Unsupervised text segmentation

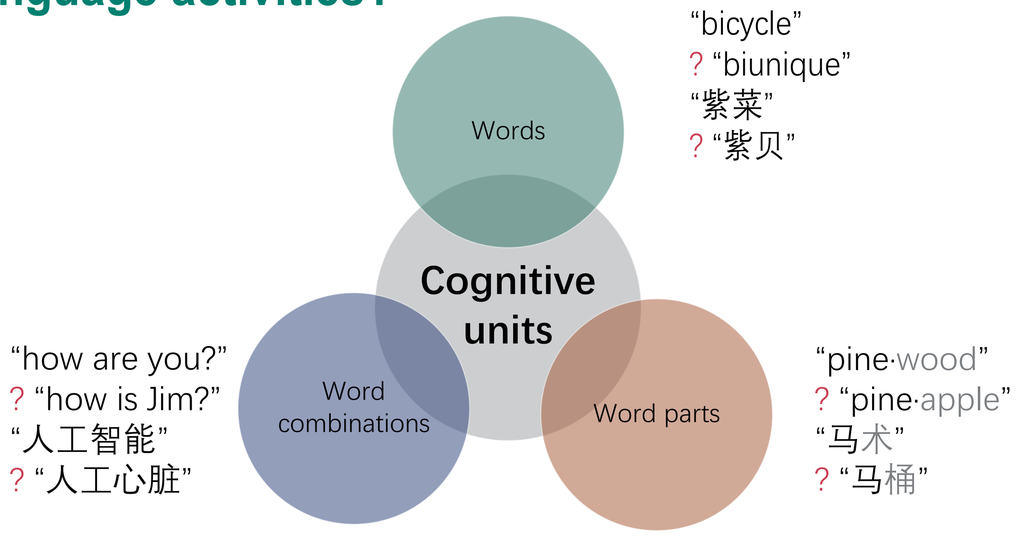
Predict ↓ ↑ *Evaluate*

Eye fixations during reading

Jinbiao Yang[1,2]
Stefan L. Frank[2]
Antal van den Bosch[3]

[1)] MPI for Psycholinguistics
[2)] CLS, Radboud University
[3)] KNAW Meertens Institute

What are the "*cognitive units*" in our daily language activities?



An example of cognitive segmentation of language:
"*I am |a |paleo|anthropolog|ist|.*"

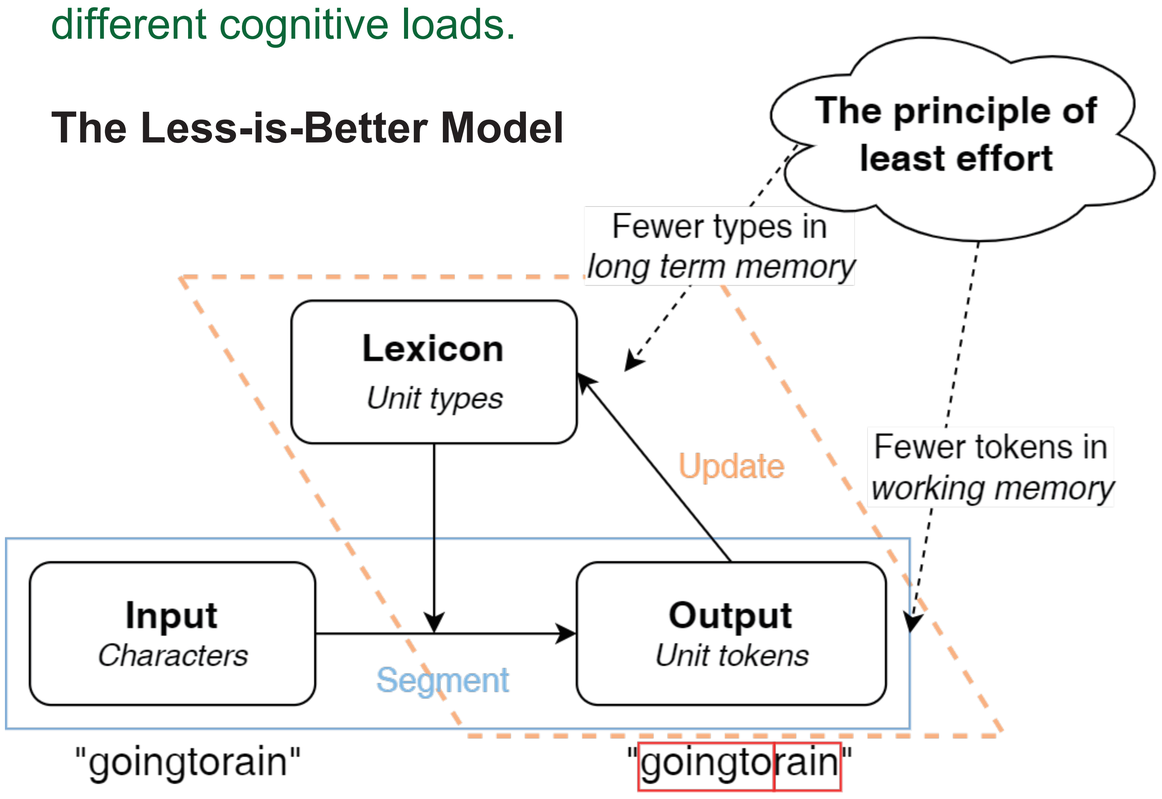
The cognitive units of language:
1. **diverse forms**;
2. **implict boundaries**.

