近接単語の並びに着目した形態素解析多義の絞り込み

5B－5

白井 諭† 池原 悟† 井上 みづほ‡
NTT コミュニケーション科学研究所†
慶応義塾大学 政策・メディア研究科‡

1 はじめに

多段解析法に基づく日本語形態素解析では，形態素単位99.5% の解析精度を達成したことを既に報告している [1]. その際，構文的・意味的関係を見て形態素の正誤を判断すべき場合は解析多義の中に正解が含まれれば解析成功としていた．形態素解析単体としての利用を考えると，これらの多義は少ないほどよい．そこで本稿では，個々の品詞・単語の特性に基づく解析多義の解消を目的として，新聞記事965 文に現れた解析多義 520 件を詳細に分析し，解析多義の絞り込み規則について提案する．

2 多義の分類

日経産業新聞 965 文（情報欄リード）に対する形態素解析結果に含まれる多義 520 事例を対象とした．全体を分類し，その中から出現件数の多い 10 事象を抽出した (表 1 参照)．現在の多義を把握するとともに，それを絞り込む規則を考える上で戦略を立てるという目的がある．

抽出した事象は以下の通り．

1. 自動詞，他動詞の区別 (ex. 開く)
2. 自動詞の連用形，自動詞転生名詞 (ex. 受け)
3. 他動詞の連用形，他動詞転生名詞 (ex. 取り付け)
4. 自動詞転生名詞，他動詞転生名詞の区別
5. ひらがなの入力に対し，複数の漢字表記があるもの (ex. なる → 成る，鳴る，生る)
6. 漢字の入力に対し，複数の読みがあてはまるもの (ex. 開く → ひらく，あく)
7. 形容語に対し，述語として使われているか副詞的に使われているかの区別 (ex. 近く)
8. 品詞は同じで，意味属性のマッピングが違うもの (ex. 日 (= 暦日 / 非暦日))
9. 一般名詞，形式名詞の区別 (ex. もの)
10. 助動詞，格助詞の区別 (ex. で (だ))
11. その他

*Disambiguation of Japanese Morphological Analysis Supported by Local Syntax and Semantics

†Satoshi Shirai and Satoru Ikehara, NTT Communication Science Laboratories

‡Mizuho Inoue, Keio University Media and Governance

3 多義の絞り込み

形態素多義の多くは深い解析 (構文解析，意味解析等) により解消すべきであるが，部分的に深く解析することによりある程度の多義は解消できあることが知られている [2]. そこで，この観点から再度検討を加えることにした．具体的には，多義が生じた語を含む文節の直前直後の文節，あるいは数個前までの文節を見ることにより多義絞り込みの可能性を検討した．

事象 1,2,3,4,7,10 については，品詞 (各々の事象) について適応される規則を考えることができる．それ以外の事象については，各々の語に別個の規則を考えた．

3.1 規則適応の例

1 つ例を示す．

「光ファイバーは米国メーカー製品の流入などもあり価格低下が急ピッチで進んでいる．」

この文を形態素解析すると，次のように単語分割される．[1]

「光ファイバー / は // 米国 / メーカー / 製品 / の // 流入 / など / も // あり // 価格 / 低下 / が // 急ピッチで // 進ん / でいる /．」

この文では「急ピッチで」の品詞解釈に，1) 形容動詞連用形 2) 形容動詞副詞形の多義がある，事象 7 に分類される．これに対し考えた規則は，

直前の文節に数詞がなく，かつ直後の文節に動詞があれば，副詞的に使われている方を選ぶ．

というものである．この文では，直前の文節は数詞でなく，直後の文節に「進む」という動詞があるので，規則が該当し多義は取り除かれる．このように，全ての事例について，規則が考えられるかどうか，またその規則が該当するかを調べた．

3.2 規則の例

まず，規則に現れる言葉の定義を挙げる．

[1]「/」は文節区切り以外の形態素区切り，「//」は文節区切りである形態素区切りを表す．

直前 − 同じ文節内で，多義語の前
直後 − 同じ文節内で，多義語の後ろ
直前の文節 − 多義語がある文節の，1つ前の文節
直後の文節 − 多義語がある文節の，1つ後ろの文節
自分の前 − 多義語がある文節と，そこから前をたどって，文頭にたどり着く，または句読点が出てくる，または動詞が出てくるまでの間

また，特記がない場合は，列挙された規則以外には，多義候補には手を加えない(消極的規則).

