



自然言語処理と記述言語研究のインターフェイス

Universal Dependencies のすすめ

田口 智大 (Chihiro Taguchi)

Department of Computer Science and Engineering, University of Notre Dame



KELLOGG INSTITUTE
FOR INTERNATIONAL STUDIES

A LEGACY OF EXCELLENCE SINCE 1982



UNIVERSITY OF
NOTRE DAME

自然言語処理と記述言語学にまたがる研究の紹介(言語学者向け)

1. 自己紹介
2. 言語学と自然言語処理
3. Universal Dependencies (UD)
4. タタール語UDの紹介
5. 今後の展望

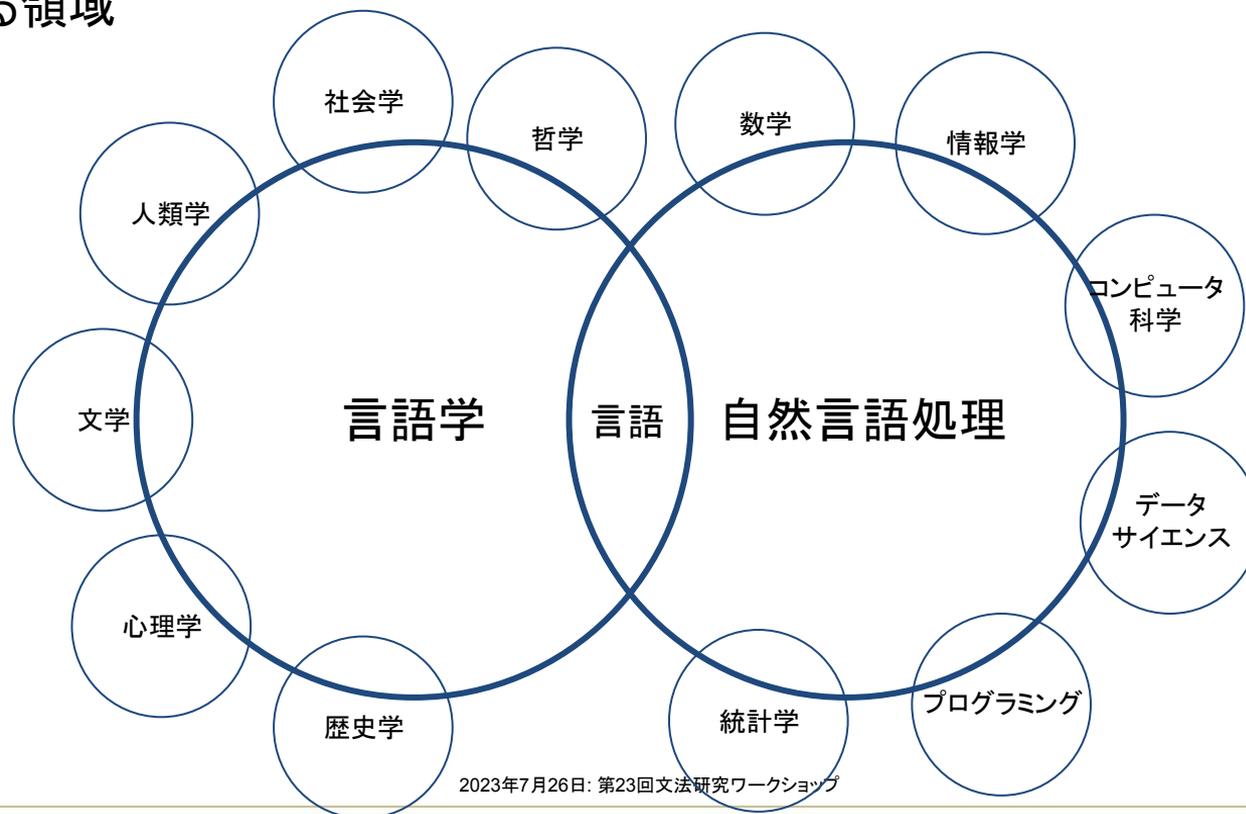
田口 智大 (TAGUCHI Chihiro)

