BERT/RoBERTa/DeBERTaモデルによる多言語係り受け解析

京都大学人文科学研究所附属東アジア人文情報学研究センター
安岡孝一

われわれの（究極の）目標

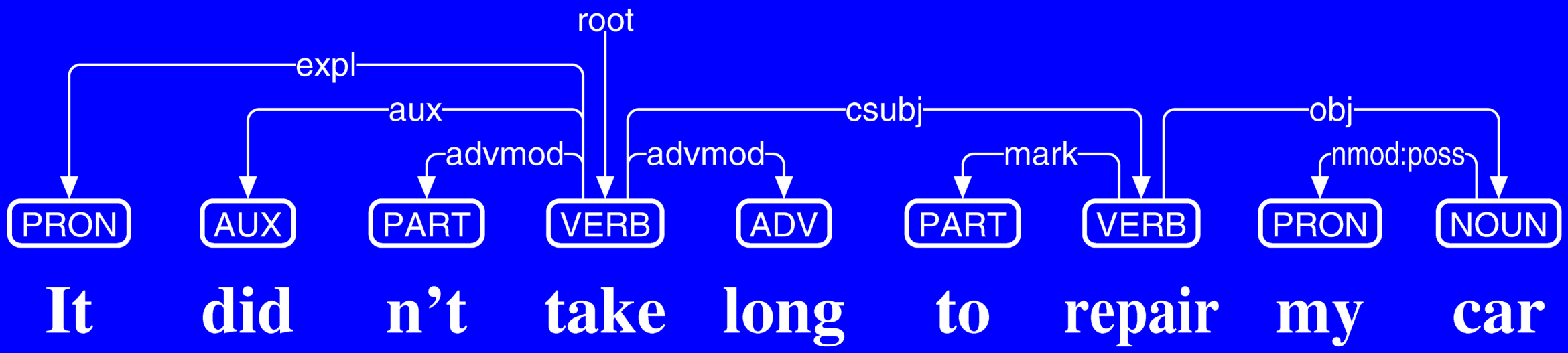
- 書写言語の「構造」を知りたい
  - 多言語にまたがる「構造」は存在するのか？
  - それは機械（コンピュータ）が解けるような「構造」なのか？
- 現時点ではUniversal Dependencies (UD)が筋がよさそう
  - ただし各言語への適用は、それぞれに問題あり

Universal Dependencies (UD)

- 141言語の係り受けコーパスを公開
  - Игорь Мельчукの依存文法(Dependency Grammar)を拡張
  - 全ての係り受けを単語間の有向グラフで表現
  - 各単語に入るリンクは1本に限定
    - 動詞から項（主語・目的語）や斜格補語へリンク
    - 被修飾語から修飾語へリンク（前置詞は修飾語の一種とみなす）
    - リンクのループは許さない

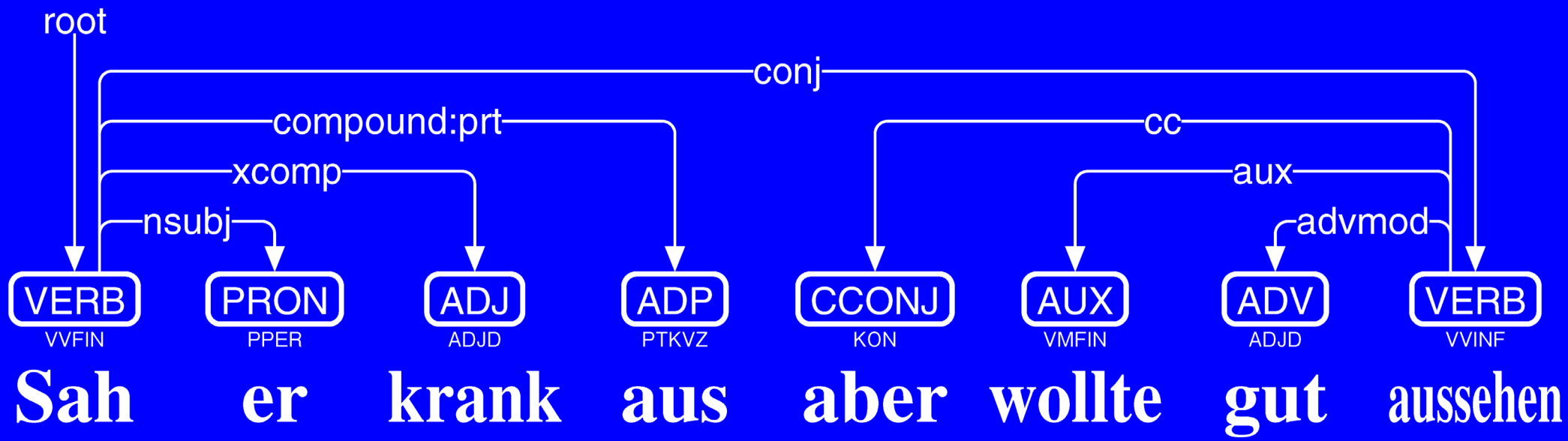
Universal Dependencies (UD)