1. 自動詞・他動詞の区別
   (a) 受動態ならば他動詞を選ぶ.
   (b) 自分の前に「を」格があれば自動詞を候補から落す.
2. 自動詞，自動詞転生名詞の区別
3. 他動詞，他動詞転生名詞の区別
   (a) 直後に助詞があれば名詞を選ぶ.
   (b) 直前の文節に「の」格があれば名詞を選ぶ.
   (c) 上記以外は動詞を選ぶ.
4. 自動詞転生名詞，他動詞転生名詞の区別

   この事象は1,2,3との複合で起こることが多く，それらのルールの組み合わせによって多義は取り除かれる.
7. 形容語に対し，述語として使われているか副詞的に使われているかの区別
   (a) 直前の文節に数詞がなく，かつ直後の文節に動詞があれば，副詞的に使われている方を選ぶ.
10. 助動詞，格助詞の区別
   (a) 直前に形式名詞(「うえ」「こと」を除く)があれば助動詞を選ぶ.
   (b) 直前に形式名詞(「うえ」「こと」を対象)があれば格助詞を選ぶ.
   (c) 直後の文節に動詞があれば格助詞を選ぶ(但し連体形は除く).
   (d) 直前が時・期間・今昔に属する語で，直後の文節が「数詞＋助数詞」であれば格助詞を選ぶ.

以上，例には品詞(事象)について規則を考えられるものを挙げた．10事象について立てた規則数としては，品詞に基づいたもの6事象10規則，単語に基づいたもの4事象37規則である.

3.3 結果

表1に事象ごとの事例数と規則の適用された数，ならびに事例の多義の合計と規則の適用による多義の削減数を示す.

但し，1つの文節内に複数の多義の事象が現れる場合については，それぞれについて分類したので，その分重複して数えられている．また，1つの事例に複数の事象が同時に現れる場合についても重複して数えられている．そのため，本検討で取り上げた事例数(520件)と表1の事例数の総計(564件)は一致しない.

別の事象におけるルールと組み合わさって多義が減少する場合については，それぞれの事象において数えた．また，あらわれた多義の候補の中に，正しい解釈が存在しない場合は，errorに分類した.

表1：多義絞り込み規則の適用情況と多義の削減効果

事象	事例数	適用数	多義合計	削減数
1	79	57	196	90
2	10	10	25	12
3	18	18	39	19
4	6	4	20	12
5	90	67	297	117
6	63	52	155	76
7	29	17	58	17
8	78	46	164	46
9	48	47	114	63
10	93	24	208	40
11	50	24	119	28
err	(5)		(12)	
計	564	366	1395	520

このように，事例数で見ると全体の64.9%(= 366/564)に，多義が除去されるまたは減少するなど，何らかの効果が見られる．多義削減効果を計る指標として，1事例に現れる平均多義候補数を考えると，2.47(= 1395/564)から1.55 (= (1395-520)/564)に減少しており，これも効果が認められる.

また，上の評価は同一文節内の事例の重複を許したものであるので，実質総計での評価を示すと，事例数で520事例のうち効果が見られたものは320事例(61.5%)で，多義数では2.39(1242)から1.52(789)への減少となる.

4 おわりに

多段解析法による形態素解析の多義を11事象に分類し，多義を絞り込むのに効果のある47規則を提案した．机上実験によれば，多義の生じた520文節のうち61.5%に規則が適用され，平均多義数は2.39から1.52に減少する見込みである．今後は規則を改良するとともに，処理系を実現する予定である.

参考文献

[1] 白井, 池原, 横尾, 奥山, 宮崎：多段解析法による日本語形態素解析の精度，情報処理学会第50回全国大会講演論文集 1R-2(1995)

[2] 宮崎, 大山：階層的単語属性を用いた同形語の自動読み分け法，電子通信学会論文誌 Vol.J68-D No.3(1985)