# NLP に重要なのは、学習データの量なのか、言語学知識体系の質なのか？
## ―中国語単語分割タスクで検証する―

王向莉
一般財団法人日本特許情報機構
{xiangli_wang@japio.or.jp }

## 1. はじめに

中国語単語分割のようなNLP分野では、言語学知識コーパスを学習データとして利用する統計機械学習手法の研究が盛んに行われているが、精度上の限界が見られ、それを如何に打破することが課題となる。

統計機械学習手法の研究では、“data sparseness”という問題が常にあり、“robust”なパーフォマンスを実現するために、学習データの量がよく求められ、強調される。高品質の NLP タスクを実現するのに、大量の学習データさえあれば、あとは機械学習の技術に任せばよいと思う人が少なくない。ところが、NLP に重要なのは統計機械学習手法とその“data sparseness”問題を解決するための学習データの量だけであろうか？中国語単語分割タスクを例とすると、統計機械学習手法がすでにたくさん提案され(Sun et al. 2009; Gao et al. 2005; Xue 2003)、また、大量の学習データも複数構築されたのにもかかわらず(Yu S. et al. 2002; Zhou Q. 2004; Xue, Fei 2000)、特許分野での実用レベルの中国語単語分割の精度がいまだに満足できないというが現状である。中国語単語分割のようなNLPタスクの精度上の限界を打破するためにどうすればよいのか、真剣に考えなければならない時がやってきたと思われる。

NLP タスクの精度を向上させるのに、統計機械学習の性能を上げる一方、言語学知識体系を工夫することも考えられる。言語学知識体系は学習データとして、統計機械学習手法の優位性を検証するのに用いられるのが一般的であるが、その自身がNLPタスクにどれほど貢献しているのか、アピールしにくい状況である。言語に依存しない形態素解析エンジン Mecab(Mecab)の開発と公開はこのような状況に転機を与える。言語学知識体系があるならば、それをMecabに導入し、新しい単語分割ツールと品詞付けツールが作成できるため、ツールのパフォーマンスで言語学知識体系の品質を検証することが可能になる。

本稿では、NLP 研究分野では、学習データの量より、言語学知識体系の質のほうがもっと重要視されるべきだという観点が挙げられ、高品質の言語学知識体系が中国語単語分割のようなNLPタスクに大いに貢献できることが検証される。我々は、既存の中国語言語学知識体系 Chinese Sentence Structure Grammar(王,宮崎 2007; Wang et al. 2013)の単語品詞辞書と小型の品詞付け文コーパス（以下はCSSGと呼ぶ）をMecabに導入し、新しい中国語単語分割ツール（以下は「Cmecab」と呼ぶ）を実現する。さらに、Cmecabとある特許分野ですでに商用化された中国語単語分割ツール（以下は商用ツールと呼ぶ）を比較評価する。比較評価の結果によって、:1) 特許分野においては、Cmecab の現段階の正解率が商用ツールを超えることが検証される；2) 商用ツールの精度上の限界が見られる一方、Cmecabの精度が単語品詞辞書項目の追加により、さらに向上する余地があることが検証される。CSSGの単語品詞辞書に専門用語

とその品詞情報を登録することより、特許分野の中国語単語分割タスクの精度上の限界をブレークスルーする可能性が示される。中国語言語処理研究では、単語分割の基準が研究グループによって分かれているのが現状である。客観的且つ公平に比較評価を行うために、個別の基準に偏らない、中立的な評価基準が設計される。

## 2. CSSG について

中国語処理研究では、よく用いられている言語学知識体系として、Peking University Treebank (PKU)、Tsinghua University Treebank (TSU)、Penn Chinese Treebank (CTB)などが挙げられる(Yu et al. 2010)。この三つの知識体系とも Context-Free Phrase Structure Grammar (CF-PSG)という英語処理で成功した文法枠組みに基づいて、中国語文構造を解釈している。しかし、CF-PSG に基づく中国語構文解析精度が同じく CF-PSG に基づく英語構文解析精度より大幅に低く(Levy, Manning 2003)、実用レベルに及ばない。

| 見出し語 | 品詞 |
|---|---|
| 他（彼） | r |
| 把（を） | ba |
| 被（によって） | bei |
| 苹果（リンゴ） | n |
| 扔（捨てる） | V1aod |