2015年4月 – 2019年9月	慶應義塾大学法学部政治学科
2018年9月 – 2019年6月	SOAS, University of London 留学
2019年8月 – 2019年9月	東京外国語大学AA研夏期言語研修 シンポー語
空白？	
2020年10月 – 2022年9月	奈良先端科学技術大学院大学修士課程 自然言語処理学研究室
2021年9月 – 2022年8月	University of Edinburgh, MSc by Research in Linguistics
2022年8月 – 現在	University of Notre Dame, PhD in Computer Science and Engineering

現在の研究テーマ：言語記述のための自然言語処理技術の研究

- 例えば、少ないアノテーション済み ELANデータから自動転写モデルを機械学習で訓練

似て非なる領域



どちらも言語を対象にしているが...

❖ 決定的な方向性の違い: **science** 志向か、**engineering** 志向か

Science としての言語学	Engineering としての自然言語処理
<p>人間独自の能力としての言語の仕組みの探究</p> <ul style="list-style-type: none">● 普遍文法 Generative grammar● 認知言語学 Cognitive linguistics● 心理言語学 Psycholinguistics● 言語獲得 Language acquisition● 記述言語学 Descriptive linguistics <p>人間のツールとしての言語の仕組みの探究</p> <ul style="list-style-type: none">● 社会言語学 Sociolinguistics● 計算言語学 Computational linguistics	<p>言語に関して人間の役に立つ技術の開発</p> <ul style="list-style-type: none">● 機械翻訳 Machine translation Google Translate, DeepL● 音声認識 Speech recognition Siri, Alexa, YouTube● 文法誤り訂正 Grammatical error correction Grammarly● チャットボット、検索エンジン ChatGPT <p>計算機による人間言語のモデリング</p>

Engineering としての自然言語処理の方向性はここ数年で一気に明確化

- 十年前: 統計的手法が主流
 - 統計的機械翻訳 statistical machine translation (SMT)
 - 構文解析 parsing → 統語論 syntax と相性が良い
 - (日本語) 形態素分析 morphological segmentation
 - コーパスを用いた研究・開発 → コーパス言語学 corpus linguistics と相性が良い
- 現在: 深層学習 deep learning を用いた手法が主流
 - [Vaswani et al. \(2017\) “Attention is all you need”](#): Transformer の台頭
 - 半導体技術の進展により、大規模な行列演算が可能
 - ニューラル機械翻訳 neural machine translation (NMT)
 - **科学的な説明性 explainability に欠けるが、技術としてはあまりにも便利で強力**
 - 句構造規則 phrase structure rules などを用いた技術は衰退 (?)

学会文化の違い(印象)

	言語学	自然言語処理
全体的な雰囲気	文系寄り	理系寄り(特に工学)
共著・単著	単著が主流	共著が主流
パブリケーション	じっくり時間をかけて書く	たくさん実験してたくさん書く(多産多死?)
業績	論文執筆が主	トップ国際学会での発表も業績
引用	30~40年前の論文も頻繁に引用	10年以内の研究の引用が多い
設置学部	文系(言語学、文献学、文学)	理系(工学、情報学)
論文の構成	例を用いながら理論構築に寄与(分野による)	コードを書いて実験、報告

言語学と自然言語処理の溝は広まっていくばかり……？

- あくまで全体的なトレンドの話
 - 英語などの「大きな言語」中心
 - ビジネス、有用性
- 言語学の科学的な重要性は不変
- 記述言語学ができること
 - 自然言語処理においてデータは不可欠
 - 世界の99%以上の言語はデータが不足
 - (機械に可読な形で)言語を記述することが重要

本発表では、(記述)言語学と自然言語処理にまたがる取り組みの一つとして、
Universal Dependencies (UD) を紹介する

Universal Dependencies (UD) とは

Universal Dependencies



どの言語に対しても統一された形態統語的アノテーションをするプロジェクト

- 2015年に version 1.0 公開
- 今もコミュニティが拡大中
- 年2回更新(春・秋)
- Github上で管理
- 2023年7月現在、141言語・245個のツリーバンク(コーパス)
- 年に一度、Universal Dependencies Workshop (UDW) を開催

