Unsupervised Learning of the Morphology of a Natural Language

John Goldsmith, University of Chicago, 2001

https://www.aclweb.org/anthology/J01-2001.pdf

(presented by Vilém Zouhar, 2020)

Task

- “Given unannotated plaintext data, produce morphological annotations.”
- Doesn’t work well with languages with high average number of affixes
- Input: data sized 5KW-1MW
- Output: segmentation of every input word + categories of such segments (although the categories are alg. internal and don’t map to “our” classes)

- Automated tool > faster and cheaper than manual work
- Great for bootstrapping other grammar systems

- Two parts: splitting words + determining possible suffixes (+common classes)

p. 153-155

Minimum Description Length Model

- Compressing corpus ∝ Morphological analysis
- (best compression, but given a specific “morphological” structure)

First model

- At most one suffix
- signature = affix set
- *p(stem+aff.)* = *p(sig.)p(stem|sig.)p(aff.|sig.)*
- We want to target the lowest entropy distribution of *p*, so that it can be compressed effectively.

Signature 3:

$$\left\{\begin{array}{l}\textit{SimpleStem : ptr(jump)}\\ \textit{SimpleStem : ptr(laugh)}\\ \textit{SimpleStem : ptr(walk)}\end{array}\right\}\left\{\begin{array}{c}\textit{ptr(NULL)}\\ \textit{ptr(ed)}\\ \textit{ptr(ing)}\\ \textit{ptr(s)}\end{array}\right\}$$

p. 164, 166, 167

Recursive structure

- allow *{ {work}{-, -ing} }{-, -s}*
- but in this case add a flag to disallow *work-s*

MDL - motivation

- MDL cannot create morphologies itself
- But when we give it two morphologies M_1, M_2 we can say which one is better according to their probabilities (compressed sizes) in MDL (minimum description model)

Word Segmentation (0th) Heuristics

- Can suggest morphologies for MDL
- Based on Expectation-Maximization
- |*w*| stem+suff. hypothesis for a word *w* (cut at every index)
- Zeroth heuristic: normalize the cut probability as

$$\frac{pr(stem\ t = w_{1,i})pr(suffix\ f = w_{i+1,l})}{\sum_{k=1}^{N} pr(stem\ t = w_{1,k})pr(suffix\ f = w_{k+1,l})},$$

- This fails (creates suffixes sized 1)

p. 170, 171

1st Heuristic

- Named *take-all-splits*
- Models splits using a Boltzmann distribution

$$H(w_{1,i} + w_{i+1,l}) = -(i \log \mathit{freq}\,(\mathit{stem} = w_{1,i}) + (l - i) \log \mathit{freq}\,(\mathit{suffix} = w_{i+1,l})) \quad (4)$$

$$\mathit{prob}\,(w = w_{1,i} + w_{i+1,l}) = \frac{1}{Z} e^{-H(w_{1,i} + w_{i+1,l})} \quad (5)$$

$$Z = \sum_{i=1}^{n-1} H(w_{1,i} + w_{i+1,l})$$

- This promotes longer suffixes
- We now have split for every word

p. 171, 172

2nd Heuristic

- Count all suffixes of size 2-6
- (6 chosen because we don't expect suffixes larger than that)
- Model the probability of this being a morpheme as:

$$\frac{[n_1 n_2 \ldots n_k]}{\textit{Total count of k-grams}} \log \frac{[n_1 n_2 \ldots n_k]}{[n_1][n_2] \ldots [n_k]},$$

- We choose 100 most probable suffixes
- Errors *{-ting,-ing}*
- Erros *{de-}{-fense,-mand,-lete}*
- Many words will then obtain multiple different splits
  We can use MDL to choose the more probable ones

Intermediate Results 1 Cleanup

- MDL gives us the best parse for each word
- Stems and suffixes can be taken from this description
- Singleton signatures removed (>90%)
- Signatures of size 1 removed
- The rest is called regular (stems/suffixes)
- Example from *Tom Sawyer:*
  Signatures: *{-,-ed,-ing}, {-e,-ed,-ing}*

p. 173, 174

Intermediate Results Cleanup - errors

1. Two suffixes collapsed into one: *-ings, -ments*
2. Common stem endings in suffixes: *-**ts***
3. *-s* is a good suffix candidate, but not all words ending with *-s* have this suffix (in a morphological sense)
4. Same (morphological) stem has different (“infered”) stems: {abbreviate}{-,-d,-s} vs. {abbreviat}{-ing}
5. In the previous case, the stem is not split consistently. In case of {win}{-,-s} and {winn}{-er,-ing} we want the stems to be connected.

Q: What’s the difference between 4 and 5?

p. 174, 175

Cleanup - Procedure - 1, 2

- Modify the morphology and compare MDL's outputs
- If the compressed length is lower, then accept this change
- In 1 suffixes can be split into e.g. *{-ings} -> {{-ing}{-,-s}}* and checked in MDL
- This can be done by checking whether it is composed of two already existing suffixes
- 2 can be fixed by examining signature prefixes e.g. *{-te,-ting,-ts}*
  Try dropping *t* and if MDL decreases, then accept this
  (related problem "*what if -t is a suffix"* discussed later)
-

p. 176, 177

Cleanup - Procedure - 3 (triage)

- For 3 we take a look at suspicious suffixes: either with too low or too high number of stems. E.g. *{boo,loo,}{-t,-l}*
- Short suffixes are suspicious.
- How much data we need to say that a signature is plausible?
- Does adding the suffix back to the stem decrease MDL? If so, accept the change.

Q: Solution to 4, 5?

p. 177, 178, 185

Intermediate Results 2

- After implementing the cleanup solutions, we get somewhat better results
- *{-,-ed,-ing,-s}* (corresponds to verbs)
- *{-,-s}* (corresponds to nouns)

Table 10
Results (English).

Category	Count	Percent
Good	829	82.9%
Wrong analysis	52	5.2%
Failed to analyze	36	3.6%
Spurious analysis	83	8.3%

- Good = good
- Wrong = segmented, but wrongly
- Failed = not segmented
- Spurious = segmented, but shouldn't be

p. 178, 179, 183, 184

Endnotes

- This algorithm doesn't deal with relating signatures together
  (because given a signature S_1, there is likely to be $S_2 \subset S_1$ and S_1 may or not may be morphologically related to S_2)
- Allomorphs are disregarded.
- Compounds are disregarded.
- Substractive morphemes are work in progress.
  We want *{-e,-ed,-es,-ing}* to be *{-,-ed,-ing,-s}* for *lov/love*
  But for this we would need a deletion operator to do *love-<x>-ing*

p. 186-189