- 141言語の係り受けコーパスを公開
  - 英語UD



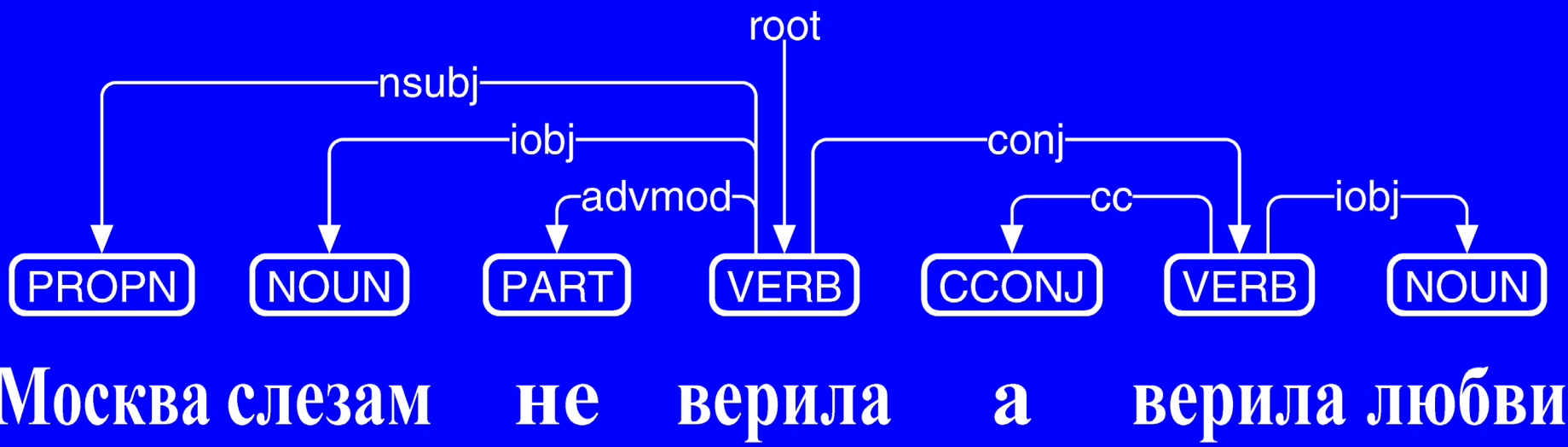
Universal Dependencies (UD)

- 141言語の係り受けコーパスを公開
  - ドイツ語UD



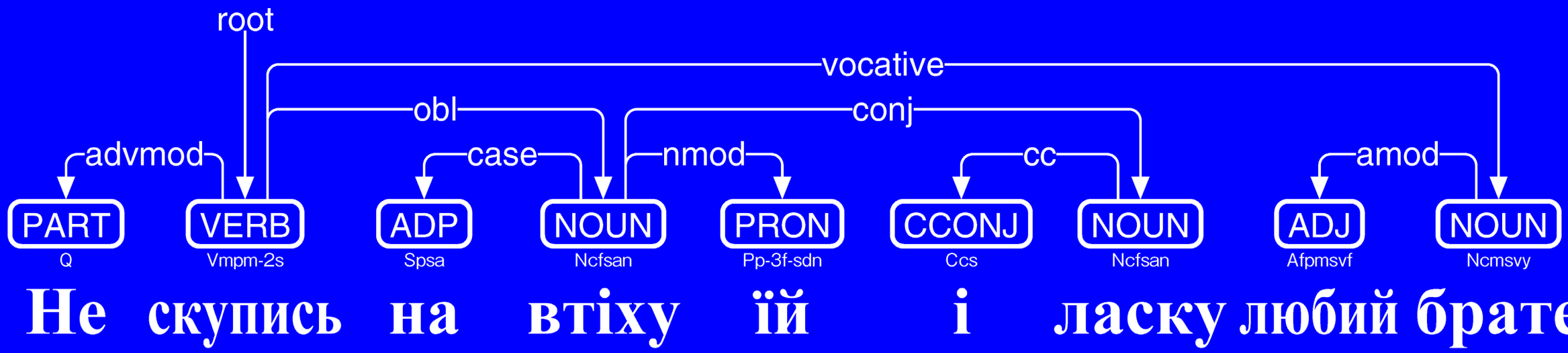
Universal Dependencies (UD)

- 141言語の係り受けコーパスを公開
  - ロシア語UD



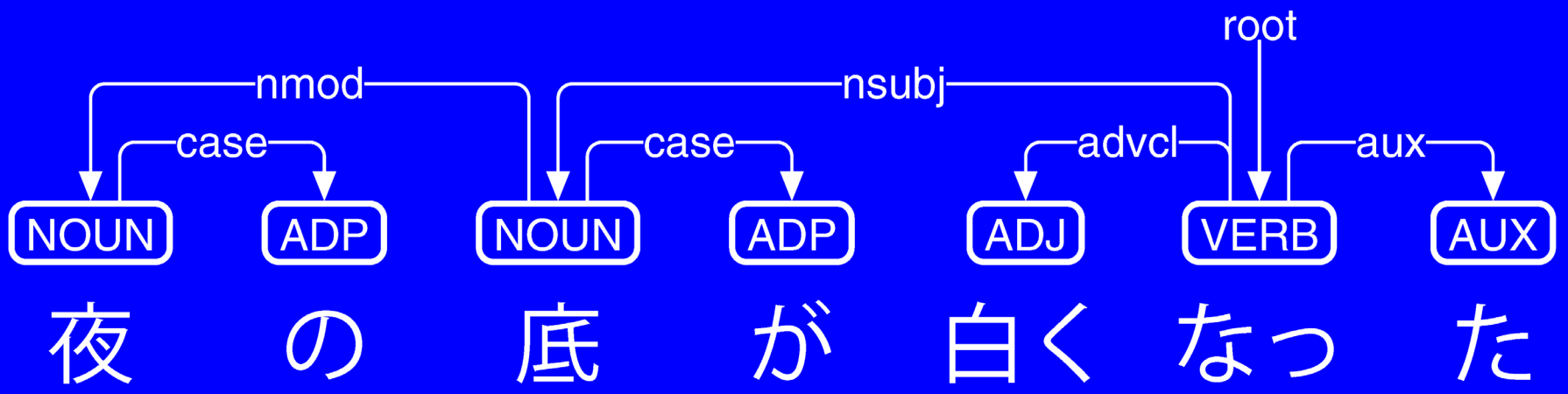
Universal Dependencies (UD)

