

D-5-1 CRFに基づく領域依存分かち書きの自動化(D-5. 言語理解とコミュニケーション, 一般講演)

福田, 拓也 / 三浦, 孝夫

(出版者 / Publisher)

一般社団法人電子情報通信学会

(雑誌名 / Journal or Publication Title)

電子情報通信学会総合大会講演論文集

(号 / Number)

1

(開始ページ / Start Page)

41

(終了ページ / End Page)

41

(発行年 / Year)

2007-03-07

CRFに基づく領域依存分かち書きの自動化

福田拓也

三浦孝夫

法政大学工学部情報電気電子工学科
東京都小金井市梶野町 3-7-2

1 前書き

インターネット上の日本語文書から有用な情報を抽出する問題では、単語の分かち書きは基本的で重要な基礎技術である。単語間に区切りを持たないため何を単語とするかの認識が計算機による自動処理においては根源的な機能を左右する。

形態素解析システムでは、形態素間の結合規則と形態素辞書あるいは文法の知識を用いる。形態素解析とは、一般に「分かち書き」、「品詞の同定」、「未知語処理」から構成される。ここでは、経験的優先規則や文法的接続可能性、接続コスト、統計的言語モデルなどに基づいて構築される。

実用上の観点から言えば、複合語への対応も大きな課題である。例えば複合名詞(「法政大学」を「法政」と「大学」に分割)や複合動詞(「乗り継ぐ」を「乗り」と「継ぐ」に分割)などの分割は分野固有に依存する問題である。例えば、著者らの所属する大学では、「情報電気電子工学科」は「情報電気電子」「工学科」としてのみ分かち書き可能である。

本研究では、訓練コーパスから学習により特徴的なパターンを抽出し、条件付確率場(CRF)にその特性を素性として定義することで、領域依存性を有する分かち書き機構を提案する。

2 Conditional Random Fields

Conditional Random Fields (CRF) は、入力系列 X が与えられたとき、それに対する出力系列 Y の条件付確率 $P(Y|X) = \frac{1}{Z(X)} \exp(\sum_{i=0}^n \sum_k \lambda_k f_k(X_i, Y_i))$ を1つの指数分布モデルによって与える。ここで $f_k(X_i, Y_i)$ は二値を返す素性関数、 λ_k は素性関数に対する重み、 $Z(X)$ は全系列を考慮した正規化項である。

これにより CRF は HMM では困難だった柔軟な素性の取り込み、MEMM で問題だった label bias を解決する。

また入力 X に対する最適出力系列 Y は Viterbi アルゴリズムを用いて効率よく探索できる。さらに、 λ_k の推定は訓練データより、反復スケール法などを用いて値を決定する。これより、素性と適切な訓練データを与えることによってラベル付けをすることができる。

3 領域依存分かち書き

本研究では CRF で使用する素性は、単語、上位品詞(名詞、動詞)、下位品詞(固有名詞、代名詞)の3つを用いて、それぞれ前後 3gram までを素性とする。

訓練データのラベルとして各形態素に B (分かち書き始まり)、I (分かち書き途中) という2つをつける。

領域依存しているものとして複合名詞に着目する。各分野から名詞+名詞(東京-大学)の組み合わせを抽出し、*+名詞、名詞+*と置き換えてカウントする。(*は名

詞に対するワイルドカードを意味する)これは教育分野では、*+大学が特徴語となるからである。さらに、名詞+名詞+名詞も同様に抽出する。抽出した語から上位一定数をその領域依存する固有名詞とみなし、分かち書きをさせない。名詞以外の分かち書きに関しては点字規則を用いて分かち書きをする。

4 実験

本実験では、「公開特許公報全文データ(98,99)」から農業分野訓練データ:13276 形態素、器械分野訓練データ:12795 形態素を使い2種類の CRF モデルを生成する。さらに、既存の分かち書き手法として点字ソフト「お点ちゃん」を使う。本研究で使用する CRF は CRF++, 形態素解析は Chasen を使用する。

上記データ集から、農業分野テストデータ:3886 形態素、器械分野テストデータ:3569 形態素を農業分野モデル、器械分野モデル、お点ちゃんそれぞれ分かち書きをし、不正解数をカウントする。その値から領域依存分かち書きがなされていることを判断する。

	農業分野モデル	器械分野モデル	お点ちゃん
不正解数	132 (3.40%)	378 (9.72%)	273 (7.03%)

図1 農業分野テストデータ

	農業分野モデル	器械分野モデル	お点ちゃん
不正解数	206 (5.77%)	67 (1.88%)	86 (2.41%)

図2 器械分野テストデータ

5 考察

図1より、農業分野テストデータの分かち書きは農業分野モデルが不正解数が一番少ない。よって、農業分野では農業分野モデルが精度のよい分かち書きができる。図2より、器械分野テストデータの分かち書きは器械分野モデルが不正解数が一番少ない。よって、器械分野では器械分野モデルが精度のよい分かち書きができる。また、お点ちゃんと比較してもテストデータとモデルの分野が同一であると、お点ちゃんより制度のよい分かち書きができています。この結果より、CRF による領域依存分かち書き手法の有効性を示せる。

6 結び

本論文では、CRF を用いて領域依存した分かち書きの有効性を示した。本研究では名詞のみを対象としたが、動詞や形容詞などに対しても領域依存性を検証する予定である。

参考文献

- [1] 工藤 拓: 形態素周辺確率を用いた分かち書きの一般化とその応用, 言語処理学会全国大会 NLP-2005, 2005
- [2] 工藤 拓, 山本 薫, 松本 裕治: Conditional Random Fields を用いた日本語形態素解析, 情報処理学会自然言語処理研究会 SIGNL-161, 2004
- [3] J. Lafferty, A. McCallum and F. Pereira: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Intn'l Conf. on Machine Learning(ICML), 2001