Word and Paradigm Morphology

David R. Mortensen

February 20, 2024

Project Ideas

It is time to start thinking about a project. Here are some ideas, based on what we've learned about so far (and some ideas from future units):

(1) **Tokenization**

a. Can we quantify the morphological well-foundedness of subword tokenization schemes (and relating this to downstream performance)?

b. Is it possible to build a better BPE/ULM/Morfessor?

c. Do character-level or byte-level models like ByT5 perform better on languages with extensive non-concatenative morphology than equivalent models that use subword tokens (like T5)? Consider also Mamba and MambaByte.

d. As part of Miniproject 1, some students found that tokenization with a VOLT vocabulary provides better morphological segmentation than Morfessor, even though the latter is designed to do morphological segmentation and the former is not. Is this true, generally?

(2) **Morphological Productivity**

a. Which LLMs can generate

b. To what degree does that explicit knowledge about morphological patterns demonstrated by LLMs correlate with the implicit knowledge that they use to engage in morphological generalization/productivity.

(3) **Orthography**

a. Can you build a G2P system for languages with complex orthographies like English or Arabic that can be deployed easily by downstream users?

b. Can you build a better input method (with autocomplete) for a low-resource language like Totonac?

(4) **Phonology**

a. When does working in IPA space of phonological feature space improve performance in multilingual modeling?

b. Can recent neural models result in better cognate detection? What about unsupervised cognate detection?

Review

(5) Theories of Morphology

a. Item and Arrangement (IA): morphology consists of items (morphemes) and constraints on how they can be combined (through concatenation)

b. Item and Process (IP): morphology consists of items (morphemes) and processes that apply to them (string-to-string functions paired with meaning-to-meaning functions)

c. Word and Paradigm (WP): morphology consists of relationships among the wordforms in paradigms (defined in terms of morphological properties).

(6) Rules in WP

a. Rules of Realization: rules that express how morphological properties are realized in wordforms (starting from the lemma and moving "outward").

b. Rules of Referral: rules that express how cells in a paradigm relate to one another (e.g., the Totonac first singular is always the third singular with *k* prepended.

Rules of Realization

Rules of referral provide instructions for predicting an inflected wordform based on a lemma or base[1] They are typically organized into blocks such the that first block's rules apply first, the second block's rules apply second, and so on **with no more than one rule applying per block.**

(7)
- **Blocks are ordered extrinsically.** The analyst chooses the order.
- **Rules in blocks are ordered intrinsically.** They are automatically ordered from most specific to least specific.

(8) Rule for realizing first person plural perfective by added -w in Totonac:

$$\begin{array}{c}\begin{bmatrix}1\\ \text{P}\\ \text{PFV}\end{bmatrix}\\ \text{X}\end{array} \rightarrow \text{Xw}$$

Relatively complex patterns can easily be modeled with defaults. For example, imagine a block like the following:

(9) Block *n*

$$\text{a.}\quad \begin{array}{c}\begin{bmatrix}1\\ \text{INCL}\end{bmatrix}\\ \text{X}\end{array} \rightarrow \text{X}$$

[1] We will define base as basic form containing only the stem and no inflection.

b. $$\begin{bmatrix} 1 \end{bmatrix} \atop \mathrm{X} \quad \rightarrow \mathrm{kX}$$

The empty string is added to the base just in case the wordform is to be first person inclusive. If it is otherwise first person (i.e., first person singular or first person plural exclusive) then *k*- is prepended.

See the in-class exercise on Totonac word-and-paradigm morphology.

Rules of Referral

A different way of looking at paradigmatic relationship is in terms of relationships between cells in the paradigm (or more generally, between morphologically related forms). For example, the relationship between first person singular and third person singular wordforms in Totonac can be expressed by the following rule:

(10) Rule for expressing the relationship between the third person singular and the first person singular in Totonac:

$$\underset{\mathrm{X}}{\begin{bmatrix} 3 \\ \mathrm{S} \end{bmatrix}} \quad \leftrightarrow \quad \underset{\mathrm{kX}}{\begin{bmatrix} 1 \\ \mathrm{S} \end{bmatrix}}$$

This kind of representation is likely more like the representation of paradigms in a seq2seq or language model than rules of realization. That is, these models learn analogical relationships between wordforms (which is what rules of referral capture). Of course, neural models do not learn rules per se—they learn statistical associations between strings—but rules of referral can serve as a useful way of conceptualizing problems in inflectional morphology (as, as we will see, derivational morphology and compounding).

References