Word and Paradigm Morphology

David R. Mortensen

February 18, 2025

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)?

(2) **Morphology**

a. Do transformer models have distinct representations for lexemes and inflectional properties? For example, given the wordforms *walk*, *walked*, *kiss*, and *kissed*, is there a part of the model where *walk* is closer to *walked* that to *kiss* and another where *walk* is closer to *kiss* than to *walked*?

b. Along the same lines, but from a mechanistic interpretability perspective, is it possible to locate inflectional properties within LLMs and manipulate the inflectional properties in outputs by manipulating neurons, etc.?

c. Which LLMs can generate novel words or wordforms in a way that resembles human language behavior?

d. To what degree does the explicit knowledge about morphological patterns demonstrated by LLMs correlate with the implicit knowledge that they use to engage in morphological generalization/productivity. That is, can LLMs engage in metacognition regarding morphological patterns.

e. Can interpretable representations of morphology like G3 be used to improve performance in neural models?

f. Can affix order be predicted using information-theoretic perspectives? Are affixes that appear close to the root likely to have more information content? Are they likely to have greater pointwise mutual information with the root?

g. How can improved morphological representations improve performance on downstream tasks like AMR parsing?

(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?

c. In what ways do scripts (and their encoding in Unicode) affect the behavior and performance of language models (under both intrinsic and extrinsic evaluation)?

(4) **Phonology**

a. When do phonologically-driven representations (phonemic transcriptions, phonetic embeddings, other kinds of learn phonological representations) improve performance on multilingual NLP tasks.

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

c. Is it possible to improve upon current approached to the phonological reconstruction of unattested ancestors of modern language families?

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): inflectional 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 root 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 realization provide instructions for predicting an inflected wordform based on a 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.**

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

(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:

$$\begin{bmatrix} 1 \\ P \\ PFV \end{bmatrix}$$
$$X \quad \rightarrow Xw$$

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

(9) Block *n*

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

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

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:

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

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