日本語形態素解析

内山将夫@NICT
mutiyama@nict.go.jp

日本語形態素解析

- 入力文を形態素に分割し，各種の情報をつけること

例

「今日は良い天気だ.」を茶筌で解析すると以下のようになる.

今日　キョウ　今日　名詞-副詞可能
は　ハ　は　助詞-係助詞
良い　ヨイ　良い　形容詞-自立　形容詞・アウオ段　基本形
天気　テンキ　天気　名詞-一般
だ　ダ　だ　助動詞　特殊・ダ　基本形
.　.　.　記号-句点

日本語形態素解析の重要性

- 日本語形態素解析は，日本語処理の最初の方のステップである．
- 機械翻訳においても，入力文は，単語に分割されていることを仮定している．
- Web 検索においても，入力質問や Web ページは，形態素解析される．

形態素解析の難しさ

- 入力文には，区切の曖昧さがある
- 入力文には，辞書にない単語がある．

問題：区切の曖昧さの例 (15分)

辞書に

- す
- すもも
- も
- もも
- の
- うち

という7個の単語があるとき,

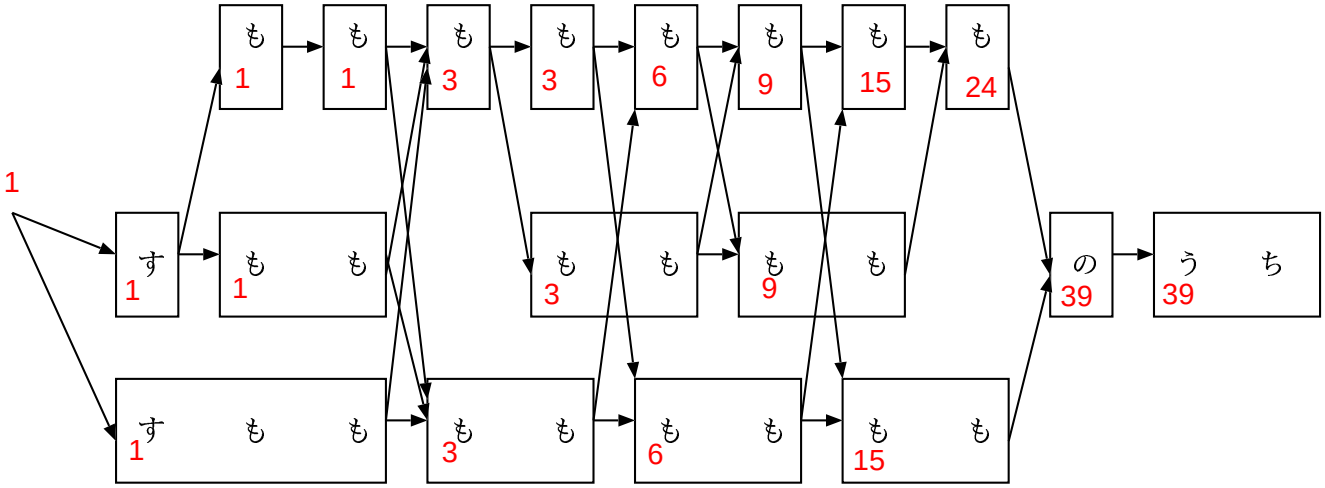
- すももももももものうち

の区切の曖昧さにはどのようなものがあるかを，あげて下さい．たとえば,

- すもも｜も｜もも｜も｜もも｜の｜うち
- す｜も｜もも｜も｜もも｜もも｜の｜うち

などがありますが，他にも色々あります.
全部で何通りの区切り方があるでしょうか.

形態素解析の曖昧さの例



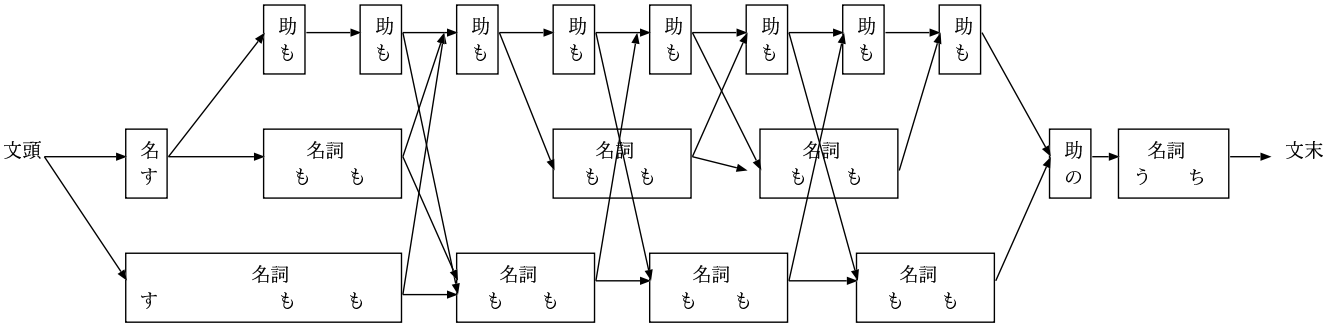
全部で39通りの区切の曖昧性があります．(実は47でした．上図では「もも」が一個ありません)

最初の時点では，区切の曖昧さは1通りです．枝わかれするにつれて，区切の曖昧さは増加します．区切の曖昧さを後の方に伝播していくことにより，区切の曖昧さの数を計算できます．

