Character Eyes: Seeing Language through Character-Level Taggers

Yuval Pinter  @yuvalpi
Marc Marone  @ruyimarone
Jacob Eisenstein  @jacobeisenstein

Bloomberg  Microsoft Research  Google

Blackbox NLP 2019
https://github.com/ruyimarone/character-eyes
Taggers

The cat walked fast
Neural Taggers

The cat walked fast
Character-level Neural Taggers

The cat walked fast

The cat walked fast
The cat walked fast
Recurrent Taggers – Good at Finding Morphemes?

The cat walked fast
Recurrent Taggers – Good at Finding Morphemes?

Agglutination
Recurrent Taggers – Good at Prefixes and Suffixes?
Recurrent Taggers – Good at Prefixes and Suffixes?

Prefixing morphology (e.g. Coptic)
Recurrent Taggers – Can They Handle diSCoNtinUiTY?

The cat waeldk fast
Recurrent Taggers – Can They Handle diScOntinUity?

The cat waeldk fast

Introflexive morphology (Hebrew, Arabic)
Main Idea(s)

Language

walked
the cat
walked
Main Idea(s)

Language

walked
the cat
waeldk

Model

measure how models encode different linguistic patterns
Main Idea(s)

Language

walked
the cat
waeldk

Model

characterize languages based on model analysis; help engineer language-aware systems
The cat walked fast
Analysis Primitive – Unit Decomposition

- Assumption: units are “in charge” of tracking morphemes that help predict POS
Analysis Primitive – Unit Decomposition

- Assumption: units are “in charge” of tracking morphemes that help predict POS
- Hypothesis: easy for **agglutinations**, difficult for **introflexions**
Analysis Primitive – Unit Decomposition

- Assumption: units are “in charge” of tracking morphemes that help predict POS
- Hypothesis: easy for agglutinations, difficult for introflexions
- Hypothesis: unit’s direction affects ease of tracking suffixes vs. prefixes
Evidence?

- Turkish is an **agglutinative** language
  - *ev* ‘house’; *evler* ‘houses’; *evleriniz* ‘your houses’; *evlerinizden* ‘from your houses’
Evidence?

- Turkish is an **agglutinative** language
  - ev ‘house’; evler ‘houses’; evleriniz ‘your houses’; evlerinizden ‘from your houses’
Evidence?

- Turkish is an **agglutinative** language
  - ev ‘house’; evler ‘houses’; evleriniz ‘your houses’; evlerinizden ‘from your houses’

![Diagram showing activation levels for units 3 and 124.](image-url)
The cat walked fast
Model & Data

- Universal Dependencies (n=24)
  - POS tags + Morphosyntactic Descriptions

The cat walked fast
Model & Data

- Universal Dependencies (n=24)
  - POS tags + Morphosyntactic Descriptions
- Linguistic diversity – morph. synthesis:
  - 5 agglutinative languages
  - 2 introflexive languages
  - 3 isolating, 14 fusional

Source for language classes: WALS
Model & Data

- Universal Dependencies (n=24)
  - POS tags + Morphosyntactic Descriptions

- Linguistic diversity – morph. synthesis:
  - 5 agglutinative languages
  - 2 introflexive languages
  - 3 isolating, 14 fusional

- Linguistic diversity – affixation:
  - (All) 1 prefixing language
  - 2 non-affixing
  - 2 equally pre- and suffixing
  - 19 suffixing

Source for language classes: WALS
Model & Data

- Universal Dependencies (n=24)
  - POS tags + Morphosyntactic Descriptions
  - Linguistic diversity (synthesis + affixation)

- Word → Tag: Bidirectional LSTM + MLP
  - (Not analyzed)
  - No word embeddings
Model & Data

- Universal Dependencies (n=24)
  - POS tags + Morphosyntactic Descriptions
  - Linguistic diversity (synthesis + affixation)

- Word → Tag: Bidirectional LSTM + MLP
  - (Not analyzed)
  - No word embeddings

- Char → Word: Bidirectional LSTM
  - Char embedding size: 256
Analysis Metrics
Analysis Metrics

- Run model on training data words
Analysis Metrics

- Run model on training data words
- Collect activation levels for each unit
Analysis Metrics

- Run model on training data words
- Collect activation levels for each unit
- Aggregate to single measure
  (e.g. average absolute or max-delta)
Analysis Metrics

● Run model on training data words
● Collect activation levels for each unit
● Aggregate to single measure (e.g. **average absolute** or **max-delta**) 
● Bin per unit over parts of speech

<table>
<thead>
<tr>
<th>Unit 42</th>
<th>[0.0,0.1)</th>
<th>[0.1,0.2)</th>
<th>...</th>
<th>[0.9,1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN</td>
<td>8</td>
<td>2</td>
<td>...</td>
<td>40</td>
</tr>
<tr>
<td>VERB</td>
<td>20</td>
<td>0</td>
<td>...</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ADJ</td>
<td>10</td>
<td>10</td>
<td>...</td>
<td>10</td>
</tr>
</tbody>
</table>
Analysis Metrics

