自然言語処理入門

岸山 健 (31-187002)

Oct. 22, 2018

1　課題 1

毎日新聞の記事『ＡＩでバブル絶頂を甘いチョコに　ＮＥＣと米社』[*1] 夏目漱石の『夢十夜』[*2]，そしてブログ『「若者のカメラ離れ」離れ』[*3]の記事[*4] から 2–3 文ずつ引き抜いた． それぞれに対し，講義で紹介された形態素解析器 (MeCab ipadic, MeCab unidic, JUMAN) の振る舞いがどう異なるかにまずは注目する．そしてそれらの結果に対して考察を与える．

まずは新聞記事中の文である．

> ＮＥＣと米菓子メーカー「ダンデライオン・チョコレート」の日本法人は２５日，人工知能（ＡＩ）を使って時代の雰囲気を甘さや苦さで表現した「あの頃はＣＨＯＣＯＬＡＴＥ（チョコレート）」を開発したと発表した．１２月２１日に発売する．強い甘さとジャスミンの香りで華やかな風味に仕上げた「１９８７　魅惑のバブル絶頂味」のほか「１９７４　オイルショックの混迷味」や「２０１７　イノベーションの夜明け味」など計５種類を用意した．

MeCab ipadic の結果では，12 月の数詞は「12」としているのに対し 12 日は「1」と「2」に分けていた．MeCab unidic 版は和/漢で，JUMAN は訓/音で読み方を区別しているのに対し，MeCab ipadic では分けていなかった． MeCab unidic はオイルショック，を「オイル」と「ショック」に分けており，また「ダンデライオン」も分けていた．このように分割する特徴は unidic 版で特に顕著であった．最後に JUMAN だが，一般的な情報だけではなくカテゴリやドメインといった列が与えられていた．さらに，自動獲得という機能があるようである．開発した研究室のページによると，Web テキストから自動獲得した辞書が実装されている．例えば「ジャスミン」に対しては「植物/しょくぶつ”」を「上位語」として引っ張ってきていた．

次に小説などの地の文である．

> こんな夢を見た．腕組をして枕元にすわっていると，仰向に寝た女が，静かな声でもう死にますと云う．女は長い髪を枕に敷いて，輪郭のやわらかな瓜実顔をその中に横たえている．

最初の一文は特に差はなかったが，MeCab unidic は動詞「見」に対して非自立可能とふっていた．MeCab ipadic と MeCab unidic は「仰向」を分割していた．しかし ipadic は「瓜実顔」を一つの名詞として参照しており，この文脈ではその解釈が正確に思える’． MeCab unidic は「瓜実顔」を「瓜実」と「顔」に分けており，JUMAN は「瓜実顔」を 3 つに分けていた．なお，JUMAN は「仰向」を未定義語としていた．

[*1] https://mainichi.jp/articles/20181026/k00/00m/020/015000c
[*2] https://www.aozora.gr.jp/cards/000148/files/799_14972.html
[*3] http://phobby.hatenablog.com/
[*4] http://phobby.hatenablog.com/entry/tokyototeienbijutsukan-tamron45mm

最後にブログ等で見られる会話調の砕けた文である．

> 東京は目黒にある東京都庭園美術館の庭園に行ってきました！ 関東に 12 年以上住んでいながら全く聞いたことがないくらい知名度は低いですが，渋谷から 15 分で行ける美術館＆庭園ということで期待大です．というわけで目黒にある東京都庭園美術館にやってきました．

MeCab ipadic は唯一「美術館」を一つの名詞とし，「美術」と「館」に分けなかった．また，「というわけで」の「という」を一つとして扱っていた． MeCab unidic は「知名度」を「知名」と「度」に分けていた．さらに，他の解析器は「やってくる」を一つの動詞と見なしていたが， MeCab unidic は「やる」と「来る」の 2 つに分けていた． 最後に JUMAN の結果では，「行く」に対して「帰る」を反義語として挙げていた．

総合すると，MeCab ipadic の結果は複合語を一つの名詞として扱っている傾向があった． 他方，MeCab unidic の結果はできるだけ単語を分割する傾向があった． JUMAN の結果からはこの解析器が新語に強いことが分かった．これは自動獲得という機能によるものだと思われる．さらに，カテゴリや反義語と言った，形態素解析には不要と思える情報も含まれていたように思える．

2　課題 2

始めに「形態素の体系を定義する」とは何を示すか，という点を考えると，これは「各形態素を表で考え，それぞれに列名と水準を決めること」であると考えた．例えば形態素の表の列名には「読み」や「表記」，「品詞」が最低限あり，さらに「品詞」のような各列には水準として「名詞」や「動詞」がある．つまり，「形態素に対してどのような列名を設けるか」だけでなく「各列にどのような水準を設けるか」が形態素の体系を定義するということだと思う．

形態素の定義が列名と水準を考える作業だとするならばゴールが必要となる．そして「そのゴールは単なる形態素の分割なのか否か」が一つの分かれ道になると思う． そう考える理由は MeCab ipadic が必要最低限の情報しか持たないのに対し， JUMAN はカテゴリ（ビジネスや料理，スポーツ）や反義語などの，形態素解析の作業自体にはあまり役割のなさそうな情報を保持しているからである．早とちりかもしれないが，もし JUMAN に含まれるカテゴリや音形の情報が形態素解析にあまり役割を持たないのならば，形態素解析以外の使用目的も念頭に置いているのではないだろうか．つまり，こうした違いは「その形態素の定義は何をゴールとするのか」を決めること自体が一つの問題，議題となることを示すと感じた．

上の話は「どのような列を設けるか」の議論だが，もちろん「各列にどのような水準をもたせるか」も難しい問題である．どのような水準をもたせればパフォーマンスが上がるのか，どのような論文を指標にするのか，加えて指標が変わった際に追従できるか，プロジェクトの途中で新しい人が入ってきて議論がバラバラになったらどうするか等，水準を決める際の問題も尽きないと思う．

以上の様に，形態素を定義する際は列と水準，それら両方に困難があるように思えた． また，列の問題は水準の問題よりも議論は少なさそうだが，プロジェクト全体の方向性を決める重要な議題に思える． 他方，水準の問題はプロジェクト全体と比べるとより現実的で具体的な，パフォーマンスの向上や環境の変化への対応のような印象を受けた．