コンピュータで処理しやすい形式(タブ区切り形式 .tsv)

UDの形式: CoNLL-U

語 (syntactic word) 単位のアノテーション

一文中の各語に対して、

- ID: 語ID
- FORM: 語
- LEMMA: レンマ (辞書形、原形、語幹)
- UPOS: 品詞
- FEATS: 形態的素性
- HEAD: 修飾先 (ヘッド)
- DEPREL: 修飾関係

をアノテーション

UDの形式: 例

UDドイツ語 (PUDツリーバンク) より

```
412 # newdoc id = n01006
413 # sent_id = n01006011
414 # text = Ein Zeuge berichtete der Polizei, dass das Opfer den Verdächtigen im April angegriffen hatte.
415 # text_en = A witness told police that the victim had attacked the suspect in April.
416 1      Ein      ein      DET      DT      Case=Nom|Definite=Ind|Gender=Masc|Number=Sing|NumType=Card|PronType=Art 2      det      _
417 2      Zeuge     Zeuge     NOUN     NN      Case=Nom|Gender=Masc|Number=Sing      3      nsubj     _
418 3      berichtete berichtete VERB     VBC     Mood=Ind|Number=Sing|Person=3|Tense=Past      0      root      _
419 4      der      der      DET      DT      Case=Dat|Definite=Def|Gender=Fem|Number=Sing|PronType=Art      5      det      _
420 5      Polizei  Polizei  NOUN     NN      Case=Dat|Gender=Fem|Number=Sing      3      obl:arg    _      SpaceAfter=No
421 6      ,        ,        PUNCT    ,      _      15      punct     _
422 7      dass     dass     SCONJ    CC      _      15      mark      _
423 8      das      der      DET      DT      Case=Nom|Definite=Def|Gender=Neut|Number=Sing|PronType=Art      9      det      _
424 9      Opfer    Opfer    NOUN     NN      Case=Nom|Gender=Neut|Number=Sing      15     nsubj     _
425 10     den      der      DET      DT      Case=Acc|Definite=Def|Gender=Masc|Number=Sing|PronType=Art      11     det      _
426 11     Verdächtigen Verdächtige NOUN     NN      Case=Acc|Gender=Masc|Number=Sing      15     obj      _
427 12-13   im      _      _      _      _      _      _      _
428 12     in      in      ADP      APPR    _      14      case      _
429 13     dem     der      DET      ART     Case=Dat|Definite=Def|Gender=Masc|Number=Sing|PronType=Art      14     det      _
430 14     April   April   NOUN     NN      Case=Dat|Gender=Masc|Number=Sing      15     obl      _
431 15     angegriffen angreifen VERB     VBN     Tense=Past      3      ccomp     _
432 16     hatte   haben   AUX      VBC     Mood=Ind|Number=Sing|Person=3|Tense=Past      15     aux      _      SpaceAfter=No
433 17     .        .        PUNCT    .      _      3      punct     _
```

Tokenization and Word Segmentation

FORM: 表層形を記述

- 統語的語 syntactic word を基準とする
 - 正書法上で区切られた語や音韻的な語ではない
 - 例: いわゆる接語 clitic は別の語として扱う
 - `dámelo` (“Give it to me”; スペイン語)
`dá=me=lo`
`give.IMP=OBJ.1SG=OBJ.3SG.M`
→3語として分析。他にもフランス語の `au < à + le` やドイツ語の `zum < zu + dem` など

LEMMA: 辞書形、原形を記述

- 言語によって方針は異なりうる
 - 英語の動詞: 辞書形、ドイツ語の動詞: 不定形、日本語の動詞: ウ段形、
テュルク諸語の動詞: 語幹

品詞を記述

開いた類 open class	閉じた類 closed class	その他
ADJ 形容詞	ADP 側置詞	PUNCT 句読点など
ADV 副詞	AUX 助動詞	SYM 記号など
INTJ 感嘆詞	CCONJ 等位接続詞	X その他
NOUN 名詞	DET 限定詞	
PROPN 固有名詞	NUM 数詞	
VERB 動詞	PART 不変化詞	
	PRON 代名詞	
	SCONJ 従属接続詞	