- 141言語の係り受けコーパスを公開
  - ウクライナ語UD



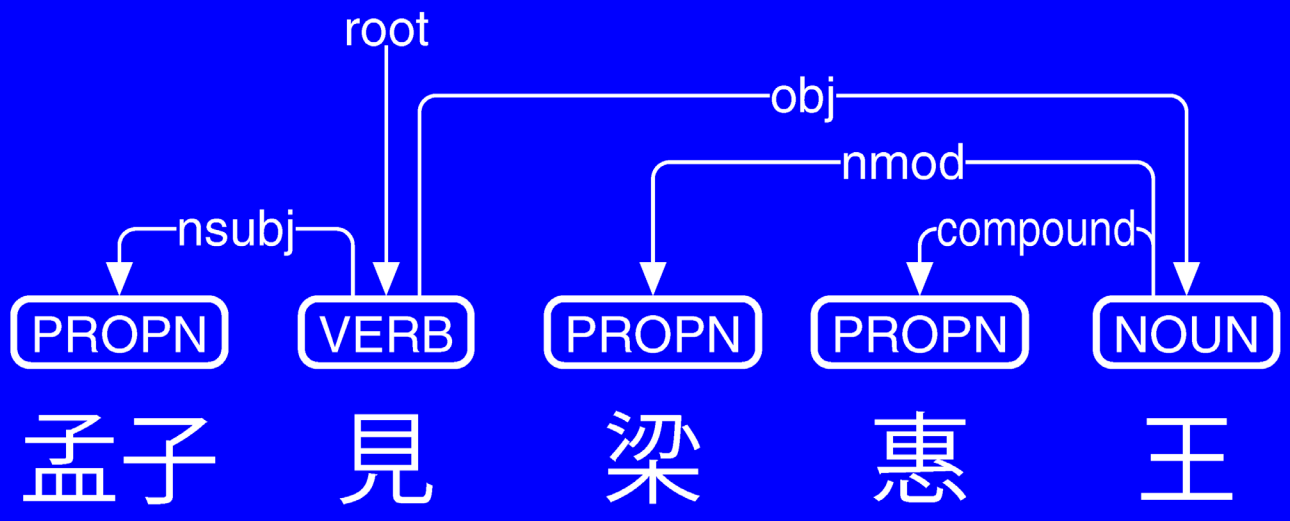
Universal Dependencies (UD)

- 141言語の係り受けコーパスを公開
  - 日本語UD



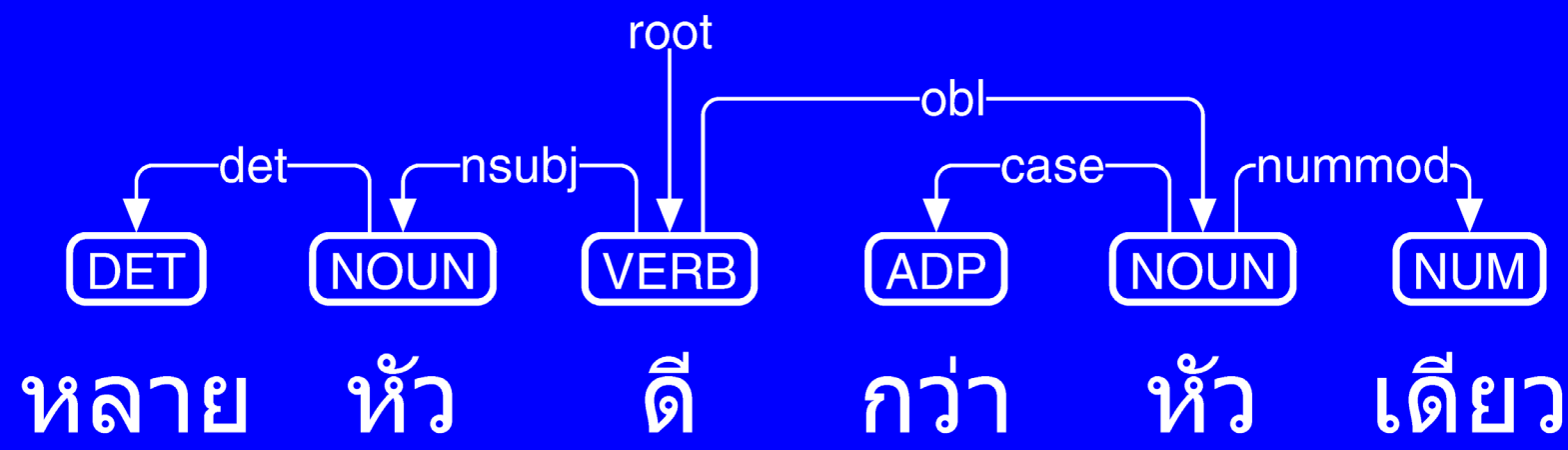
Universal Dependencies (UD)

- 141言語の係り受けコーパスを公開
  - 古典中国語（漢文）UD



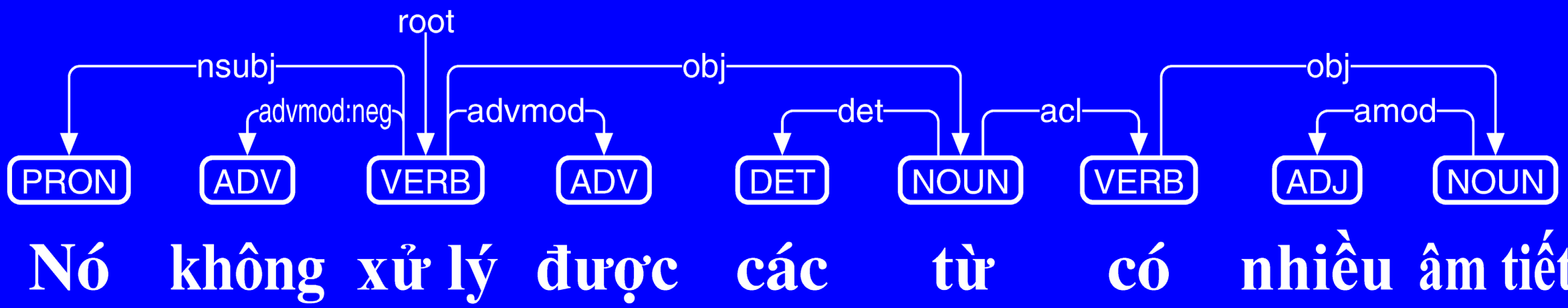
Universal Dependencies (UD)

