{tiffany.blanchet, daniel.chen-1, enora.rice} @colorado.edu
github.com/dchensta/tdvz_nlp
University of Colorado Boulder

Automatic Canonical Segmentation for Teotitlán del Valle Zapotec Documentation Tasks

Tiffany Blanchet, Daniel Chen, & Enora Rice

Teotitlán del Valle Zapotec

•**Teotitlán del Valle Zapotec (TdVZ)** is a part of the Oto-Manguean family and is mostly spoken in Oaxaca, Mexico.
• TdVZ is morphologically complex and utilizes (Lorenzo 2021):
  • **affixes** (-) - more fixed, phonologically incorporated
  • **clitics** (=) - more flexible, phonologically removed
  • morphological compounds (+)
    • `gib+yag : metal + stick`

Canonical Segmentation

• **Surface segmentation** - splits a word into morphemes exactly as they appear orthographically
• **Canonical segmentation** divides a word into the "canonical", paradigmatic forms of its morphemes.

`cylindrically ->`
Surface Segmentation: `cylindr-ical-ly`
Canonical Segmentation: `cylinder-ical-ly`

• disambiguates between surface allomorphs of the same canonical morpheme
  • -> more accurately captures the distribution of a particular morpheme throughout corpora

Morpheme Preprocessing

• Source Text: Dissertation (Lorenzo 2021)
  • `python-docx` library - converts from `DOCX` into `TXT` for faster text extraction
  • We use regular expressions to extract all the IGT forms that follow a 4-line format:

```
(1)     bǽll.'dxî                          'txi̱w?
        bǽll=dxi                           tx´-æ=u̱
        intg.how.many=day          pot-go=2sg.if
        'How many days are you going?' (elic.)
```

Each morphological word is represented as a dictionary with four key-value pairs:

TdVZ Text (Original)	Segmentation	Gloss	EN Translation
bǽll.'dxî	bǽll=dxi	intg.how.many =day	How many days are you going?' (elic.)

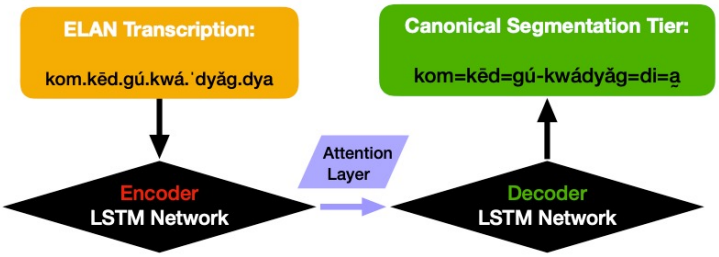
Applications

Application 1: Lookup Functionality for Manual Segmentation
  • Consult bidirectional lexicons of canonical morpheme <-> list of surface forms

```
Surface -> Canonical Lexicon:        Canonical -> Surface Lexicon:
bǽll.'dxî   :  bǽll=dxi              bǽll=dxi    :  bǽll.'dxî
'txi̱w?     :  tx´-æ=u̱              tx´-æ=u̱    :  'txi̱w?, 'txi̱w]
'xtę̂       :  xtę̂ny                 xtę̂ny       :  'xtę̂n, 'xtę̂
```

Application 2: Adding Automatic Canonical Segmentation Tier to TdVZ ELAN Transcription Files

Sequence-to-Sequence Language Models



Encoder-Decoder Model for producing automatic canonically segmented form of input text in surface form

• Sequence to sequence (seq2seq) models are a family of neural networks
  • Input: a sequence of text
  • Output: a new, transformed text sequence
• Canonical segmentation can be treated as a seq2seq task:
  • Input text: ELAN Transcription Tier surface form
  • Output text: canonical form of the composite morphemes
    `dya → di=a̱`
• To model the segmentation task, we leverage interlinear glossed text (IGT) in TdVZ to train an LSTM with attention

Preliminary Results

Architecture	Whole-Word Accuracy
Attentive-LSTM	49%

Accuracy from 10-fold cross-validation with 663 non-duplicated training samples

Next Steps:
• *Data*
  • Handling different orthography for ELAN transcriptions
    • Standardization of affix-clitic boundaries (wrong symbol results in inaccuracy label)
    • e.g. vowel duplication
• *Model*
  • Looping in training data from other Zapotec varieties
  • Testing new architectures (pointer-generator)
  • Hyperparameter tuning
  • Leveraging translations? Requires automatic word-alignment

Related Works
On Modeling Morphological Segmentation

• **Kann et al. 2016 -** bidirectional RNN encoder-decoder with neural reranker

• **Mager et al. 2020 - pointer-generator network** vastly improves the performance of the LSTM canonical segmentation model in the low-resource setting

• **Moeng et al. 2021 -** transformer performs best from list of sequence-to-sequence models on 4 Nguni languages