形態論情報を記述

- 形態素境界は示さない(屈折語を考慮)
- 形態素の順番も考慮しない
- FORM(語)に含まれる形態的素性を set として記述
 - 慣習的に素性をアルファベット順に記述

例: артырдым (タタール語: “I increased (it)”)

art-tir-di-m

increase-CAUS-PST-1SG

→ FEATSは **Number=Sing | Person=1 | Tense=Past | VerbForm=Fin | Voice=Cau**

素性一覧

統語的依存関係の主要部のIDを記載

- 依存先がない場合 0

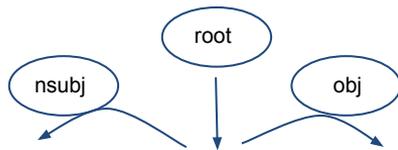
例: Clinton defeated Dole.



ID	FORM	...	HEAD
1	Clinton		2
2	defeated		0
3	Dole		2
4	.		2

統語的依存関係の種類を記載

- 依存先がない場合 root



例: Clinton defeated Dole.

ID	FORM	...	HEAD	DEPREL
1	Clinton		2	nsubj
2	defeated		0	root
3	Dole		2	obj
4	.		2	punct

[依存関係一覧](#)

UDの形式:まとめ

トルコ語UD(GB)より

- syntactic word による分解
- 品詞情報、形態論情報、統語情報は必ず明記
- 現在もアノテーション方針は議論が進行中

```
# sent_id = GK12-0012
# text = Siz çok iyi bir doktorsunuz.
# en = You're a very good doctor
1      Siz      siz      PRON      _      Case=Nom|Number=Sing|Person=2|PronType=Prs      5      nsubj      _      _
2      çok      çok      ADV      _      _      3      advmod      _      _
3      iyi      iyi      ADJ      _      _      5      amod      _      _
4      bir      bir      DET      _      Definite=Ind|PronType=Art      5      det      _      _
5-6    doktorsunuz      _      _      _      _      _      SpaceAfter=No
5      doktor  doktor  NOUN      _      Case=Nom|Number=Sing      0      root      _      _
6      sunuz    i      AUX      _      Mood=Ind|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin      5      cop      _      _
7      .      .      PUNCT     _      _      5      punct     _      _
```

- 全言語共通のアノテーション基準を用いて通言語的な比較ができる
 - 類型論
 - [Marneffe et al. \(2014\)](#), [Futrell \(2020\)](#)
 - 統語論の比較
 - [Alzetta et al. \(2018\)](#), [Kanayama et al. \(2018\)](#), [Choi et al. \(2021\)](#)
 - 形態論の比較
 - [Cöltekin and Rama \(2022\)](#), [Chen and Gerdes \(2017\)](#), [More and Tsarfaty \(2016\)](#)
- 他の言語理論との互換性
 - CCG: [Tran and Miyao \(2022\)](#), LFG: [Patejuk and Przepiórkowski \(2018\)](#)
- 多様なテキストへの応用
 - 数式: [Levine \(2023\)](#)
- ツール開発
 - 日本語形態素分析: [GiNZA](#)

少数言語の記述も兼ねたUDの活用例 ([Zariquiey et al. 2022](#))