茶筌による解析結果

すもも	スモモ	すもも	名詞-一般
も	モ	も	助詞-係助詞
もも	モモ	もも	名詞-一般
も	モ	も	助詞-係助詞
もも	モモ	もも	名詞-一般
の	ノ	の	助詞-連体化
うち	ウチ	うち	名詞-非自立-副詞可能

区切の曖昧性の解消法



上図のパスのうちで，もっとも確からしいパスを通ることにより，最適な分割ができる．

パスの確からしさは，上図のノードとエッジにコストを与えることにより，パスのコストを，ノードのコストとエッジのコストの和で表現し，そのパスのコストが最小のパスを選択する．

スコアの例

ノードのコスト

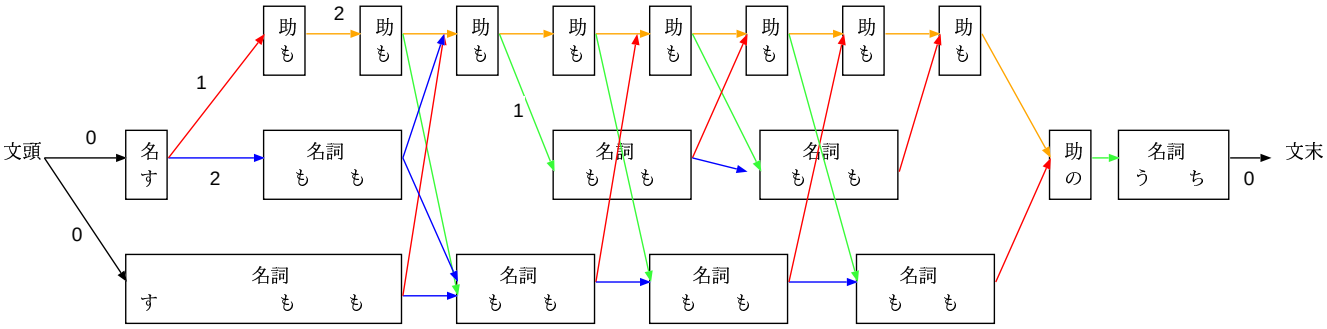
- も (助詞) $= 1$
- の (助詞) $= 1$
- もも (名詞) $= 1$
- うち (名詞) $= 1$
- す (名詞) $= 1$
- すもも (名詞) $= 1$

単純に，どのコストも 1 と考える．

エッジのコスト

- 文頭に名詞が接続するとき $= 0$ (とてもよくある)
- 名詞に文末が接続するとき $= 0$ (あまりないが簡単のため)
- 名詞に助詞が接続するとき $= 1$ (よくある)
- 助詞に名詞が接続するとき $= 1$ (よくある)
- 名詞に名詞が接続するとき $= 2$ (やや少ない)
- 助詞に助詞が接続するとき $= 2$ (やや少ない)

問題：区切の曖昧さの例 (15分)



名詞＝＞名詞（2）、助詞＝＞助詞（2）、名詞＝＞助詞（1）、助詞＝＞名詞（1）

最小コストパスはどれでしょうか．そのコストはいくらですか．最小コストパスを求める一般的な方法についても考えて下さい．

回答

- 最小コストパス

  すもも (名詞) も (助詞) もも (名詞) も (助詞) もも (名詞) の (助詞) うち (名詞)

- 合計コスト $= 13$
- 最小コストパスを求める一般的な方法 (課題とします)

  動的計画法により求めることができる．フレーズベースのSMTのデコーダーと類似のアルゴリズムを利用できる．

どのようにコストを決めるか

- コーパスベースでコストを決めることができる

$$\begin{aligned}&\arg \max_{\text{形態素列}} P(\text{形態素列} \,|\, \text{文字列}) \\ &= \arg \max_{\text{形態素の文字列=文字列}} P(\text{形態素列})\end{aligned}$$

$$\begin{aligned}P(\text{形態素列}) &= P((w_1,t_1),(w_2,t_2),...) \\ &= P(w_1,w_2,...,w_n|t_1,t_2,...,t_n)P(t_1,t_2,...,t_n) \\ &= \prod_i P(w_i|t_i)P(t_i|t_{i-1})\end{aligned}$$

ただし，w_iは単語で，t_iは品詞とする．これより

$$\begin{aligned}&\arg \max_{w_1^n,t_1^n} \prod_i P(w_i|t_i)P(t_i|t_{i-1}) \\ &= \arg \min_{w_1^n,t_1^n} \sum_i(-\log P(w_i|t_i)) + \sum_i(-\log P(t_i|t_{i-1}))\end{aligned}$$

つまり，

- 単語コスト $= -\log P(w_i|t_i)$
- 接続コスト $= -\log P(t_i|t_{i-1})$

とすれば，コストを学習できる．ただし，そのためには，形態素に区切られ，かつ，品詞が付けられたコーパスが必要である．

どのようなコーパスがあるか

- 代表的な形態素解析済みコーパスとしては，京都テキストコーパスがある

これは，京都大学黒橋研究室が，作成したものである．
良質な形態素解析済みのコーパスを作成するのは，非常に大変である．
また，辞書を作成するのも非常に大変である．

どのような形態素解析器があるか

ChaSen, Juman, Mecab

まとめ

- 形態素解析における区切の曖昧性を取り扱う方法を述べた

残された問題

- 辞書にない語 (未知語) をどう扱うか
- コーパスをどう作るか
- 辞書をどう作るか

どれも難しい問題である.