The 25th 人工智慧研討會
TAAI 2020
The 25th International Conference on Technologies and Applications of Artificial Intelligence

Probabilistic Segmentation of Word Forms into Affixes and Word Roots

Tsen Hsieh, Jason S. Chang

Background

- Previous morphological segmentation approaches often use statistics of hypothesized stems and affixes to identify the word parts.
- However, roots and affixes may be identified incorrectly (e.g., “ally” stemming to “all” or splitting “rated” into “rat”, “ed” instead of “rate”, “ed”).
- We present a new method which can generate segmentation of a given word using a probabilistic model.
- We adjust the probabilistic model to improve the precision rate.

Method

- Define a probabilistic model to calculate the probability for a segmentation of a string of letters.
  1. The probability of a certain root is:

$$P(R) = \frac{frequency\ of\ R}{total\ amount\ of\ roots}$$

  2. The probability of a non-existing segment is:

$$P(S) = 1/(N * 10^{L})$$

  3. The probability for the segmentation of a word is:

$$P(S_{1:n}) = \prod_{k=1:n} P(S_k)$$

- Improve the model by considering rules for affiliating suffixes to word roots and positions of affixes in a word.
- Extract the most probable segmentation of a word:
  1. ~~Stem~~ Lemmatization the word.
  2. Segment the given stem into possible segmentation candidates.
  3. Use the probabilistic model we developed to assign probability to each candidates.
  4. Choose the one with the highest probability.

Result

- Testing data: 50 words from 《英文字根字首神奇記憶法》, denoted as *Testing-Kang*, and 250 words from *Concise Dictionary of Roots for English Words*, denoted as *Testing-Seya*.
- Evaluation results: *Testing-Kang* 94%, *Testing-Seya 86%.*

Table I

Correct segmentation results

word	result	word	result
technically	techn+ical+ly	finance	fin+ance
neglect	neg+lect	contradict	contra+dict
capture	cap+ture	venture	vent+ure
receive	re+ceive	community	commun+ity
seduce	se+duce	transmit	trans+mit
insult	in+sult	duplicate	du+plic+ate
oppress	op+press	introspection	intro+spect+ion
textile	text+ile	substitute	sub+stit+ute
direct	di+rect	expel	ex+pel
passage	pass+age	discover	dis+cover
otherwise	other+wise	basement	base+ment
central	centr+al	notify	not+ify
affluence	af+flu+ence	along	a+long
absurd	ab+surd	allocate	al+loc+ate
handful	hand+ful	reinforce	re+in+force
epidemic	epi+dem+ic	guardian	guard+ian
succeed	suc+ceed	amid	a+mid
recede	re+cede	innate	in+nate
novel	nov+el	descend	de+scend
principle	prin+cip+le	rectangle	rect+angle
airline	air+line	temperature	temper+ate+ure
festival	fest+ive+al	cultivate	cult+ive+ate
geometric	geo+metry+ic	climate	clim+ate
prosecute	pro+sec+ute		

Table II

Incorrect segmentation results

word	correct segmentation	result
continue	con+tin+ue	contin+ue
amity	ami+ty	a+mit+y
pastime	past+time	pas+time