Finely cutting the stem/suffix boundary using MDL

John Goldsmith

October 2003

Starting point…

- Familiarity with information theory
  - Information complexity of referring to an entity X is –log freq X
- Unsupervised learning of grammar…and in particular, of natural language morphology

- MDL Minimum Description Length
  - Goal of analysis is to maximize the probability of the data
  - Prob (data) = prob (data|model)*prob(model)
  - Prior probability distribution over models exponential in the length of the model in its minimal formulation

- MDL Minimum Description Length
  - …Prior probability distribution over models exponential in the length of the model in its minimal formulation
  - So minimize the sum:
    - log probability of data + length of model
- Can linguists *seriously* use the notion of *length of a grammar*? (Householder 1965, Chomsky and Halle 1965)

That’s what we’ll show…do.

- The Zellig Harris *successor frequency* suffix-finding bootstrapping algorithm is good, but far from perfect.
- Can MDL catch its errors?

Some errors on 250K words

- on & ve:
  - affirmati agressi attenti comprehensi conclusi decisi destructi evasi …15 more
- l & tion:
  - differentia inaugura
- NULL & rs
  - ringside teenage
- ous & ty
  - tenaci vivaci
- e & y > le & le > ble & bly
  - admirabl audibl conceivabl considerabl equitabl formidabl honorabl impeccabl impossibl incomparabl incredibl indelibl irredeemabl justifiabl notabl predictabl preferabl reasonabl remarkabl terribl unavoidabl (4 more)

Let us consider each signature σ

- And evaluate its description length;
- Then consider slicing each of its words 1,2,3, or 4 letters further to the left.
- We compute the grammar length of the signature(s) in each case, and choose the one with the smallest DL.

DL of a signature σ

- Sum of:

1. The description length of each stem in the signature (actual phonological substance)
2. The description length of the pointer to the suffix in the signature
3. The (prorated) *portion* of the phonological substance of the suffix
4. The length of all of the pointers to that signature σ found on each of its stems

ed.ing.s

- With stems *jump, walk*
  - Length of *jump*: 4 log(26)
- Length of pointer to *–ed*:
  -log freq (ed) =

$$-\log \frac{\#\, words\, ending\, in - ed}{\#\, analyzed\ words\, in\, corpus}$$

Entropy of the ends of the stems

- Measure how much variety there is among the last 1 (or 2,3,4) letters of the stems
- If there’s too much variety (= entropy), it’s unlikely that the varying material ought to be in the suffixes.
- Entropy threshold : 1.5

stem entropy for on.ve

Shift # letters: 1: Entropy sufficiently small: 0

Shift # letters: 2: Entropy sufficiently small: 0.987693 (why?)

Shift # letters: 3: Entropy too large: 3.23619 (Threshold 1.5.)

Shift # letters: 4: Entropy too large: 4.26269 (Threshold 1.5.)

suffix use by this signature:

+on use count: 26 DL: 7.685

Information for this suffix is owned by this sig in this proportion: 0.885 ; i.e. 8.316 bits

+ve use count: 23 DL: 7.862

Information for this suffix is owned by this sig in this proportion: 1.000 ; i.e. 9.401 bits

By the way…

This information is generated automatically by *Linguistica* when you turn on its log.

Length of pointers to this sig:	180.833
Current signature's DL:	214.098

Entropy tells us to consider moving 1 or 2 letters to the right

affirma
atten
co-opera
destruc
imagina
introspec
posi
provoca
recep
representa

“ti” cases...

tion and ***tive***

tion existed; old count was 15; New DL for this affix: 7.138

tive did not exist before; DL for this affix is 26.664

26.664 is a lot bigger, because this signature would have to pay for *all* of the new suffix.

- Pointers to this sig: 80.639
- That’s 10 times 8.0639 – one pointer for each of its stems.
- Total for this signature: 114.441bits

Now, ***sion*** and ***sive***

sion did not exist before; DL for this affix is 26.664

sive did not exist before; DL for this affix is 26.664

“**si”** cases

aggres
comprehen
conclu
deci
eva
exclu
expan
explo
indeci
percus
permis
persua
repres

sion.sive

Pointers to this sig: 99.910

Total for this sig: 153.239

So total for tion.tive and sion.sive:

267.680

compared to the original 214.098

That’s a loser…

Let’s add *one* letter to the suffixes

New signature: ion.ive

- ion existed; old count was 85; New DL for this affix: 5.631
- ive existed; old count was 5; New DL for this affix: 7.579

Nice!

New stems…

affirmat
aggress
attent
co-operat
comprehens
conclus
decis
destruct
evas
exclus
expans
explos

imaginat
indecis
introspect
percuss
permiss
persuas
posit
provocat
recept
representat
repress

Pointers to this sig: 157.833

Total for this sig: 171.042

That’s better than the original, which was 214.098

We’ve left out so far stem-content information

- There are two aspects of this:
  - As you shift material from the stems, each of them is shorter, and hence has a smaller information content;
- And if the new stem that is created is one that exists independently, then the new signature is responsible for only part of it, not all of it.

Both of these are important considerations.