- [Akuntsu語UD](#) (トゥピ語族)
- [Aurina語UD](#) (アラワク語族)
- [Beja語UD](#) (アフロアジア語族クシ語派)
- [Bororo語UD](#) (マクロ・ジェ語族)
- [チュクチ語UD](#) (チュクチ・カムチャツカ語族)
- [ユプト語UD](#) (アフロアジア語族エジプト語派)
- [Erzya語UD](#) (ウラル語族モルドヴィン語派)
- [Guajajara語UD](#) (トゥピ語族)
- [Kapor語UD](#) (トゥピ語族)
- [Kangri語UD](#) (印欧語族インド諸語)
- [キチェ語UD](#) (マヤ語族)
- [Livvi語UD](#) (ウラル語族バルト・フィン諸語)
- [Madi語UD](#) (アラワ語族)
- [Makurap語UD](#) (トゥピ語族)
- [マン島語UD](#) (印欧語族ケルト語派)
- [Mbya Guarani語UD](#) (トゥピ語族)
- [Moksha語UD](#) (ウラル語族モルドヴィン諸語)
- [Munduruku語UD](#) (トゥピ語族)
- [Nayini語UD](#) (印欧語族イラン諸語)
- [Nheengatu語UD](#) (トゥピ語族)
- [North Sami語UD](#) (ウラル語族フィン・ウゴル語派)
- [Pomak語UD](#) (印欧語族スラヴ語派)
- [Skolt Sami語UD](#) (ウラル語族フィン・ウゴル語派)
- [Soi語UD](#) (印欧語族イラン諸語)
- [西シエラ・プエブラ・ナワトル語UD](#) (ユト・アステカ語族)
- [Xavante語UD](#) (マクロ・ジェ語族)
- [シベ語UD](#) (ツングース諸語)
- [Yupik語UD](#) (エスキモー・アレウト語族)
- [Zaar語UD](#) (アフロアジア語族チャド諸語)

- プロジェクトとしてのUDはまだ始まって8年
- ガイドラインは頻繁に更新されている
 - 経験論的にガイドラインを更新
 - できるだけ多くの言語を考慮する必要がある
 - 生成文法、記述言語学、類型論などと同様

Our experience tells us that every language adds something to the general program of the scientific study of grammar.

Hale, K. (1998). On endangered languages and the importance of linguistic diversity.

長所

- 活気がある
- 言語の多様性
 - 言語数
 - 語族
 - 手話言語、死語、コードミキシング
- 貢献者の多様性(言語学者、プログラマ、計算言語学者...)
- 手軽に貢献できる

短所

- アノテーションの不一致(次スライド)
 - ミス、放置、言語学的知見の不足

アノテーションの不一致の問題

- 他のツリーバンクからの自動変換で生成されたUDツリーバンクも少なくなく、UDにそぐわないアノテーションになっていることがある
- 同じ言語のツリーバンクでも、貢献者によってアノテーション方針に食い違いがあったりする
- ツリーバンク間・言語間の貢献者の間での議論が活発でない
- 貢献者の都合で更新がストップしてしまう

とはいえ...

- まだ開始して数年のプロジェクトで、問題提起や議論の場も増えつつある
- 今後に期待

UDへ貢献するには？

Release checklist

1. GitHubアカウントの作成
2. UDレポジトリ管理担当者に連絡
 - a. 言語名、ツリーバンク名
 - b. GitHub上にレポジトリを作ってくれる
3. ツリーバンクの編集 ([How to start](#))
 - a. テキストを選定
 - b. アノテーション
 - c. 正しいフォーマットかどうか自動でテスト
 - d. GitHubレポジトリにファイル(.conllu)をアップロード
 - e. 年2回(5月・11月)に更新、公開される
4. ドキュメンテーションを作成

研究紹介: タタール語UD

Татар теле / Tatar tele

- タタール語 < キプチャク(北西)語群 < テュルク語族
- ロシア連邦タタールスタン共和国
 - 旧ソ連諸国、新疆、欧米にディアスポラ
- 約500万人の話者
 - ロシア語とのバイリンガル
- キリル文字正書法が主流
- **ロシア語の語彙の流入**
 - 典型的な借用語と異なり、ロシア語の音韻を保っている

Коронавирустан саклануның төп ысулы **вакцинация** булып тора — башка юл юк.

Koronavirustan saqlanuniñ töp isulı **vaksinatsiyä** bulıp tora — başqa yul yuq.

