語彙データベースとしての Sudachi 辞書

株式会社ワークスアプリケーションズ・システムズ
ワークス徳島人工知能NLP研究所
高岡 一馬

辞書ベースで形態素解析をする意義

WORKS APPLICATIONS
徳島人工知能ＮＬＰ研究所

分割するだけでいい？

- ニューラルなら単語分割でじゅうぶん（？）
- 産業利用ではまだ辞書ベースが主流

辞書ベースだと語を直接同定できる

- 語をアンカーとして既存の知識と結びつける

Sudachi形態素辞書と結びつく情報

WORKS APPLICATIONS
徳島人工知能ＮＬＰ研究所

形態素情報

- 表記、読み、品詞

語彙正規化情報

- 表記ゆれ
  引越し　引っ越し　引越

語構成

- 複数分割粒度
- より下位の構成素の情報

Sudachiによる複数粒度分割の例

Sudachi形態素辞書と結びつく情報

WORKS APPLICATIONS
徳島人工知能ＮＬＰ研究所

Sudachi同義語辞書

- 同義語
- 旧称、別称、対訳
- 略称、略語
- 異表記、翻字
- 誤用、誤表記
- ドメイン情報

単語埋め込み

- chiVe [真鍋+ NLP2020]
  https://github.com/WorksApplications/chiVe

シソーラス

- 分類語彙表番号（UniDic経由）

活用のとりくみ

機械学習と辞書情報の融合

- 人が知っていることは書いた方がいい
- 例）事前学習モデルchiTraでの表記ゆれ対応（PT1-6)

データベースとして共有

- OSS、オープンデータとして公開
- ユーザとともに情報を集積していきたい

chiTra (Sudachi Transformers)

WORKS APPLICATIONS
徳島人工知能ＮＬＰ研究所

Hugging Faceで利用できるSudachiトークナイザ

- Hugging Face Transformersフレームワークで利用可能
- Sudachi.rsにより高速動作可能

日本語事前学習モデル

- 表記ゆれなど日本語テキストに固有な問題の発見と解決

商用利用可能なApache 2.0ライセンスで公開中

- https://github.com/WorksApplications/SudachiTra