morphogen: Translation into Morphologically Rich Languages with Synthetic Phrases
Eva Schlinger  Victor Chahuneau  Chris Dyer
School of Computer Science  Carnegie Mellon University  Pittsburgh, PA, USA

Motivation
morphogen is a tool that improves translation into morphologically rich languages by modeling inflectional morphology using rich source context and either supervised morphological analyses or unsupervised morphological segmentations.

```
| Думайтье | “think!” [pl] |
| +mood    |                |
| +number  |                |
| +tense   |                |
| +person  |                |

Думать (V) ➞ “to think”
+tense +gender +number
dумала ➞ “she thought”
+aspect

“I will think”
```

We want a single technique that helps all morphologically rich languages (fusional and agglutinative), and makes use of whatever language specific information that is available.

Modeling Inflection
We model each target word \( f \) as a combination of a stem \( e \) and an inflectional pattern \( \mu \) by some operation \( \star \), which is dependent on the morphological analyzer used.

e.g. \( \text{run} \rightarrow \text{PAST} = \text{ran} \)

We decompose the probability of generating each target word \( f \) given a source sentence \( e \) and alignment \( i \) in the following way:

\[
p(f | e, i) = \sum_{e \rightarrow f} p(e | i) \cdot p(\mu | e, i)
\]

where:
- \( p(e | i) \): source context feature vector function
- \( p(\mu | e, i) \): target morphology feature vector function

\( \Omega^* \): set of allowed inflections. Restricted to those seen in parallel + monolingual training data

\( W \) and \( V \) are the parameter matrices that we want to learn:

\[
p(\mu | e, i) = \frac{\exp \left[ \phi(e, i)^T \cdot W \cdot \psi(\mu) + \psi(\mu)^T \cdot V \cdot \phi(e, i) \right]}{\sum_{\mu \in \Omega} \exp \left[ \phi(e, i)^T \cdot W \cdot \psi(\mu) + \psi(\mu)^T \cdot V \cdot \phi(e, i) \right]}
\]

(1)

Evaluation
We evaluate our approach in translation as our decoder and perform MIRA decoding.

- A baseline system, using a 4-gram and bilingual data available.
- An enriched system with a class-based system and Brown clusters.
- The enriched system further enriched.

Table: Translation quality

```
<table>
<thead>
<tr>
<th>Baseline</th>
<th>+Class</th>
<th>+Synth</th>
<th>unsup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Running morphogen
Our Python implementation is available at https://github.com/jhclark/ducttape
We provide sample workflows that augment the grammars produced by the standard translation model. The resulting augmented grammars are then added to the grammars produced by the standard translation model. The resulting augmented grammars are then added to the grammars produced by the standard translation model.

Synthetic Phrases
- **Learning**
  - Automatic: \( \text{pytat\’ya} \) \( \rightarrow \text{pytat\’ya} \)
  - Synthetic: \( \text{pytat\’ya} \) \( \rightarrow \text{pytat\’ya} \)

References
- https://github.com/vchahun/fast_umorph