**MORSE:** Semantic-ally Drive-n MORpheme SEgment-er

Samuel MORSE minimized the number of on-off clicks for non-verbal communication.

This MORSE minimizes the vocabulary size for Natural Language Processing systems.
1

Morpheme Segmentation
Hope

ful

ly

Hopefully
Not a trivial task

Players
Playing
Beijing
Butterflies
Applications

Machine Translation

Model:

Quickly
• Rapidement

Sad
• Triste

Test:

Sadly

Model:

Quickly
• Rapide

Sadly

Test:

Tristement
Here at Toyota World, we have the **cheapest cars** in town. We are proudly called the first and last stop.

...
Previous Work
Helplessly

Letter Successor Variety (Harris, 1970)
Morfessor (Creutz and Lagos, 2005)

Help: 2387
Helping: 1586
Helper: 498
Helps: 2437

Jump: 1847
Jumping: 1664
Jumper: 1290
Jumps: 2987
Downsides

Freshman

Butterflies

Butterfly ies ✅
Locally Semantic

Cosine similarity

car  car
caring
cars

(Schone and Jurafsky, 2000)
(Narasimhan et al., 2015)
(Luo et al., 2017)
Distinguishing criteria

car \xrightarrow{\text{}} cars
player \xrightarrow{\text{}} players
runner \xrightarrow{\text{}} runners
goal \xrightarrow{\text{}} goals
play \xrightarrow{\text{}} plays
fine \xrightarrow{\text{}} fines
wheel \xrightarrow{\text{}} wheels
hand \xrightarrow{\text{}} hands
laptop \xrightarrow{\text{}} laptops
lab \xrightarrow{\text{}} labs
MORSE

**Input:** Word Embeddings

**Segmentation:** Optimization Problem

Unsupervised Morphology Learning

4 hyperparameters: Small tuning dataset
Step 1

Learning Morphology
Collecting candidate morphological rules

Vocabulary:

- jump
- play
- buy
- jumping
- playing
- buying
- jumper
- player
- buyer
- and
- stand

Examples:

- (suf, ∅, ing):
  - (jump, jumping)
  - (play, playing)
  - (buy, buying)

- (suf, ∅, er):
  - (jump, jumper)
  - (play, player)
  - (buy, buyer)

- (pre, ∅, st):
  - (and, stand)
  - (ore, store)
  - (one, stone)

(Soricut and Och, 2015)
Signals

Orthographic

quick                quickly
beautiful           beautifully
confident           confidently
wrong               wrongly

Semantic

Word Embeddings

quick                quickly
beautiful           beautifully
confident           confidently
wrong               wrongly
What makes a good rule?

Signal 1: Orthography

Rule = (suf, ∅, ly)  Size = 8723
(quick, quickly)  (beautiful, beautifully)
(confident, confidently)
.............................................
............................... (wrong, wrongly)

Rule = (pre, ∅, st)  Size = 16
(ore, store) ......
(amp, stamp)
What makes a good rule?

**Signal 2: Semantics**
What makes a good member of a rule?

**Scope: Vocabulary-Wide**

- Quick
- Quickly
- Confident
- Confidently
- Wrong
- Wrongly
- Beautiful
- Beautifully
- Only
- On
What makes a good member of a rule?

**Scope: Local**

confidently

only

on
Step 2
Segmenting
Linear Optimization Problem

<table>
<thead>
<tr>
<th></th>
<th>(ring, uncaring)</th>
<th>(caring, uncaring)</th>
<th>(uncaring, uncaring)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₁</td>
<td>![Green Icon]</td>
<td>![Red Icon]</td>
<td>![Green Icon]</td>
</tr>
<tr>
<td>t₂</td>
<td>![Green Icon]</td>
<td>![Yellow Icon]</td>
<td>![Green Icon]</td>
</tr>
<tr>
<td>t₃</td>
<td>![Green Icon]</td>
<td>![Red Icon]</td>
<td>![Green Icon]</td>
</tr>
<tr>
<td>t₄</td>
<td>![Green Icon]</td>
<td>![Red Icon]</td>
<td>![Yellow Icon]</td>
</tr>
</tbody>
</table>

uncaring
Iterate

(car, caring) (care, caring) (carol, caring)

un + caring
Iterate

un + care + ing
Experiments
Experimental Setup

Training

Languages

Gold Datasets

Morpho Challenge

- jumping
- playing
- jumps
- calls
- rooms

- jump
- play
- jump
- call
- room
Experiments

![Bar chart showing performance metrics for English, Turkish, and Finnish.](chart)

- English
  - Morfessor: 64.35
  - MORSE: 70.32
- Turkish
  - Morfessor: 31.01
  - MORSE: 38.07
- Finnish
  - Morfessor: 34.06
  - MORSE: 14.98
Morpho Challenge downsides

- Non-compositional
- Trivial instances
- Human error

Business
Turning-point
Player’s
Turning
Experiments

New Dataset: SD17

- 2000 words
- Compositional
- 91% inter-annotator agreement
- In canonical (butterfly + ies) and non-canonical version (butterfly + ies)
Results on SD17

- Morfessor: 57.31
- MORSE (tuned on MC): 81.01
- MORSE (tuned on SD17): 83.96

F-score
Against state-of-the-art

<table>
<thead>
<tr>
<th>Method</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MORSE</td>
<td>83.96</td>
</tr>
<tr>
<td>MorphoChain</td>
<td>79.9</td>
</tr>
<tr>
<td>Morfessor S + W</td>
<td>67.4</td>
</tr>
<tr>
<td>Morfessor S + W+ L</td>
<td>67.14</td>
</tr>
</tbody>
</table>
Negative Dataset

- 100 words like: honeymoon, passport, outdoors
- Checks for robustness
Looking forward

- Robustness to highly agglutinative languages
- Extending to other languages (non-concatenative)

k̂aṭābā
Looking forward

- Morphological mappings across languages

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>(suf, ∅, ly)</td>
<td>(suf, ∅, ment)</td>
</tr>
<tr>
<td>(suf, ∅, s)</td>
<td>(suf, ∅, s)</td>
</tr>
<tr>
<td></td>
<td>(suf, ∅, es)</td>
</tr>
</tbody>
</table>
Links

https://morse.mybluemix.net

https://github.com/yoonlee95/morse_segmentation
Thank you

Questions?
Effect of Hyperparameters

Recall

Precision
Prerequisite

Morpho-syntactic regularities in word vectors

Valid rule with an invalid instance

\[(\text{suf}, \emptyset, \text{ing}) \quad (s, \text{sing})\]

Invalid rule

\[(\text{pre}, \emptyset, s)\]
Demo

morse.mybluemix.net