UD Tatar NMCTT

- 2021年11月15日公開
- 2280トークン
- テキスト:タタール語のニュースサイト
 - [Tatar Inform](#)
 - テキストの使用許可を取得
- 貢献者:私のみ(協力者募集中です)
- コード・ミキシングを明示的にアノテーション

タタール語UD: 見たい目

コード・ミキシングが起きている箇所をアノテーション

```
# sent_id = 5840560_0
# link = https://tatar-inform.tatar/news/health/08-10-2021/mishustin-koronavirustan-t-p-saklanu-charasy-vaktsinatsiya-bashka-yul-yuk-5840560
# genre = health
# text = Коронавирустан саклануның төп ысулы вакцинация булып тора – башка юл юк.
1   Коронавирустан   коронавирус   NOUN   _   Case=Abl|Number=Sing   2   obl   _   CSPoint=Коронавирус$тан|LangID=MIXED [RU$TT]
2   саклануның       сакла       VERB   _   Case=Gen|Number=Sing|VerbForm=Vnoun|Voice=Pass   4   nmod   _   LangID=TT
3   төп             төп        ADJ    _   _   4   amod   _   LangID=TT
4   ысулы           ысул       NOUN   _   Case=Nom|Number=Sing|Person[psor]=3   6   nsubj   _   LangID=TT
5   вакцинация     вакцинация  NOUN   _   Case=Nom|Number=Sing   6   xcomp   _   LangID=RU
6   булып         бул        VERB   _   VerbForm=Conv   0   root    _   LangID=TT
7   тора           тор        AUX    _   Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   6   aux     _   LangID=TT
8   –              –          PUNCT  _   _   11   punct  _   LangID=OTHER
9   башка         башка     ADJ    _   _   10   amod   _   LangID=TT
10  юл             юл        NOUN   _   Case=Nom|Number=Sing   11  nsubj   _   LangID=TT
11  юк             юк        ADJ    _   _   6   parataxis   _   LangID=TT|SpaceAfter=No
12  .             .         PUNCT  _   _   11   punct  _   LangID=OTHER
```

タタール語UD作成の流れ

- ニュースサイトよりテキストの使用許可を取得
- タタール語UDを作りたい旨を管理者に連絡
- アノテーション開始
 - 他のツリーバンク(特にテュルク諸語)を参照しながら進める
 - Google Colab上で統語依存関係を可視化しながら、手作業でアノテーション
- ツリーバンクファイル(.conllu)をGitHubにアップロード
- ドキュメンテーションを編集

タタール語内のロシア語コード・ミキシングの予測 ([Taguchi et al. 2022](#))

- 語の内部(形態素単位)で起きることがある

коронавирустан

koronavirus-tan

coronavirus-ABL

- ロシア語部分を予測することは、ラテン文字正書法への翻字などで重要
- タタール語コーパス構築や辞書構築などでも有用
- タタール語の文中に混ざったロシア語を定量的に調べることができる

タタール語UDの応用

```
# sent_id = 5840560_0
# link = https://tatar-inform.tatar/news/health/08-10-2021/mishustin-koronavirustan-t-p-saklanu-charasy-vaktsinatsiya-bashka-yul-yuk-5840560
# genre = health
# text = Коронавирустан саклануның төп ысулы вакцинация булып тора – башка юл юк.
1   Коронавирустан   коронавирус   NOUN   _   Case=Abl|Number=Sing   2   obl   _   CSPoint=Коронавирус$тан|LangID=MIXED [RU$TT]
2   саклануның       сакла       VERB   _   Case=Gen|Number=Sing|VerbForm=Vnoun|Voice=Pass   4   nmod   _   LangID=TT
3   төп               төп         ADJ    _   _   4   amod   _   LangID=TT
4   ысулы             ысул       NOUN   _   Case=Nom|Number=Sing|Person[psor]=3   6   nsubj   _   LangID=TT
5   вакцинация       вакцинация  NOUN   _   Case=Nom|Number=Sing   6   xcomp   _   LangID=RU
6   булып           бул        VERB   _   VerbForm=Conv   0   root   _   LangID=TT
7   тора             тор         AUX    _   Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   6   aux    _   LangID=TT
8   –                 –           PUNCT  _   _   11   punct  _   LangID=OTHER
9   башка           башка     ADJ    _   _   10   amod   _   LangID=TT
10  юл               юл        NOUN   _   Case=Nom|Number=Sing   11  nsubj   _   LangID=TT
11  юк               юк        ADJ    _   _   6   parataxis   _   LangID=TT|SpaceAfter=No
12  .                 .         PUNCT  _   _   11   punct  _   LangID=OTHER
```