- Run model on training data words
- Collect activation levels for each unit
- Aggregate to single measure (e.g. average absolute or max-delta)
- Bin per unit over parts of speech
- Mutual Information metric – POS Discrimination Index, or PDI
  - (Higher PDI = better discriminator)

\[
\sum_{t=1}^{T} \sum_{b=1}^{B} \left( \ln P(t, b) - \ln P(t) - \ln P(b) \right)
\]
Analysis Metrics

- Run model on training data words
- Collect activation levels for each unit
- Aggregate to single measure (e.g. average absolute or max-delta)
- Bin per unit over parts of speech
- Mutual Information metric – POS Discrimination Index, or PDI
  - (Higher PDI = better discriminator)
- Aggregate across units by
Analysis Metrics

- Run model on training data words
- Collect activation levels for each unit
- Aggregate to single measure (e.g. **average absolute** or **max-delta**)
- Bin per unit over parts of speech
- Mutual Information metric – POS Discrimination Index, or **PDI**
  - (Higher PDI = better discriminator)
- Aggregate across units by
  - Summing total mass
  - Reporting % of **forward** units before **mass median**

\[
\sum_{t=1}^{T} \sum_{b=1}^{B} P(t, b) \left[ \ln P(t, b) - \ln P(t) - \ln P(b) \right]
\]
Findings (Cherry Pick)

- Coptic: agglutinative, prefixing
  - Large mass (easy to distinguish POS based on char sequence)
  - Forward-heavy (71%)

- English: fusional, suffixing
  - Small mass (hard to capture POS)
  - Backward-heavy (80%)
Findings (General Trends)

<table>
<thead>
<tr>
<th>Language</th>
<th>Total PDI mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamil</td>
<td>71.0</td>
</tr>
<tr>
<td>Irish</td>
<td>62.0</td>
</tr>
<tr>
<td>Coptic</td>
<td>58.1</td>
</tr>
<tr>
<td>Hungarian</td>
<td>47.9</td>
</tr>
<tr>
<td>Greek</td>
<td>31.2</td>
</tr>
<tr>
<td>Turkish</td>
<td>30.1</td>
</tr>
<tr>
<td>Russian</td>
<td>25.9</td>
</tr>
<tr>
<td>Thai</td>
<td>25.9</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>25.0</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>24.2</td>
</tr>
<tr>
<td>Chinese</td>
<td>23.8</td>
</tr>
<tr>
<td>Danish</td>
<td>21.7</td>
</tr>
<tr>
<td>Swedish</td>
<td>20.8</td>
</tr>
<tr>
<td>Basque</td>
<td>20.6</td>
</tr>
<tr>
<td>Indonesian</td>
<td>20.3</td>
</tr>
<tr>
<td>Latvian</td>
<td>17.0</td>
</tr>
<tr>
<td>Spanish</td>
<td>16.1</td>
</tr>
<tr>
<td>English</td>
<td>16.0</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>15.6</td>
</tr>
<tr>
<td>Italian</td>
<td>14.1</td>
</tr>
<tr>
<td>Arabic</td>
<td>12.6</td>
</tr>
<tr>
<td>Hebrew</td>
<td>11.4</td>
</tr>
<tr>
<td>Persian</td>
<td>10.3</td>
</tr>
<tr>
<td>Hindi</td>
<td>8.4</td>
</tr>
</tbody>
</table>
Findings (General Trends)

- 4/5 agglutinatives hold 4/6 top total-mass positions

<table>
<thead>
<tr>
<th>Language</th>
<th>Total PDI mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamil</td>
<td>71.0</td>
</tr>
<tr>
<td>Irish</td>
<td>62.0</td>
</tr>
<tr>
<td>Coptic</td>
<td>58.1</td>
</tr>
<tr>
<td>Hungarian</td>
<td>47.9</td>
</tr>
<tr>
<td>Greek</td>
<td>31.2</td>
</tr>
<tr>
<td>Turkish</td>
<td>30.1</td>
</tr>
<tr>
<td>Russian</td>
<td>25.9</td>
</tr>
<tr>
<td>Thai</td>
<td>25.9</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>25.0</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>24.2</td>
</tr>
<tr>
<td>Chinese</td>
<td>23.8</td>
</tr>
<tr>
<td>Danish</td>
<td>21.7</td>
</tr>
<tr>
<td>Swedish</td>
<td>20.8</td>
</tr>
<tr>
<td>Basque</td>
<td>20.6</td>
</tr>
<tr>
<td>Indonesian</td>
<td>20.3</td>
</tr>
<tr>
<td>Latvian</td>
<td>17.0</td>
</tr>
<tr>
<td>Spanish</td>
<td>16.1</td>
</tr>
<tr>
<td>English</td>
<td>16.0</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>15.6</td>
</tr>
<tr>
<td>Italian</td>
<td>14.1</td>
</tr>
<tr>
<td>Arabic</td>
<td>12.6</td>
</tr>
<tr>
<td>Hebrew</td>
<td>11.4</td>
</tr>
<tr>
<td>Persian</td>
<td>10.3</td>
</tr>
<tr>
<td>Hindi</td>
<td>8.4</td>
</tr>
</tbody>
</table>
Findings (General Trends)