表 1：CSSG 単語品詞辞書の内容の一部

中国語は孤立語であり、英語と違って、動詞の語尾変化などの表面上の構文制約が少ない。中国語のような孤立語の構造をうまく処理するために、表層制約(構文要素の位置制約)と深層制約(動詞と格要素の共起制約）の両方を取り扱う文法枠組み Sentence Structure Grammar (SSG) が提案された（王,宮崎 2007; Wang et al. 2013)。CSSG は前記の 3 つの知識体系 PKU、TSU、CTB と違い、SSG という文法枠組みに基づいて構築された中国語言語学知識体系の一部である。CSSG が独自な品詞体系を用いる単語品詞辞書(306,447 語;表 1)と小型品詞付け文コーパス(3,180 文；表 2)からなる。

PKU、TSU、CTB の場合では、設計者と作業者が別々であり、作業者が複数であった。それに対して、CSSG の設計者と作業者が同じ人である。そのため、CSSG のほうが知識体系としての一貫性がほかの 3 つの知識体系よりよいと考えられる。

```
他/r 朝/chao2 约翰/nm 扔/V1aod 苹果/n
他/r 把/ba 苹果/n 扔/V1aod 向/xiang4 约翰/nm
苹果/n 被/bei 扔/V1aod 向/xiang4 约翰/nm
苹果/n 被/bei 他/r 扔/V1aod 向/xiang4 约翰/nm
苹果/n 被/bei 他/r 扔/V1aod 给/to 约翰/nm
苹果/n 被/bei 扔/V1aod 给/to 约翰/nm
他/r 把/ba 苹果/n 扔/V1aod 回/back 筐/tn 里/f1
苹果/n 被/bei 扔/V1aod 进/into 筐/tn 里/f1
他/r 把/ba 苹果/n 扔/V1aod 在/at 桌/tn 上/f1
苹果/n 被/bei 他/r 扔/V1aod 在/at 桌/tn 上/f1
…
```

表 2：CSSG 品詞付け文コーパスの内容の一部

我々が、CSSG を言語に依存しない形態素解析エンジン Mecab (Mecab) に導入し、新しい中国語単語分割ツール Cmecab を作成した。

## 3. 評価

言語学知識体系 CSSG がどれほど中国語単語分割タスクに貢献しているのかを検証するために、特許の電気、機械、化学などの各分野を網羅する 120 文をテストセットとし、Cmecab と商用ツールを比較評価する。

### 3.1 評価基準

中国語言語処理研究では、単語分割の基準が研

究グループによって分かれている。客観かつ公平に比較評価を行うために、個別の基準に偏らない、中立的な評価基準を設計する必要がある。中国語単語分割の結果では、例１～7 の赤い部分に示すようなものがあり、どの基準に従ってもエラーだと判断するしかない。ここでは、このようなものだけをエラーとして数える。表 3 に示すように、分割結果が違っても、前記のエラーでない限り、どちらも正解だと判断される。

例１:反馈 到泵 转速 的 控制 中
　　ポンプ回転数の制御にフィードバックさせる
例２：再 经侧 吹风 冷却
　　さらに片側から風で冷却する
例３：特别 适合于 就 寝时 穿戴
　　特に就寝時の着用に適している
例４：带破碎 盖 铲斗
　　破砕蓋付バケット
例５：有机化　合物
　　有機化合物
例６：抑制 振荡 的 陷 波 电路
　　発振を抑圧するトラップ回路
例７：防 细 菌 滋生
　　細菌の繁殖を防止し

| 中国語 | 日本語訳 | 判定 |
|---|---|---|
| 支撑　杆 | 支持　バー | ○ |
| 支撑杆 | 支持バー | ○ |
| 薄　膜 | 薄　膜 | ○ |
| 薄膜 | 薄膜 | ○ |
| 适合 于 | ～に　適する | ○ |
| 适合于 | ～に適する | ○ |

表３：中国語単語分割の比較評価基準

### 3.2　評価結果と分析

テストセット文を商用ツールと Cmecab でそれぞれ単語分割し、前記の評価基準で比較評価を行った。その結果を表 4 に示す。Cmecab の正解率が 96.41%であり、商用ツールよりやや高い。

2 つのツールで生じたエラーを比べると、商用ツールのエラーが不規則であるのに対して、Cmecab のエラーの大多数は専門用語が分けられてしまうようなケースが多いことがわかる。例えば、例 8 を商用ツールと Cmecab で単語分割した結果が表 5 に示される。

| ツール名 | 分割された総単語数 | エラー数 | 正解率 |
|---|---|---|---|
| 商用ツール | 3,824 | 143 | 96.26% |
| Cmecab | 3,982 | 143 | 96.41% |