- 141言語の係り受けコーパスを公開
  - タイ語UD



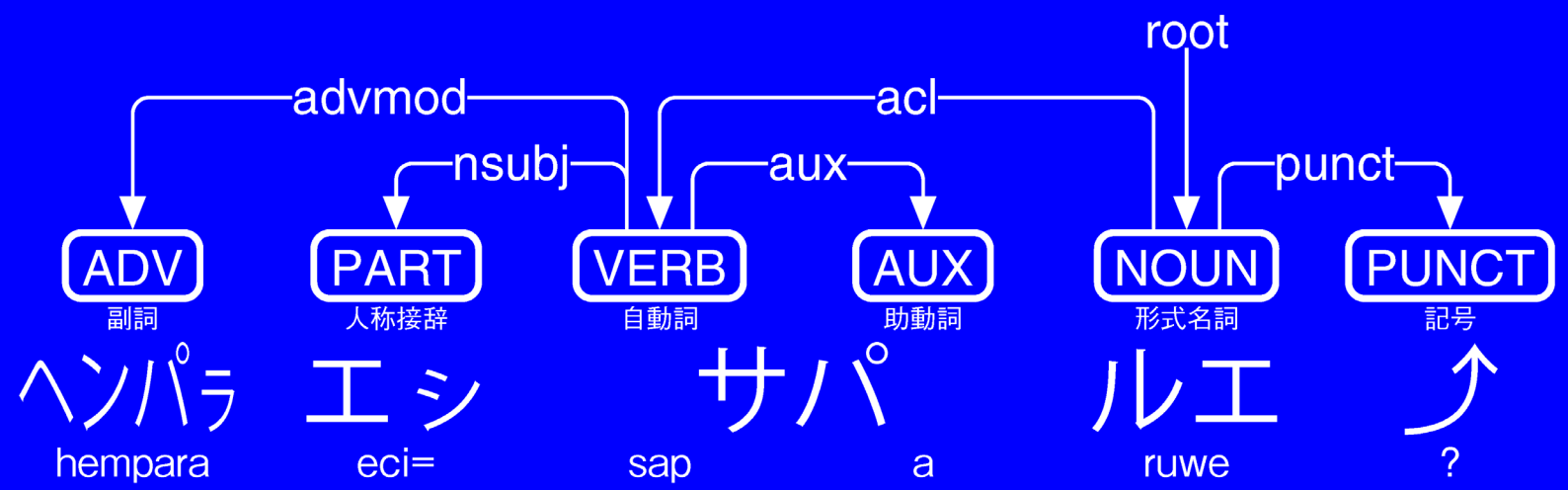
Universal Dependencies (UD)

- 141言語の係り受けコーパスを公開
  - ベトナム語UD



Universal Dependencies (UD)

- 141言語の係り受けコーパスを公開
  - アイヌ語UD（上川アイヌ語、公開準備中）



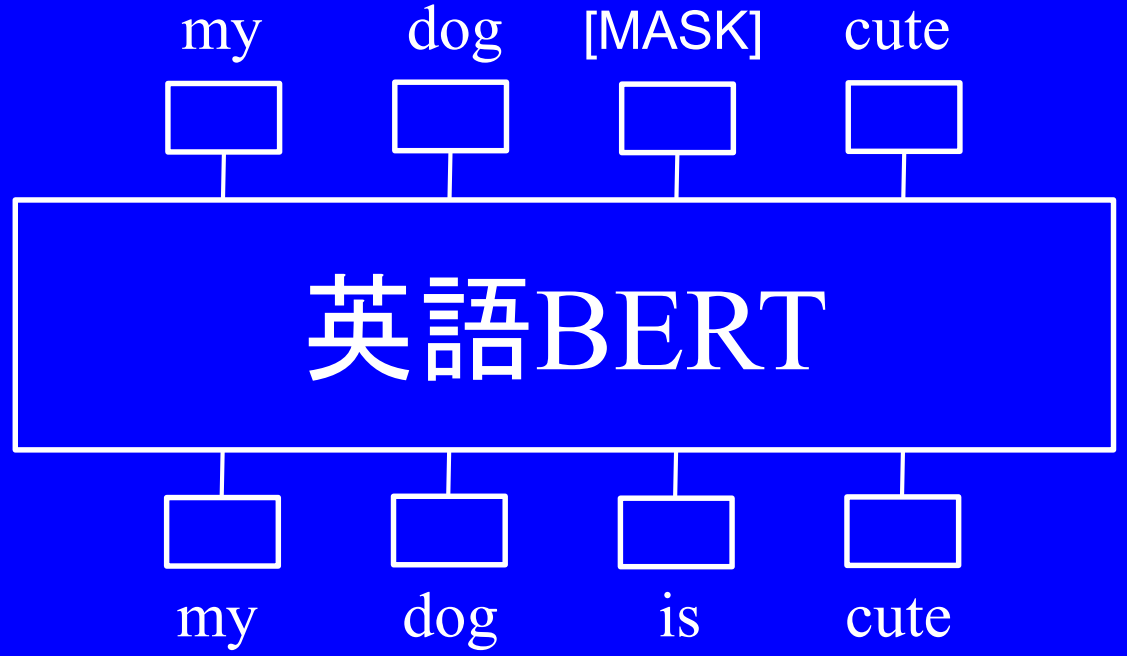
UDを用いた形態素解析と係り受け解析

- BERT/RoBERTa/DeBERTaモデルを製作
  - 穴埋めゲームを用いて大量の文章を丸暗記
  - 系列ラベリングを用いて形態素解析向けチューニング
  - 系列ラベリングを拡張して係り受け解析向けチューニング

