# 大規模コーパスを用いた高速な形態素解析

岡野原大輔　辻井潤一

東京大学情報理工学系研究科コンピュータ科学専攻

School of Computer Science, University of Manchester / NaCTeM

## 概要

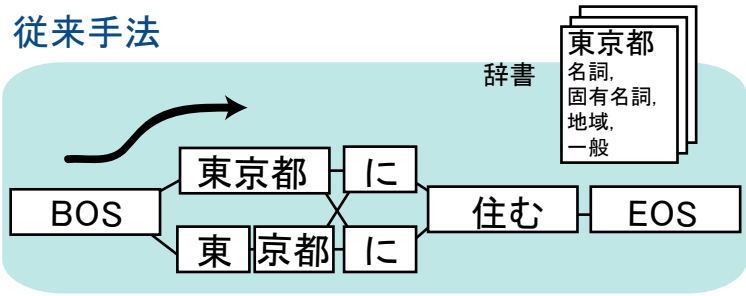
Shift/Reduce操作に基づく高速な形態素解析を提案する. 従来の形態素解析では辞書を用いて形態素ラティスを構築しラティス上での最適パスを求めて解析する方法がとられてきた. しかし, 未知語処理を行う場合、多くの候補ノードが必要になり効率的に処理できない. 本提案手法は未知語処理を含めて効率的に形態素解析を行うために文字単位のShift, 単語単位のReduce操作を組み合わせることを提案する. さらに大規模コーパスの情報を利用したフィルタリングを適用し,効率的に解析できることを示す.

## 背景

形態素解析：文中の単語とそれらの品詞を同定するタスク

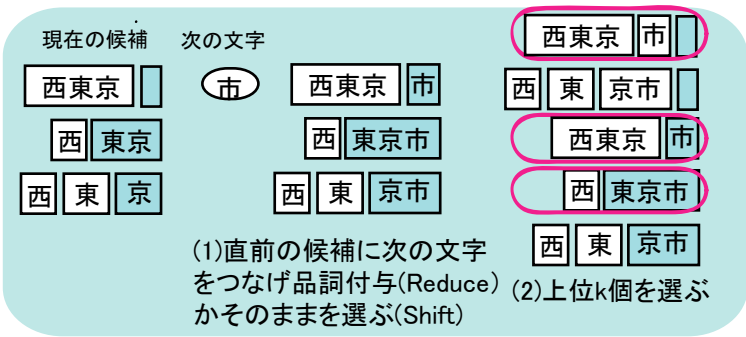
日本語の形態素解析の特徴として (1) 明示的な単語境界が存在しない、(2)品詞の種類数が多い ことが挙げられる
大規模な辞書は存在しているが、近年評判分析をはじめとして、未知語処理がますます重要になっている.

## 従来手法




大規模な辞書を基に可能な系列ラベリングを形態素ラティスで表現. 重みが最大(最小)になるパスを求めることで形態素解析を行う. ノード,ノード間の重みはHMM, CRF, MEMMなどを用いて学習. 文字単位のラベリングに比べ高速であり、かつ辞書情報が利用できるため精度が高い。しかし、未知語を考えるためには多くの候補を考えなければならない.

## 提案手法




Shift/Reduce操作を行い単語,品詞同定を行う。各ステップでは上位k個の解析候補を保持する。1文字ずつ読み込み各候補の最後のチャンクにつなげ、品詞付与(Reduce)と、そのまま(Shift)の両方を候補に加える. 品詞付与は品詞階層に基づき再帰的に行う.

長所：無駄な辞書マッチングが必要がない。辞書情報を必要に応じて利用可能. 未知語を自然に扱える.

短所：全候補を網羅できない、後方の情報を前方に使えない.

## 分類モデル

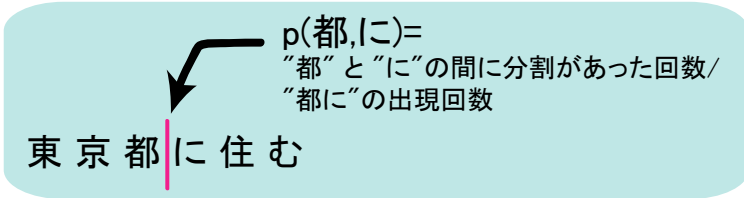
Φ(y): 形態素解析結果yに対する素性ベクトル
w　: 学習結果の重みベクトル
としたとき、<w,Φ(y)>をスコアとして線形分類器を用いる. 素性ベクトルには遠距離素性も利用可能であるが、解析候補をコンパクトに表現できなくなる.

## 学習手法

Averaged Perceptron により学習を行う. 各訓練事例について,現在の学習器の予測した結果y'と正解yが違う場合,重みwを
w += Φ(y) − Φ(y')
と更新する. 最終的に得られる重みベクトルは全ステップでの重みの平均を用いる.

## フィルタリングによる高速化

各文字間での分割確率を文字Bi-gramで求める。その確率が閾値a以下だったらShift操作のみ、閾値b以上だったらReduce操作のみを行う。どちらでもなければ、両方適用し両方候補に入れる. 文字Bi-gramの情報は、大量のコーパスを既存形態素解析器で処理した結果を用いる。




## 予備実験

はじめにフィルタリングの実験を行った. 日本語Wikipediaに対しMecab (ver 0.96)を用いて形態素解析を行い,文字間分割確率を求めた. 閾値aと閾値bはそれぞれ0.05, 0.999を用いた. 分類器の結果はS(Shift)かR(Reduce)かU(unknown)のいずれかである. これを京大コーパスの1月1～8日分に対し適用した結果は以下の通りである.

| | |
|---|---|
| Shiftの精度 | 99.0% (98507/99498) |
| Reduceの精度 | 99.9% (134599/134607) |
| U/(S+R+U) | 28.5% (93389/327494) |

この結果からほぼ間違えずに候補を約1/4にできていた.

また、京大コーパスのはじめの1000文を訓練データとして本手法を学習し、1月9日分でテストデータとして用いた。JUMAN品詞体系を用いた. パラメータ調整などはしていなく、暫定的な結果である。結果、(単語区切り/品詞/細品詞)の精度/再現率はそれぞれ (83.4/78.5/74.8) (83.9/79.0/75.3) であった.

## まとめ・今後の予定

Shift Reduce型の形態素解析を提案し、フィルタリングを用いることで効率的に行うことができることを示した. 今後は実装・実験を進め本手法の評価を進めるとともにブログなど実際に未知語が多いデータ上で精度測定を行う予定である。