

複合名詞内アクセント句境界を用いた アクセント結合予測の高精度化に関する実験的検討*

○高野克弥, 清水信哉, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

日本語テキスト音声合成システムにおいて、文中のアクセント核位置を適切に推定することは、自然な読み上げ音声の出力に対する重要な要件である。各々の単語は独自のアクセント核位置情報を持つが、それらの単語を連続して発声するとアクセント核の位置が頻繁に変化する(アクセント結合)。我々は、CRFを用いてアクセント核位置の変化予測を試みてきたが、複合名詞に関して十分な精度が得られていなかった。そこで本稿では、4形態素以上の複合名詞に対して係り受け解析を行い、そこから得られる(複合名詞内)アクセント句境界を利用することで精度向上を検討した。

2 CRFによるアクセント核位置変化予測

Unidic1.3.8を参照させた形態素解析システムChasenによる解析結果に、「アクセント句境界」と「アクセント核位置」の情報を単独ラベラに付与させ、学習データ(6,753文)と評価データ(527文)を用意する。素性としては、[基本形/基本形読み/書字形/品詞/活用型]、[品詞]、[活用型]、[活用形]、[モーラ数]、[単独発声アクセント型]、[組み合わせ素性](当該形態素の単独発声アクセント型と直前・当該・直後形態素のアクセント結合様式および結合アクセント価)、[単独型種類ラベル]、[アクセント句内の相対位置]、[特定位置のモーラ]を用いている。

CRF++によりアクセント核位置を推定すると、全てのアクセント句単位での正解率が92.6%であるのに対し、名詞連続部分のアクセント句においては正解率が89.0%であった。[2]

3 複合語名詞のアクセント格位置

3.1 複合名詞データの準備

4から8つの形態素からなる複合名詞を800語(割合は、4:50%、5:25%、6:20%、7:4%、8:1%)を新聞記事より抽出した。Mecabを用いて形態素解析を行い、「単独発音アクセント核位置」をUnidic1.3.12から参照し、連続発声時のアクセント核位置情報を人手でラベリングを施した。これを、学習データ650語と評価データ150語に分割した。

3.2 CRFによるアクセント核位置推定

CRF++によるアクセント核位置を推定すると、形態素単位で77.7%の精度が得られた。素性としては、基本的には2章で用いた素性をベースに定義している。2章における複合名詞部分と比べて精度が極端に低いのは、より長い(4~8形態素)複合名詞を用いているためである。

3.3 複合名詞に対する係り受け解析の利用

4つ以上の形態素から形成される複合語は、アクセント核を複数所持することがあるため、複合語に適切なアクセント句境界を設けることが精度向上へと繋がると期待される。実際に日本人が複合名詞を朗読する際は、通常2,3単語を一つのアクセント句として認識することが多いので[1]、適切にアクセント句境界を挿入する必要がある。

適切なアクセント句に分割するためには、複合名詞の係り受け解析によって得られる木構造(Fig.1)を基に、[1]の手法に従って複合名詞のアクセント句抽出を試みた。具体的には、構成する単語間の係り受け関係を支配する語彙的性質に着目し、木構造を人手で作成した。その後、その木構造を参照して、下記の手順でアクセント句を抽出した。

- 1) 最上位節点をはじめ、順に下位節点を見ていき、
- 4単語以上の単語列が境界の設定によって無くなる

* Experimental study on improving the performance of accent sandhi prediction using accentual phrase boundaries in a compound noun, Katsuya Takano, Shinya Shimizu, Nobuaki Minematsu, Keikichi Hirose(University of Tokyo).