言語モデル製作にGPU (A100)を長時間使用

英語BERTの穴埋めゲーム

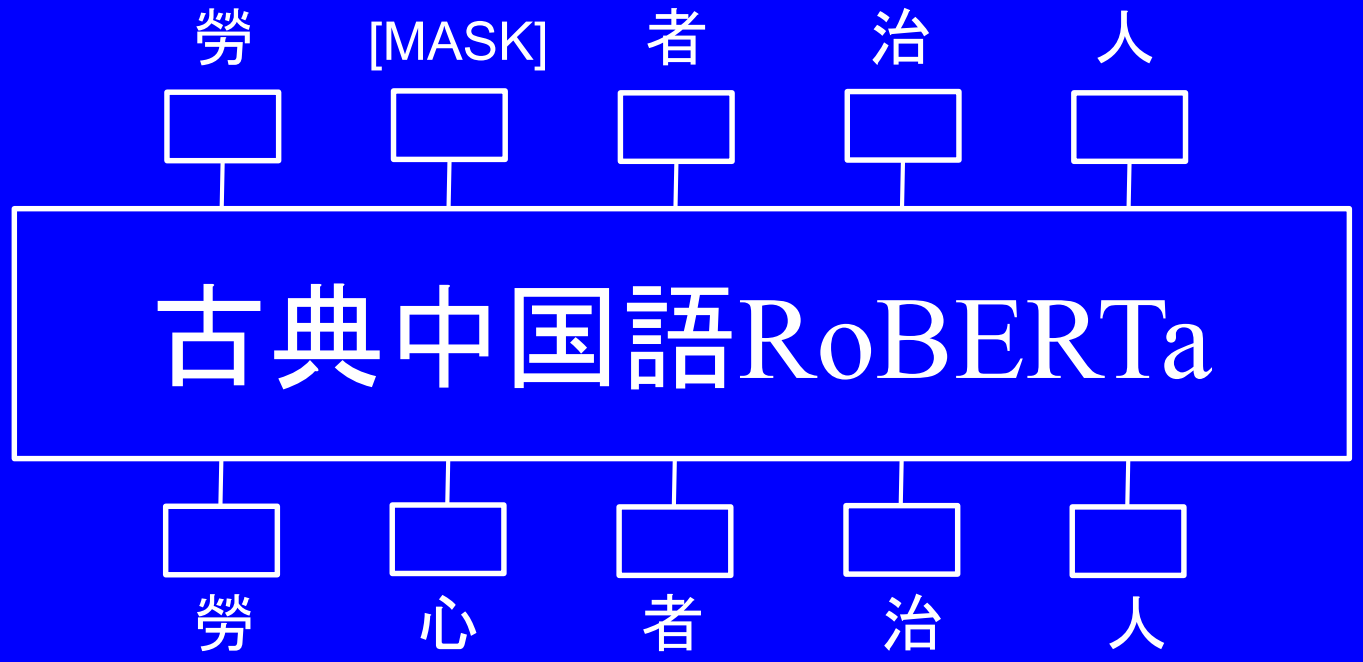
- 文中の単語15%を[MASK]にして学習



内部に言語空間ができる

古典中国語RoBERTaの穴埋めゲーム

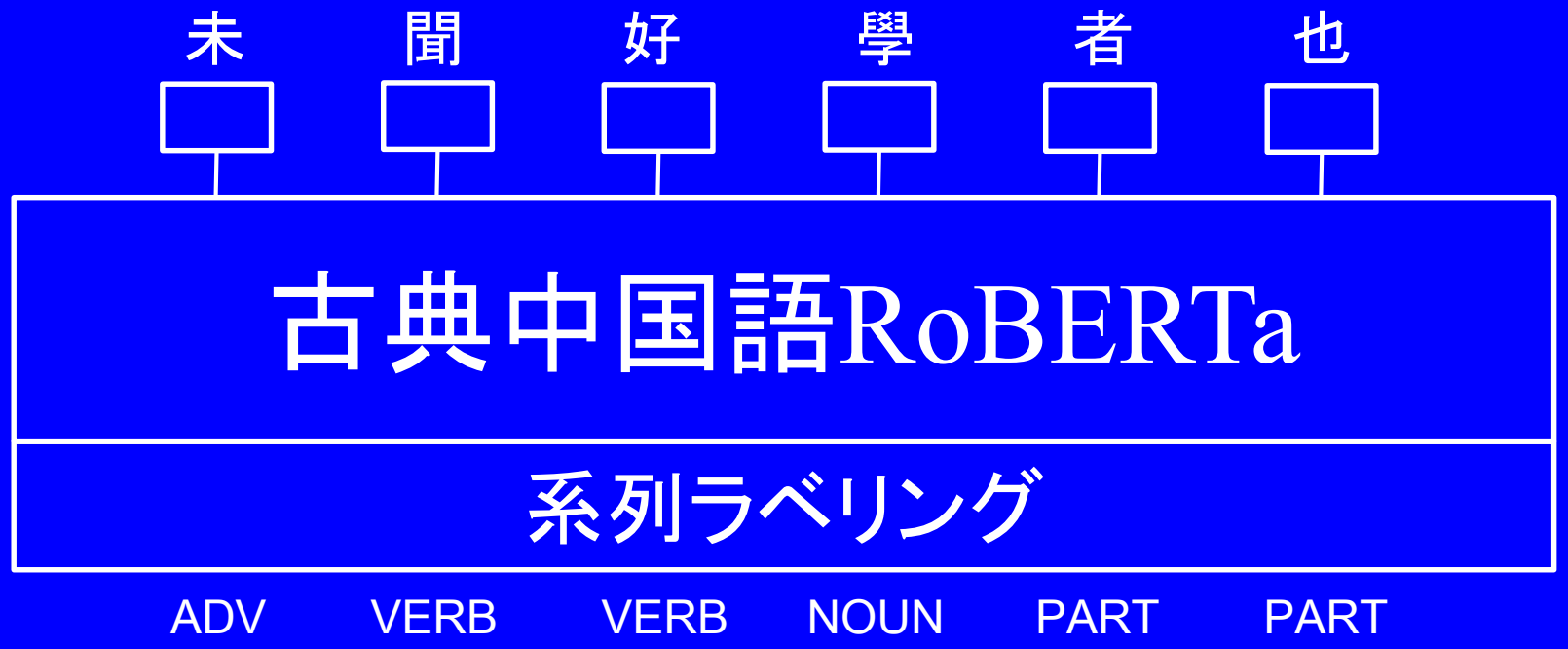
- 文中の漢字15%を[MASK]にして学習



四庫全書を中心に17億字を学習

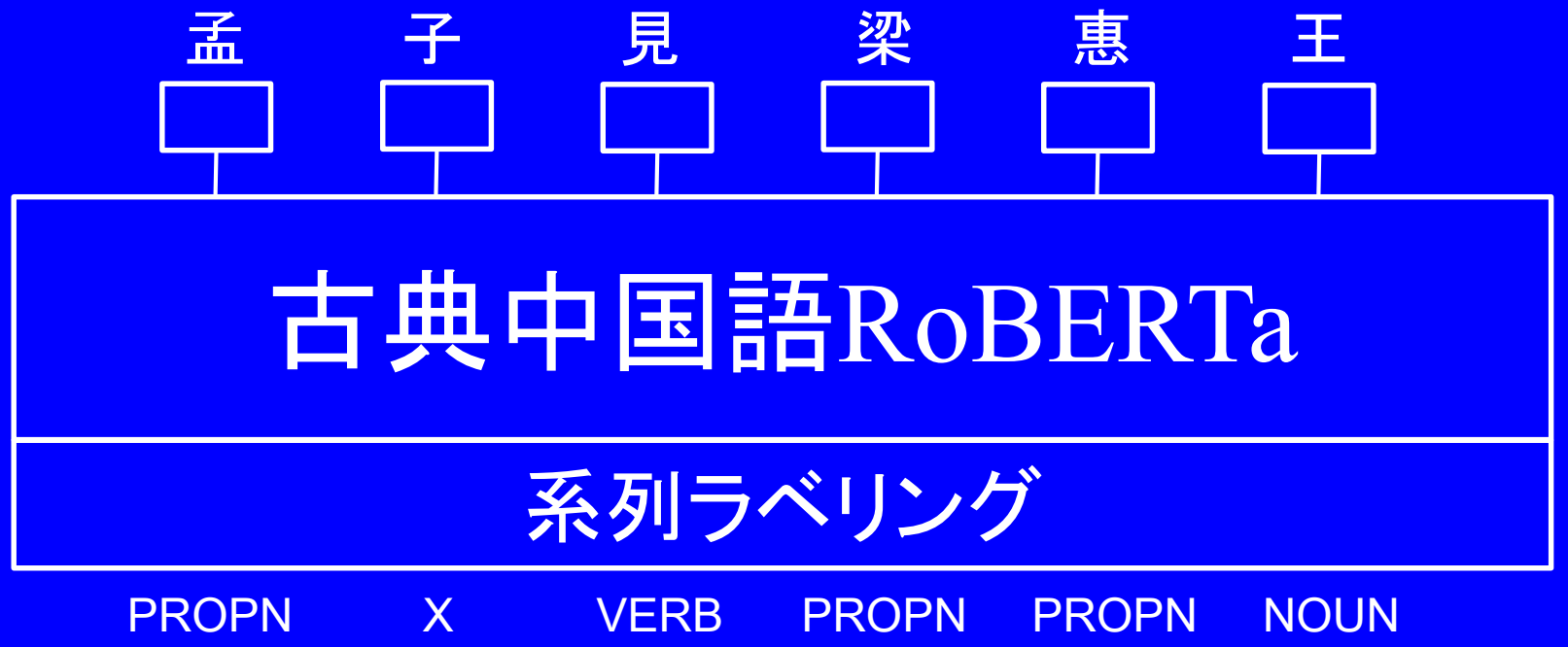