表４：中国語単語分割の比較評価結果

例 8：带有所述夹片的所述下管的一端插置于所述上管内，所述夹片卡固于所述半片止档部。
　　前記クリップを有する前記下パイプの一端が前記上パイプ内に挿設され，前記クリップが前記半片ストッパ部に係止される

| ツール | 単語分割結果 | エラー数 |
|---|---|---|
| 商用ツール | 带有 所述 夹片 的 所述 下管 的 一 端 插 置于 所述 上 管内 ，所述 夹片卡固于 所述 半片止 档部 | 7 |
| Cmecab | 带有 所述 夹 片 的 所述 下 管 的 一端 插置于 所述 上 管 内 ，所述 夹 片 卡固于 所述 半 片 止档 部 | 6 |

表５：商用ツールと Cmecab の単語分割結果

| 見出し語 | 品詞 |
|---|---|
| 夹片（クリップ） | n |
| 半片（半片） | n |

表６：追加された用語とその品詞

商用ツールの精度をさらにアップするのが難しいのに対して、Cmecab のほうが単語品詞辞書に用語とその品詞を登録することで、精度がさらに向上する傾向が見られる。例えば、「夹片/クリッ

プ」と「半片/半片」との２つの単語とその品詞（表6）を CSSG 単語品詞辞書に登録し、再び Cmecab に学習させると、例８のエラーの数が６から２まで減り、Cmecab の精度がさらに向上する余地が示された（表 7）。

| ツール | 単語分割結果 | エラー数 |
|---|---|---|
| 登録前 | 带有 所述 夹 片 的 所述 下 管 的 一端 插置于 所述 上 管 内，所述 夹 片 卡固于 所述 半 片 止档 部 | 6 |
| 登録後 | 带有 所述 夹片 的 所述 下 管 的 一端 插置于 所述 上 管 内，所述 夹片 卡固于 所述 半 片 止档 部 | 2 |

表 7：用語登録前後の Cmecab の単語分割結果

## ４．おわりに

統計機械学習手法は中国語単語分割のような NLP 研究分野では主流となっているが、精度上の限界が見られる。“data sparseness”が一般的な課題となり、学習データの量の重要性が強調されるが、学習データの量を増やすだけで、精度上の限界を打破しにくいというのが現状である。我々は、NLP に対しては、学習データの量より、言語学知識体系の質のほうがもっと重要視されるべきだと主張し、既存の中国語言語学知識体系 CSSG を言語に依存しない形態素解析エンジンに導入することで、高品質の言語学知識体系が中国語単語分割のような NLP タスクに大いに貢献することを検証できた。さらに、CSSG 単語品詞辞書に用語とその品詞情報を登録することより、特許分野の中国語単語分割タスクの精度上の限界をブレークスルーする可能性を示した。

**謝辞**

CSSG を Mecab に導入するにあたり、東京大学の範暁蓉氏に協力していただいたので、深く感謝を申し上げる。

**参考文献**

Sun X., Y. Zhang, T. Matsuzaki, Y. Tsuruoka (2009). A Discriminative Latent Variable Chinese Segmenter with Hybrid word/character Information. In the proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 56-64. Association for Computational Linguistics, Boulder, Colorado.

Gao J., M. Li, A. Wu, CN. Huang (2005). Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. In the proceedings of Computational Linguistics.

Xue N. (2003). Chinese Word Segmentation as Character Tagging. International Journal of Computational.Linguistics and Chinese Language Processing, 8(1).

MeCab: Yet Another Part-of-Speech and Morphological Analyzer. “Web Page.” http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

Wang Xiangli, Yi Zhang, Yusuke Miyao, Takuya Matsuzaki, Junichi Tsujii (2013). Deep Context-free Grammar for Chinese with Broad Coverage. In the Proceedings of SIGHAN-7.

Yu, S. et al. (2002). The Basic Processing of Contemporary Chinese Corpus at Peking university specification. Journal of Chinese Information Processing, 16 (5).

Zhou, Q. (2004). Annotation Scheme for Chinese Treebank. Journal of Chinese Information Processing, 18(4).

Xue Nanwen, Xia Fei. (2000) The Bracketing Guidelines for the Penn Chinese Treebank.” Tecknical Report. University of Pennsylvania.

Yu Kun, Yusuke Miyao, Takuya Matsuzaki, Xiangli Wang, Yaozhong Zhang, Kiyotaka Uchimoto, Junichi Tsujii (2010). Comparison of Chinese Treebanks for Corpus-oriented HPSG Grammar Development. Journal of Natural Language Processing (Special Issue on Empirical Methods for Asian Language Processing).

王向莉，宮崎正弘(2007).文構造文法に基づく中国語構文解析．言語処理論文誌．vol.14，No.2.

Roger Levy, Christopher Manning (2003). Is it Harder to Parse Chinese, or the Chinese Treebank? Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.