An unsupervised computational exploration of cognitive units

Backgroud -- The principle of least effort:
"Each individual will adopt a course of action that will involve the expenditure of the probably least average of his work." (Zipf, 1949)

Linking hypothesis:
We assumes cognitive units are the units that minimize different cognitive loads.

The Less-is-Better Model



Segmentation:

Larger-first selection

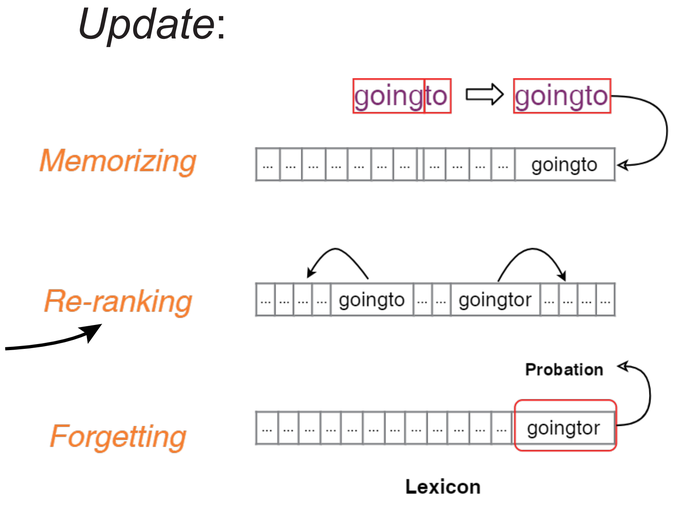
goingtorain...

Counterfactual evalution

goingtorain... Bad

goingtorain... Good ✓

Update:



Model outputs:

	English	*Dutch*
Input	i was trying to make up my mind what to do	was ik nog aan het overleggen wat ik zou gaan doen
Output	i was | trying to | make | up | my mind | what | to do	was ik | nog | aan het | over| leggen | wat ik | zou gaan | doen

Computational advantages (Yang et al., 2020):
1. Minimal description length;
2. Minimal Bits-per-character scores on tiny language models;
3. Can be used for Chinese word segmentation.

An emprical exploration of cognitive units

Backgroud -- Eye fixations during reading
Some words are skipped, and some words are re-fixated.

Linking hypothesis:
We assumes that reading is cognitive-unit-wise rather than word-wise, therefore:
Eye fixations = Centers of cognitive units.

A demo:

Experiment

1. **Train** the Less-is-Better model on the Ghent Eye-Tracking Corpus (GECO) corpus;
2. **Segment** the corpus by the model;
3. **Predict** the locations of eye fixations based on segmentation;
4. **Compare**:
   a. the predicted number of fixations on each word,
   b. the observed number of fixations on each word.

Results (Yang et al., 2022)

Model	*English*	*Dutch*
Less-is-better	53.06	**51.87**
Adaptor Grammar (collocation)	**53.35**	51.45
Chunk-Based Learner	52.20	50.04
Fixation counts determined by word length	50.82	50.57
Word-by-Word reading	38.32	38.68
Adaptor Grammar (word)	30.10	28.95

The weighted F1 scores between different models and the ground truth.

Take-home message

- Cognitive units of language are :
  (×) words/morphemes/phrases.
  (√) the units that can minimize the cognitive load.
- Reading is cognitive-unit-by-cognitive-unit.
- Unsupervised model can learn cognitive units.

Jinbiao.Yang@mpi.nl • twitter.com/itsbiao • github.com/ray306