系列ラベリングによる品詞付与・単語切り

- BERT/RoBERTa/DeBERTaの出力をファインチューニング



系列ラベリングによる品詞付与・単語切り

- BERT/RoBERTa/DeBERTaの出力をファインチューニング
  - 複数トークンにまたがる単語は品詞に「X」を付与



系列ラベリングの拡張による係り受け解析

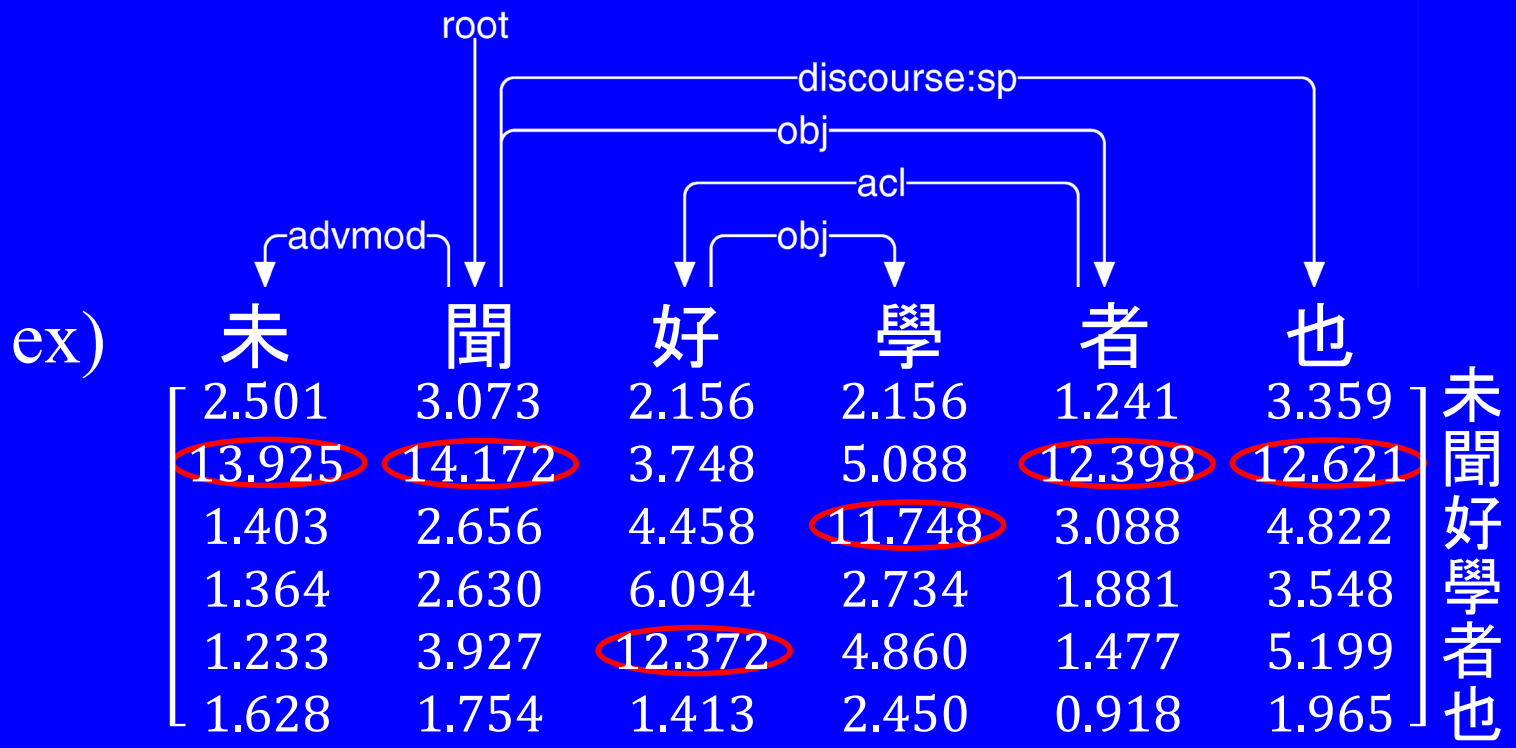
- 有向グラフの隣接行列を系列ラベリングで導出
  - 有向枝の重みにlogits（対数オッズ）を使用