タタール語UDからわかること

- 書き言葉のタタール語では、コード・ミキシングを起 open class の語に集中
- 約20%の名詞がロシア語(またはその派生形)

※ ニュース記事なので、他のジャンルや話し言葉の分布とは異なる可能性が大きい

Class	UPOS	Total	Russian	Mixed
Open	NOUN	413	21	62
	PROPN	79	34	8
	VERB	169	0	1
	ADJ	117	8	0
Closed	AUX	18	0	0
	DET	9	0	0
	ADV	40	0	0
	SCONJ	8	0	0
	ADP	35	0	0
	CCONJ	26	0	0
	PRON	26	0	0
NUM	12	0	0	
Other	PUNCT	167	0	0

Table 4: The distribution of UPOS tags in the treebank with respect to language code. The first column specifies whether the UPOS tag is an open class or a closed class.

どのようにして自動でロシア語部分を抜き出すか？

→ 2つの問題を同時に解く

- スパン同定タスク

- テキスト中のどこからどこまでを抜き出すか
- B(eginning), I(ntermediate), O(utside), S(ingleton) の4つのタグを各文字に付与

к о р о н а в и р у с т а н
B I I I I I I I I I O O O

- 言語識別タスク

- 文字列の言語を判定
- この場合は三つの候補: タタール語、ロシア語、その他)

スパン同定タスクと言語識別タスクを文字単位で解くモデルを訓練したい

→**条件付き確率場** (Conditional Random Fields; CRFs)

- 前後の情報やその他の関連する情報(素性)を考慮することができる(文脈)
 - コード・ミキシングは周辺の単語や文字列、品詞とある程度の関連がみられるため、これらを活用したい

訓練データ: タタール語UDのコード・ミキシングのアノテーション

形態素単位での自動識別

Precision (適合率): Xと予測したもののうち、実際にXであ

Recall (再現率): 全てのXのうち、Xであると予測できた確

F1スコア: 適合率と再現率の調和平均

Features	Precision	Recall	F1
Default	90.9	90.0	88.9
[-POS]	87.3	86.5	84.3
[-word]	86.4	86.5	84.9
[-POS, -word]	86.7	87.0	85.7

Table 8: Ablation study of features on NMCTT. Scores are calculated at a character level.

結果

- 少ないデータでも言語識別できている
- 品詞タグ、語 (FORM) などのUDに備わった追加の情報を入れると精度が上昇している

→ UDを利用した応用ツールの開発やコーパス研究ができそう

- 言語学と自然言語処理
 - トレンドや研究目標の面で分野の壁が大きくなっているように見えるが、根本の関心は言語
 - 言語処理では、大言語 (< ~20言語?) 以外はデータ(言語資源)が足りていない
 - (記述)言語学と言語処理技術を合わせた研究の機運も高まっている
- UDは今後もどんどん発展していく
 - 言語多様性の上昇 → より普遍性の高いフレームワークへ
 - 応用研究の機運が高まる: 類型論、コーパス言語学、ツール開発
 - 言語記述の一つの手段としての UD
 - ただし問題も多く残る
 - (記述)言語学者が必要とされている

おまけ: 今後やってみたいこと

発話データからUD構築

- ELANのアノテーション + UDのアノテーション
- 自動音声認識モデルを訓練 → UD形式でアウトプット
- 音声からコーパスへのパイプライン
→ フィールドワークをする記述言語学者に結びつけたい

現状のUDは書記言語に偏っている

- 無文字言語や、正書法が確立されていない言語は？
- 「話された形式」をもっと重視すべき

時間が余ったらフィールドワーク × 音声認識の話をしてします...

フォーマット変換

ELAN形式(.eafファイル = .xmlファイル)

フォーマット変換を自動化することはできるか

- xmlファイル → Python辞書形式
- → .tsv に成形してエクスポート(.conlluファイル)

スクリプト共有？

ありがとうございました

Email: ctaguchi@nd.edu