Flexica01: non-neural, alignment based

Lemma-to-inflected form transformation are generated directly by the following simple process:

**Step 1.** Find maximal continuous matches between lemma and inflected form.

*Example: understand → understood*

Extracted rules: \( \{ \text{\textbackslash{}v\textbackslash{}am\textbackslash{}t} \rightarrow \{ \text{\textbackslash{}v\textbackslash{}am\textbackslash{}t} \} \), where \{\text{\textbackslash{}v\textbackslash{}am\textbackslash{}t} and \{\text{\textbackslash{}v\textbackslash{}am\textbackslash{}t} are groups.*

**Step 2.** Starting with previously generated transformation pattern(s), generate a set of patterns more specific to a given training word by treating a limited number of characters as concrete (i.e. standing outside any group).

For the example from previous step and a limit of one character: \( \{ \text{\textbackslash{}v\textbackslash{}am\textbackslash{}t} \rightarrow \{ \text{\textbackslash{}v\textbackslash{}am\textbackslash{}t} \} \), where \{\text{\textbackslash{}v\textbackslash{}am\textbackslash{}t} and \{\text{\textbackslash{}v\textbackslash{}am\textbackslash{}t} are groups.*

When predicting a form, score matching candidate patterns using the following three components:

1. A frequency of transformation occurrence in a training set;
2. The diversity \( \Delta \) of marginal (the first one and the last one) letters in groups as they occurred in different fits of a given transformation found in the training set.
3. Specificity \( s \) which here means the number of concrete characters in the pattern (without counting characters falling into groups).

We were using the following empirical formula:

\[
G = \frac{1}{2} \log f + 6 \log \Delta + 12 s
\]

Near-misses (the second scored transform was correct)

<table>
<thead>
<tr>
<th>G</th>
<th>0.00</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\Delta)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Flexica02: Hard attention, multilingual (family-based)**

This neural system is based on hard monotonic attention model proposed in [Aharoni and Goldberg](2017), with the same loss function, but with the following differences:

1. We combined all the languages belonging to a given family into a single dataset, having added two extra features such as language and genre.
2. We used maximal continuous sub-string search to organize alignment between lemma and its inflected form in order to advance hard attention state (abolishing one-by-one alignment of mismatching characters).

**Flexica03: Adding hallucinated data**

Inspired by [Anastassopoulos and Neninić](2019), we added 200 samples per language per part-of-speech (POS) in order to produce hallucinated inflection samples that look like real. We reused the predictor from Flexica01. We also enriched the model with word-generator [Shcherbakov et al.](2016)Shcherbakov, Vylomova, and Thierberg, http://regexus.com/ag.php to produce more phonotactically plausible forms: 1) Word generator trained on inflected forms for a given POS produces samples of hallucinated inflected forms (without distinction of grammatical features); 2) The reverse Flexica01 predictor produces tags/lemma for each hallucinated inflected form. Accuracy was significantly improved in low-resource languages (such as Maori, Zarma, Tajik, Anglo-Norman, Middle High/Low German).

**Conclusion**

We proposed and tested (1) multilingual training, and (2) pattern-based hallucinated inflections as possible enhancements of sequence-to-sequence morphology modeling for diverse low-resource languages. We also developed a simple non-neural approach based on multi-variant search of common inflection patterns.