ex)

	未	聞	好	學	者	也	
	2.501	3.073	2.156	2.156	1.241	3.359	未
	13.925	14.172	3.748	5.088	12.398	12.621	聞
	1.403	2.656	4.458	11.748	3.088	4.822	好
	1.364	2.630	6.094	2.734	1.881	3.548	學
	1.233	3.927	12.372	4.860	1.477	5.199	者
	1.628	1.754	1.413	2.450	0.918	1.965	也

  - 対角の最大値・各列の最大値を抽出
    - ループ発生時にはChu-Liu-Edmonds法を適用

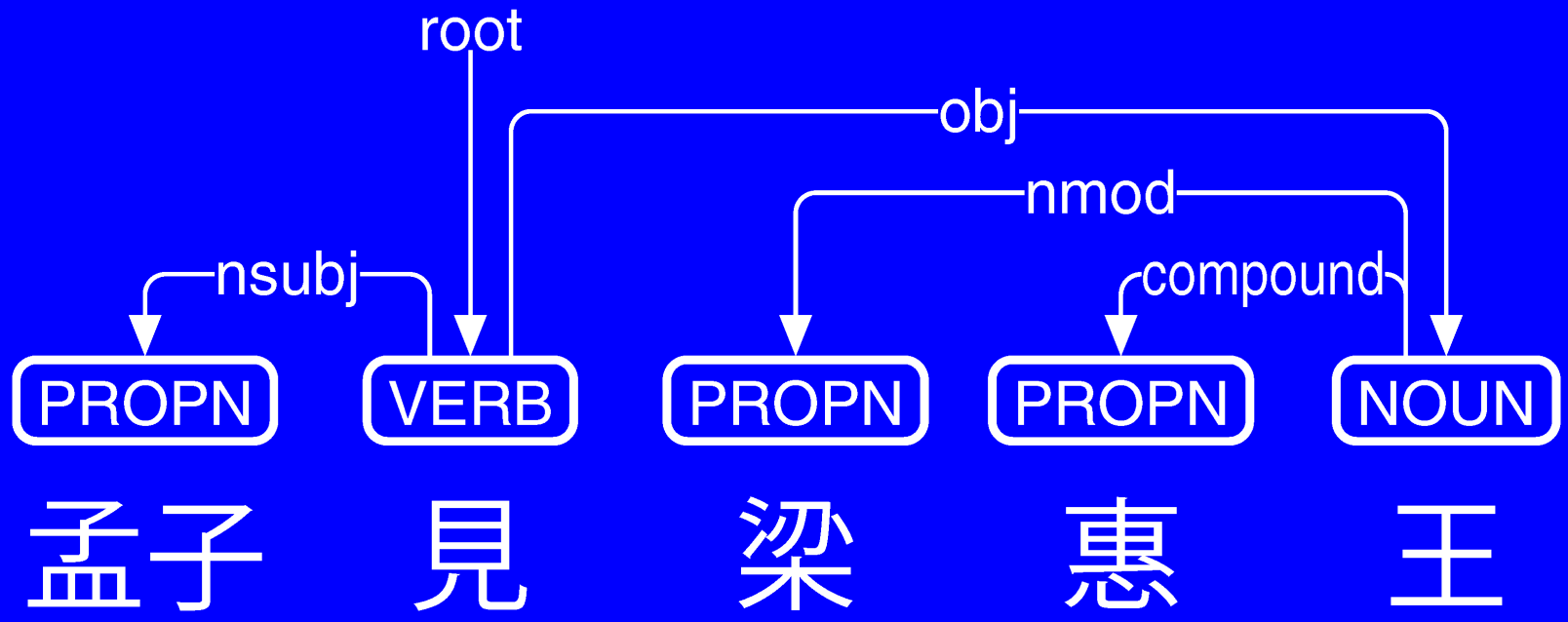
系列ラベリングの拡張による係り受け解析



- 対角の最大値・各列の最大値を抽出
  - ループ発生時にはChu-Liu-Edmonds法を適用

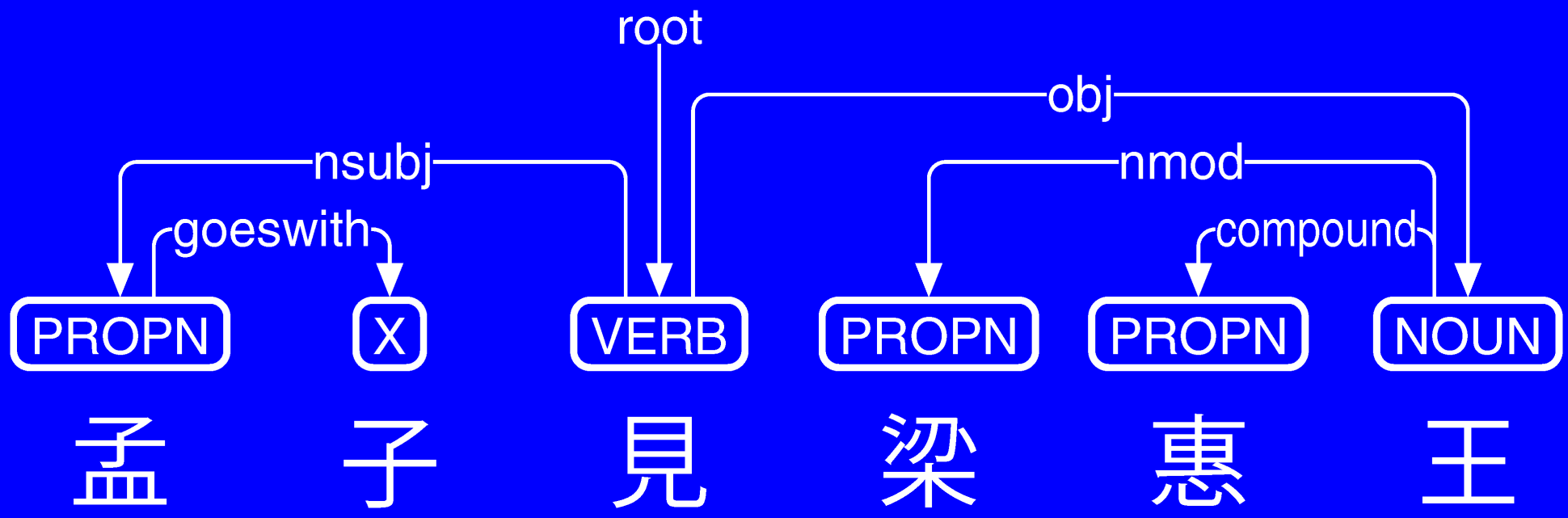
単語間に区切りのない言語の難しさ

- BERT/RoBERTa/DeBERTaのトークンと単語長の不一致
  - 古典中国語RoBERTaは漢字1文字が1トークン
  - 古典中国語（漢文）には2文字以上の単語もある



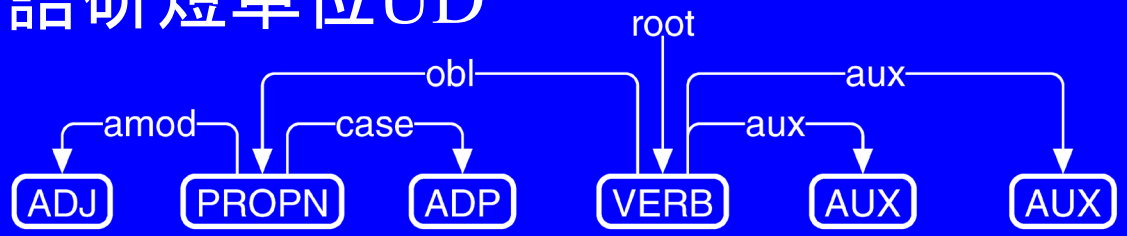
単語間に区切りのない言語の難しさ

- BERT/RoBERTa/DeBERTaのトークンと単語長の不一致
  - goeswith（泣き別れ）リンクと品詞「X」で全単語をトークンに分解

