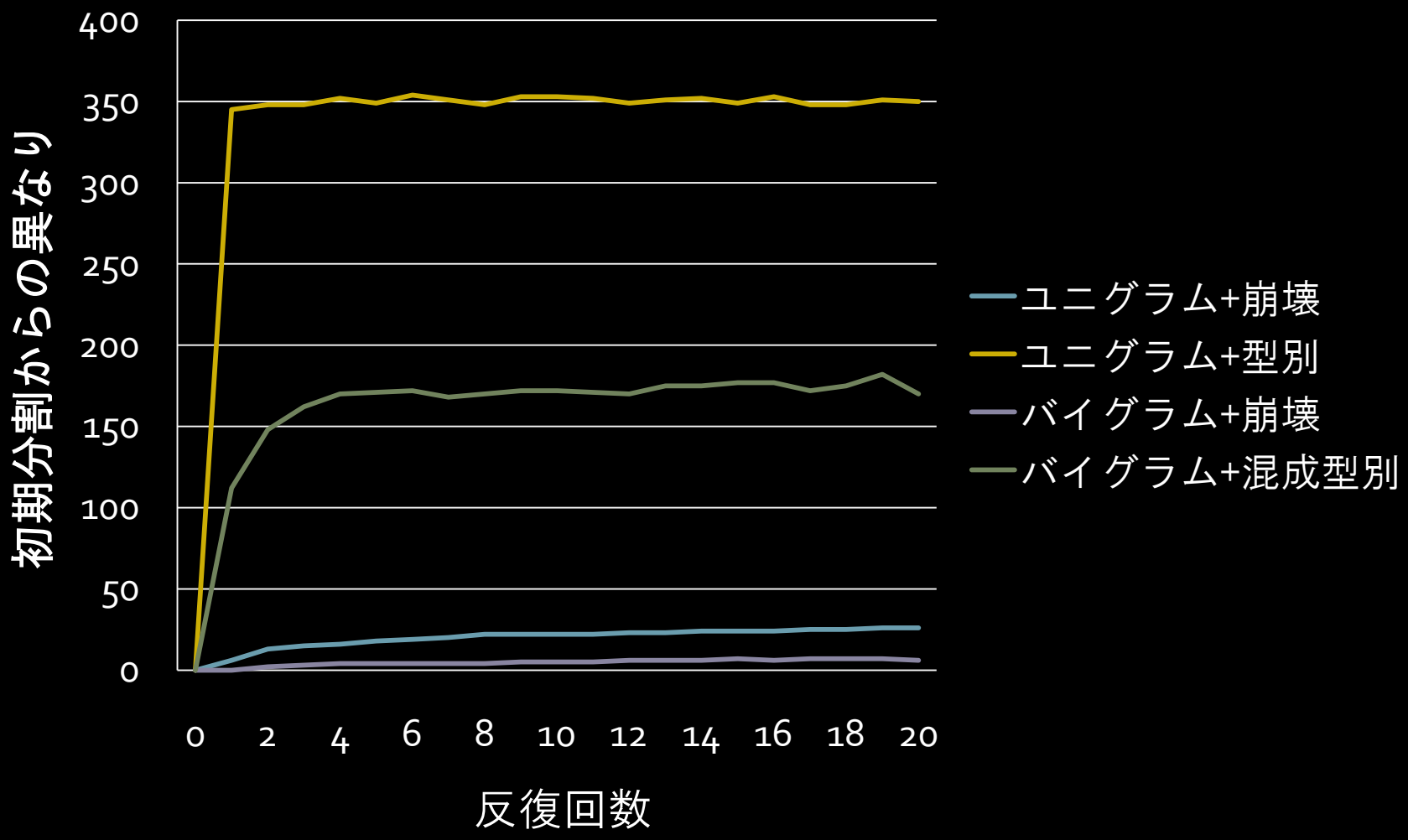
- データ
  - 名詞句候補: 日本語版ウィキペディアの見出し語
    - 500個に人手で正解を付与
  - 関連テキスト: 記事本文
- 手順
  - バーンイン10回の後、10反復から得られた分割サンプルの多数決

実験: 名詞句の分割精度

モデル	F値 (適合率/再現率)
ユニグラム + 崩壊サンプリング	80.94 (77.41/84.80)
ユニグラム + 型別サンプリング	68.84 (77.61/61.85)
バイグラム + 崩壊サンプリング	80.23 (75.76/85.27)
バイグラム + 混成型別 (提案手法)	**83.81** (82.39/85.27)
形態素解析結果 (ベースライン)	80.09 (75.80/84.89)

ただしハイパーパラメータの設定に敏感

収束速度


初期分割からの異なり
400
350
300
250
200
150
100
50
0
0
2
4
6
8
10
12
14
16
18
20
反復回数
ユニグラム+崩壊
ユニグラム+型別
バイグラム+崩壊
バイグラム+混成型別

実験: 結果例

- 改善（初期分割 ⇒ 修正後）
  - 宇部 + 興 + 産 ⇒ 宇部 + 興産
  - こな + みる + く ⇒ こなみるく
  - ちん + すこう ⇒ ちんすこう
- 誤り
  - 保土 + ケ谷 + 区 (「保土ケ谷区」、「瀬戸ケ谷」の影響)
  - 東 + ソー

結論

- 統計的単語分割による名詞句分割
- 形態素解析結果を効率良く修正する推論手法を提案
- 課題
  - 頑健性の向上
  - テキストからの語彙獲得への統合