まで2)の作業を繰り返す。

2) 木構造の最上位節点から、A) 両端の枝が木構造(複数の形態素から形成される纏り)の場合、B) 左枝が単語で右枝が木構造の場合、C) 左枝が木構造で右枝が単語の場合のいずれであるかを判断し、以下の処理を行う。

A) の場合：最上位節点に境界を設定する。

B) の場合：最上位節点に境界を設定する。

C) の場合：最上位から3つの節点までが左枝が木構造で右枝が単語のとき、Table 1に従って、それぞれの形態素間の結合力を求める。ただし、この途中でA)やB)の構造が現れた時は、その節点に境界を設定する。3つの結合力が得られたところで、前方から相対的に結合力が弱い場所を探し、その節点に境界を設定する。

3.4 アクセント核位置推定実験

木構造 (Fig. 1) の様子を素性化し、3.2 で用いている素性に追加した。Fig. 2 の左から2番目の枠内に木構造を1/0で表現している。アクセント核の推定精度は、81.2%という結果が得られた。

次に、木構造から得られるアクセント句境界を3.2の素性に追加した。Fig. 2の左から3番目の枠内であり、「日本」と「弁護士連合会」に分割されることを意味する。アクセント核の推定精度は、85.7%という大進展が見受けられた。

3.5 考察

特定領域研究「日本語コーパス」電子化辞書班により提供された中単位解析器を用いて係り受けを自動抽出した。得られた境界情報を、3.4と同様に素性化すると、79.1%という精度が得られた。Fig. 2の一番右の枠内であり、「日本弁護士」と「連合会」に分割されることを意味する。

3.4で高精度が得られたのは、木構造を手で作成したためだと考えており、高い精度での自動境界抽出が今後必要になってくると思う。

Table1 2単語間の結合力

前方単語	後方単語	結合力
非用言性名詞	非用言性名詞	3
用言性名詞	非用言性名詞	2
名詞	用言性名詞	1
名詞	接尾辞	4

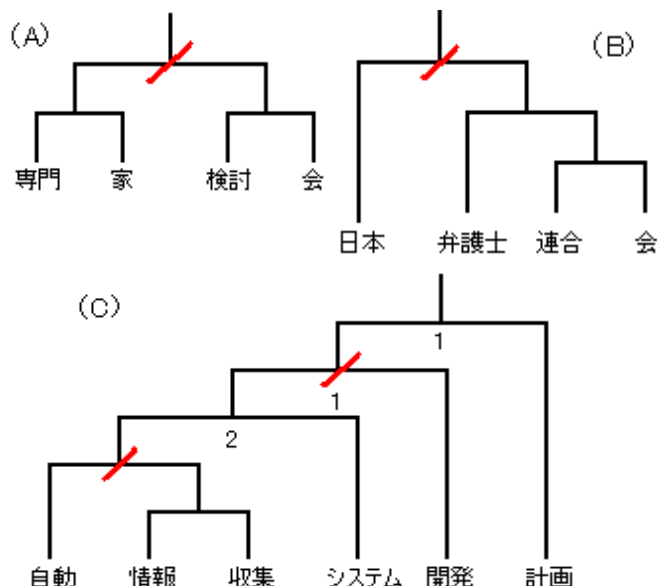


Fig. 1 複合名詞の木構造

日本	4	1	*	*	1	*	*	1	1	2	1
弁護士	4	2	*	1	0	*	*	3	1	2	2
連合	4	3	1	0	*	*	*	3	2	2	1
会	4	4	0	*	*	*	*	3	3	2	2

Fig. 2 複合名詞の素性

4 まとめ

複数のアクセント核が存在し得る複合名詞について、複合名詞の木構造を基に抽出されるアクセント句境界を素性に加えることで精度の向上を実現した。

接尾辞(「-的」「-性」「-向け」など)として分類される形態素の直前の形態素内のアクセント核位置誤推定は、全誤推定のうち52%を占めた。規則を用いて別の処理を施したり、重点的にエラー解析を行うことで更なる精度向上へと繋げたい。

謝辞

本研究にご協力を下された「日本語コーパス」電子化辞書班の皆様方や富士通の印南様には、深く御礼を申し上げます。

参考文献

- [1] 藤石知夏, 宮崎正弘, 日本語複合語構造解析に基づく複合語アクセント句の自動抽出法, 情報処理学会第49回全国大会, 2-51, 1994
- [2] 印南圭介, 規則処理を参考にしたCRFによるアクセント結合処理モデル, 日本音響学会講演論文集, 2-P-13, 2009