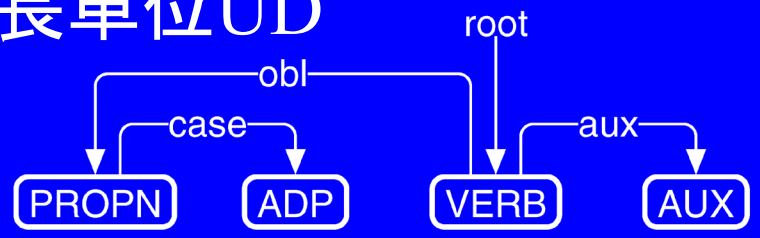


単語に区切りのない言語の難しさ

- 日本語では単語長に合意がない
  - 国語研短単位UD



  - 国語研長単位UD



単語に区切りのない言語の難しさ

- 日本語では単語長に合意がない
  - 日本語DeBERTaにおけるトークン長はどう設計すべきか
    - トークンが長すぎると系列ラベリング不可能に
    - トークンが短すぎると解析精度が上がらない
  - 文字の途中でトークンを「切る」ことも試してみるべきか

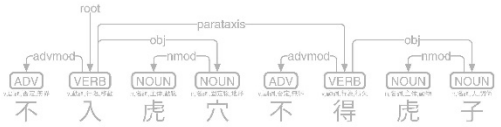
    ex) 「試ス」「ィ」「て」「みル」「ゥ」「べき」「か」
- タイ語やアイヌ語はさらに難しそう

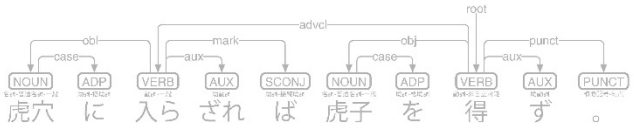
今後の研究の進展に乞うご期待

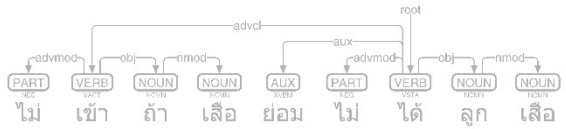
Universal Dependencies と BERT/RoBERTa/DeBERTa モデルによる

多言語情報処理

安岡孝一







2023年8月版

まとめ

- BERT/RoBERTa/DeBERTaによる形態素・係り受け解析
  - 穴埋めゲームを用いて大量の文章を丸暗記
  - 系列ラベリングを用いて形態素解析向けチューニング
  - 系列ラベリングを拡張して係り受け解析向けチューニング
- 単語間に区切りのない言語ではトークン長の設計が難しい