- 4/5 agglutinatives hold 4/6 top total-mass positions

- 2/2 introflexives in bottom 2/4 spots (Persian and Hindi below, both fusional with non-Latin charsets)

<table>
<thead>
<tr>
<th>Language</th>
<th>Total PDI mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamil</td>
<td>71.0</td>
</tr>
<tr>
<td>Irish</td>
<td>62.0</td>
</tr>
<tr>
<td>Coptic</td>
<td>58.1</td>
</tr>
<tr>
<td>Hungarian</td>
<td>47.9</td>
</tr>
<tr>
<td>Greek</td>
<td>31.2</td>
</tr>
<tr>
<td>Turkish</td>
<td>30.1</td>
</tr>
<tr>
<td>Russian</td>
<td>25.9</td>
</tr>
<tr>
<td>Thai</td>
<td>25.9</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>25.0</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>24.2</td>
</tr>
<tr>
<td>Chinese</td>
<td>23.8</td>
</tr>
<tr>
<td>Danish</td>
<td>21.7</td>
</tr>
<tr>
<td>Swedish</td>
<td>20.8</td>
</tr>
<tr>
<td>Basque</td>
<td>20.6</td>
</tr>
<tr>
<td>Indonesian</td>
<td>20.3</td>
</tr>
<tr>
<td>Latvian</td>
<td>17.0</td>
</tr>
<tr>
<td>Spanish</td>
<td>16.1</td>
</tr>
<tr>
<td>English</td>
<td>16.0</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>15.6</td>
</tr>
<tr>
<td>Italian</td>
<td>14.1</td>
</tr>
<tr>
<td>Arabic</td>
<td>12.6</td>
</tr>
<tr>
<td>Hebrew</td>
<td>11.4</td>
</tr>
<tr>
<td>Persian</td>
<td>10.3</td>
</tr>
<tr>
<td>Hindi</td>
<td>8.4</td>
</tr>
</tbody>
</table>
Direction Balance Study
Direction Balance Study

- Some languages might not need two equal LSTM directions
Direction Balance Study

- Some languages might not need two equal LSTM directions
- What if... they don’t need one of them at all?
Direction Balance Study

● Some languages might not need two equal LSTM directions

● What if they need them in a different balance? Somewhere in the middle?

● What if... they don’t need one of them at all?
Balance Study – Results
Balance Study – Results

- Can unidirectional models outperform bidirectional ones?
Balance Study – Results

- Can unidirectional models outperform bidirectionals?
  - Yes.
    - Especially on agglutinative languages and on suffixing languages
    - Fully-forward better than fully-backward
Balance Study – Results

- Can unidirectional models outperform bidirectional models?
  - Yes.
    - Especially on agglutinative languages and on suffixing languages
    - Fully-forward better than fully-backward
    - MAJOR caveat – 128x128 > 2^*(64x64)
Balance Study – Results

- Can unidirectional models outperform bidirectional models?
  - Yes.
    - Especially on agglutinative languages and on suffixing languages
    - Fully-forward better than fully-backward
    - MAJOR caveat – 128x128 > 2*(64x64)

- Is there a sweet spot in the middle?
Balance Study – Results

● Can unidirectional models outperform bidirectionals?
   ● Yes.
      ○ Especially on agglutinative languages and on suffixing languages
      ○ Fully-forward better than fully-backward
      ○ MAJOR caveat – $128 \times 128 > 2^\ast(64 \times 64)$

● Is there a sweet spot in the middle?
   ● Not that we can tell.
Summary + Open Ends
Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier
Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any \(<\text{instance, unit}\)> metric on any neural classifier
- Found substantial differences between differently-balanced recurrent models
Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier
- Found substantial differences between differently-balanced recurrent models

- Are we quantifying **data** instead of **languages**?
Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier
- Found substantial differences between differently-balanced recurrent models
- Are we quantifying **data** instead of **languages**?
- Affixing: many languages (e.g. English) have higher PDI for **backward** units, but fare better with more **forward** units. Is this:
Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier
- Found substantial differences between differently-balanced recurrent models

- Are we quantifying **data** instead of **languages**?

- Affixing: many languages (e.g. English) have higher PDI for **backward** units, but fare better with more **forward** units. Is this:
  - A saturation effect?
Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier
- Found substantial differences between differently-balanced recurrent models

- Are we quantifying **data** instead of **languages**?

- Affixing: many languages (e.g. English) have higher PDI for **backward** units, but fare better with more **forward** units. Is this:
  - A saturation effect?
  - Fault in assuming PDI measures unit importance?
Thank You!

https://github.com/ruyimarone/character-eyes

uvp@gatech.edu
mmarone6@gatech.edu