LIN570: HW3 – morphological segmentation (1/2) (100pts)

YOUR NAME (`UW NetID`)

Due date: 11pm on Oct 22, 2018 (Tuesday)

For this homework, you are going to reproduce morphological segmentation inside-out results (Cotterell et al., 2016) and (partially) compound splitting (Koehn and Knight, 2003). All the required files are under `~/dropbox/19-20/570/hw3/examples`.

Rubric:

2pts `hw.tar.gz` submitted, it should contain following files:

- `convert_[hierarchy|flat].sh` (for Q1a)
- `convert_leaf.sh` and `test0.leaf` (for Q1b)
- two parsing models (hierarchy and flat) (for Q1b)
- `test0.[hierarchy|flat].parsed` and `test0.[hierarchy|flat].score` (for Q1b)
- `frequency_metric_german.sh` and `file.score` (for Q2)

2pts `readme.txt` or `readme.pdf` submitted

6pts All files and folders are present in expected locations

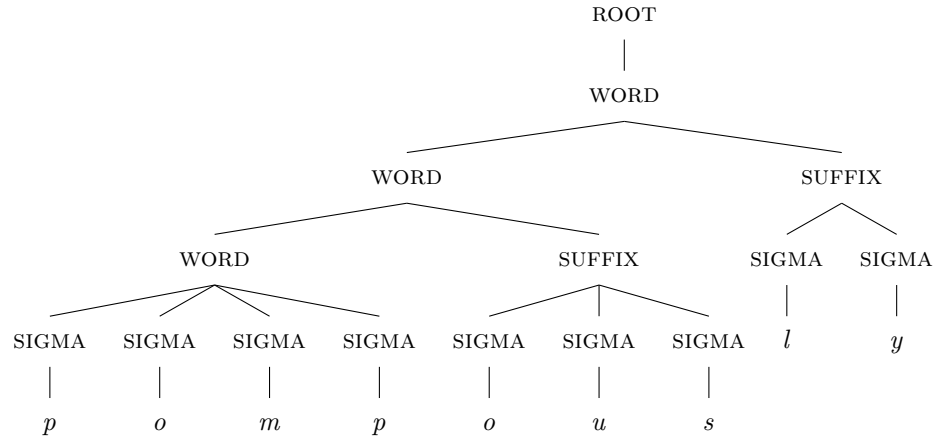
10pts Programs run to completion

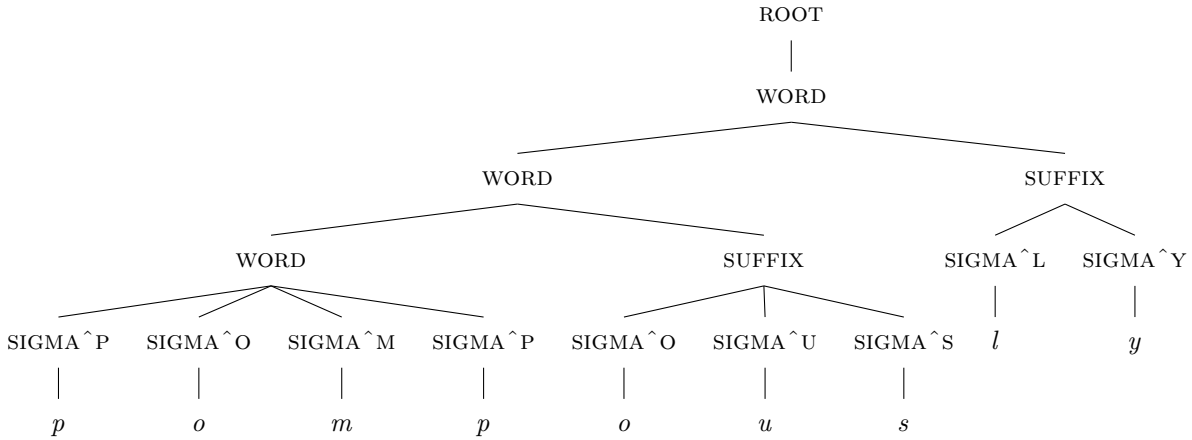
5pts The output of programs on patas match submitted output

1. (**45pts**) Reproduce morphological segmentation inside-out results (Cotterell et al., 2016)[1]

