A Chinese Dependency Parser(CNP) 用中国語解析モデル Version 1

● 概要

このデータベースは、オープンソースソフトウェアとして配布されている係り受け解析器（A Chinese Dependency Parser）（CNP）Version 1 のための中国語解析用モデルパラメータです。Language Data Consortium （LDC）[http://www.ldc.upenn.edu]より配布されているChinese Treebank 4.0 （CTB 4.0）[Catalog No. LDC2004T05], Chinese Treebank 5.0 (CTB 5.0) [Catalog No. LDC2005T01], Chinese Treebank 6.0 （CTB 6.0）[Catalog No. LDC2007T36]および Chinese Gigaword Second Edition [Catalog No. LDC2005T14]を用いて学習した、GBK/UTF-8 文字コード用のモデルを提供します。本モデルをCNPと共に用いることにより、高精度な中国語係り受け解析が可能になります[1]。本データベースはCRF++を用いた簡易な中国語形態素解析モデル（GBKエンコーディング用）を含んでおりますので、本データベースのみで、中国語の生の文を入力とした係り受け解析システムを構築することが可能です。なお、(独）情報通信研究機構はLDCのfor-profitライセンスを取得しており、本データベースを配布する権利を有しております。

使用方法につきましては、配布ファイルを展開した結果生成されるディレクトリ中の README ファイルをご覧下さい。

CNP の入手方法については、http://nlpwww.nict.go.jp/cnp をご覧下さい。

● ファイル

- CNP_ALAGIN_CHINESE_MODEL_V1_README.pdf（本文書）
- MSTModels, res, SegPOSCRF（モデルファイルが格納されたディレクトリです）
- README（使用方法）
- MSTModels/README（各モデルの詳細）

[1] 参考文献に挙げた研究では使用したテストデータに対して世界最高性能を達成しました。

これらのモデルを用いてCNPを使用するには、モデルによって異なりますが7GB程度の空きメモリが必要となります。解析速度はマシンにもよりますが、1 文/sec程度です。

なお、今回配布するモデルの解析精度は以下のようになっております（いずれも、付属の形態素解析モデルを使用し、形態素解析から係り受け解析まですべて自動で行った場合の精度です。精度の測定は学習に用いたコーパスのテストデータ部分で行っております。CTB5、6モデルの精度がCTB4モデルよりも数値上低くなっているのはそのためです。用途に合わせて、精度が良いと思われるモデルをご使用下さい)。

モデル	UAS(ラベルなし係り先精度)	LAS(ラベルあり係り先精度)	文全体正解率
CTB4モデル	0.865	0.834	0.434
CTB5モデル	0.829	0.787	0.294
CTB6モデル	0.829	0.791	0.271

● 利用条件

本データベースの利用には、(独) 情報通信研究機構と利用許諾契約を結ぶ必要があります。詳しくは、http://www.alagin.jpをご覧ください。

● 利用に関する注意

本データベースは、インターネットホームページ等、(独) 情報通信研究機構以外の第三者が作成した文書等のデータから、語彙の抽出及び統計処理等によって作成されたものです。そのため本データベースの内容は、(独) 情報通信研究機構の主体的な意思決定・判断を示すものではありません。本データベースの生成は、電子的な方法又は一様の選別基準による機械的判定によって行われています。そのため本データベースの内容の正確性、真実性及び相当性は一切保証されません。以上の理由により、(独) 情報通信研究機構は、本データベースの内容について、責任を負いかねます。本データベースの使用に関連して生

ずる損失、損害等についても、一切責任を負いかねます。本データベースには、意図せず、第三者への誹謗中傷、差別用語、個人情報などが含まれている場合があります。本データベースを利用の際はこれらによる権利侵害に十分な注意をお願いいたします。利用者においては、本データベースの以上の特質をよくご理解の上で、本データベースをご利用下さい。

● 参考文献

“Improving Dependency Parsing with Subtrees from auto-Parsed Data”, Wenliang Chen, Jun’ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa, EMNLP 2009, 2009 年.

● 本データベースに関する問い合わせ先

独立行政法人情報通信研究機構
知識創成コミュニケーション研究センター
MASTAR プロジェクト 言語基盤グループ
Email: alagin-lr@khn.nict.go.jp