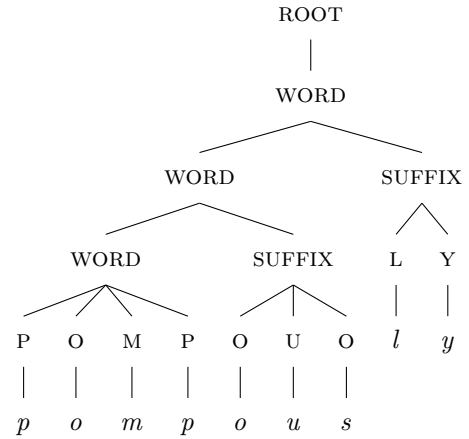
   - **corpus, original** `(pompously (S (S (pomp:stem) (ous:suffix)) (ly:suffix)))`
     corpus, treebank-like `(S (S (STEM pomp) (SUFFIX ous)) (SUFFIX ly))`
     parsing, hierarchy `(ROOT (WORD (WORD (WORD (SIGMA p) (SIGMA o) (SIGMA m) (SIGMA p)) (SUFFIX (SIGMA o) (SIGMA u) (SIGMA s))) (SUFFIX (SIGMA l) (SIGMA y))))`

[1] 10-fold data split available at `https://ryancotterell.github.io/data/splits.tar.gz`





Alternatively,



parsing, flat `(ROOT (WORD (WORD (SIGMA p) (SIGMA o) (SIGMA m) (SIGMA p)) (SUFFIX (SIGMA o) (SIGMA u) (SIGMA s)) (SUFFIX (SIGMA l) (SIGMA y))))`

(a) (20pts) convert the data set for parsing (hierarchy and flat). use only `train0, dev0, test0`: `*.hierarchy.penn` and `*.flat.penn`
   - *i.e.* `cat train0 | ./convert_hierachy_penn.sh > train0.hierarchy.penn`
   - `cat train0 | ./convert_flat_penn.sh > train0.falt.penn`

(b) (25pts) train using the Berkeley parser[2] (hierarchy and flat models) (`train0, dev0` for training) and evaluate results using EVALB[3]
   - train:

     ```
     java -cp berkeleyParser.jar edu.berkeley.nlp.PCFGLA.GrammarTrainer \
     -path hierarchy.penn -out hierarchy.model \
     -treebank SINGLEFILE
     ```

     where `hierarchy.penn` is the concatenation of `train0.hierarchy.penn` and `dev0.hierarchy.penn`
   - for test, you will need `convert_leaf.sh` which produces `test0.leaf`. It contains the input for parsing as follows:

     ```
     t h o u g h t l e s s n e s s
     u n i d e n t i f i e d
     ```
   - `cat test0.hierarchy.penn | ./convert_leaf.sh > test0.leaf`
   - parse `test0.leaf` using the trained model (hierarchy and flat models),
   - and evaluate it useing EVALB

2. (**30pts**) Reproduce a frequency based metric using the monolingual German corpus in described in Koehn and Knight (2003):

$$\underset{S}{\operatorname{argmax}} \left(\prod_{p_i \in S} C(p_i) \right)^{\frac{1}{n}} \tag{1}$$

- *Aktionsplan*:
  - aktionsplan: $C(\text{aktionsplan}) = 852 \rightarrow 852$ (KO)
  - aktion - plan: $C(\text{aktion}) = 960, C(\text{plan}) = 710 \rightarrow 825.6$
  - aktions - plan: $C(\text{aktions}) = 5, C(\text{plan}) = 710 \rightarrow 59.6$
  - akt - ion - plan: $C(\text{akt}) = 224, C(\text{ion}) = 1, C(\text{plan}) = 710 \rightarrow 54.2$
- *Freitag*:
  - freitag: $C(\text{freitag}) = 556 \rightarrow 556$
  - frei - tag: $C(\text{frei}) = 885, C(\text{tag}) = 1864 \rightarrow 1284.4$ (KO)
- About 30% of compounds require a connector between the combined words. These are most commonly **-n-**, **-en-**, **-s-**, **-es-** and sometimes **-e-**.
  - split up to only 3 words
  - avoid one-letter morphs

    freitag *fr eitag*, *fre itag*, *fr itag*, *fre itag*, ... (split into 2 words)

    fr eitag ***fr*** *ei tag*, ***fr*** *ei tag*, ***fr*** *it ag*, ... (split into 3 words)

[2] `https://github.com/slavpetrov/berkeleyparser`
[3] `https://nlp.cs.nyu.edu/evalb/`

  - segmentation using cky and bio: `https://www.overleaf.com/read/sctybppbmtqt` ()

- use `europarl-v7.de-en.true.de.gz`
- `cat file.txt | ./frequency_metric_german.sh > file.score`
- to display utf8 characters correctly in patas:

```
export LC_ALL=en_US.UTF-8
export LANG=en_US.UTF-8
export LANGUAGE=en_US.UTF-8
```

- `cat file.score`

```
aktionsplan =  aktionsplan (852) = 852
aktion plan =  aktion (960)  plan (710) = 825.6
aktions plan = aktions (5)  plan (710) = 59.6
akt ion plan = akt (224) ion(1) plan (710) = 54.2
...
```

References

Cotterell, R., Kumar, A., and Schütze, H. (2016). Morphological Segmentation Inside-Out. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2330, Austin, Texas. Association for Computational Linguistics.

Koehn, P. and Knight, K. (2003). Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 187–194.