Morfessor 2.0: Toolkit for Statistical Morphological Segmentation – Model

Probabilistic Model definition

- Full description in (Virpioja, 2012; Virpioja et al., 2013)
- Generative model
  $p(\underset{analyses}{A}, \underset{words}{W} \mid \underset{parameters}{\theta})$
  The model generates pairs of words and analysis (the segmentation of a word into morphs)
- Tokenization function
  $a = \phi(w; \theta)$
- Cost derivation

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta) p(\underset{data}{D} \mid \theta)$$

$$L(\theta, D) = -\log \underset{prior}{p(\theta)} - \log \underset{data\ likelihood}{p(D \mid \theta)}$$

The data (D) is a list of (non-segmented) words to learn the model from in unsupervised manner.

Data Likelihood

$$\log p(D \mid \theta) = \sum_{j=1}^{N} \log p(W = w_j \mid \theta)$$

$$= \sum_{j=1}^{N} \log \sum_{a \in \Phi(w_j)} p(A = a \mid \theta),$$

Morfessor Baseline assumes independence of words. Also, only valid tokenisations of need to be considered. Morfessor selects only one tokenisation (analysis) for each word at a time, by introducing a hidden variable $\mathbf{Y}$.

$$\log p(D \mid \theta, \mathbf{Y}) = \sum_{j=1}^{N} \log p(y_j \mid \theta)$$

$$= \sum_{j=1}^{N} \log \underset{selected\ analysis}{p(m_{j1}, \ldots, m_{j|y_j|}, \#_w \mid \theta)}$$

Prior

(Creutz and Lagus, 2007) The parameters of Morfessor Baseline encode the properties of the morph lexicon:

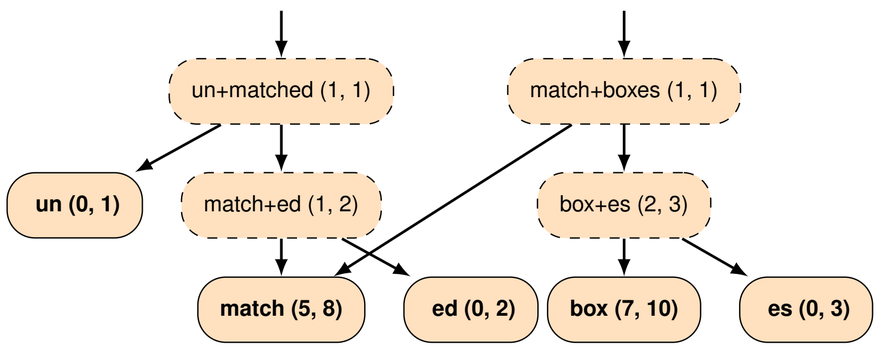
$$p(\theta) = p(\underset{\#morphs}{\mu}) \times \underset{\#morph\ permutations}{\mu!} \times p(\mathrm{properties}(m_1), \ldots, \mathrm{properties}(m_\mu)).$$

$$p(\underset{morph}{\sigma_i}) = \underset{length\ prior}{p(L = |\sigma_i|)} \prod_{j=1}^{|\sigma_i|} \underset{character\ distribution}{p(C = \sigma_{ij})}$$

Algorithm

(Creutz and Lagus, 2002)

function LOCALBATCHTRAIN(D, ϵ)
  $\theta, \mathbf{Y} \leftarrow$ INITMODEL(D_W)
  $L_{\mathrm{old}} \leftarrow \infty$
  $L_{\mathrm{new}} \leftarrow L(D_W, \theta, \mathbf{Y})$
  while $L_{\mathrm{new}} < L_{\mathrm{old}} - \epsilon$ **do**
    $J \leftarrow$ RANDOMPERMUTATION($1, \ldots, N$)
    for $j \in J$ **do**
      $\theta, \mathbf{Y} \leftarrow$ LOCALSEARCH($w_j, D, \theta, \mathbf{Y}$)
    $L_{\mathrm{old}} \leftarrow L_{\mathrm{new}}$
    $L_{\mathrm{new}} \leftarrow L(D, \theta, \mathbf{Y})$
  return θ, $\mathbf{Y}$



unmatched, **matchboxes**, **matched**, **boxes**, **match**, and **box**

Likelihood weighting and Semi-supervised training

Likelihood weighting with α (Virpioja et al., 2011)

$$L(\theta, D) = -\log p(\theta) - \alpha \log p(D \mid \theta)$$

α can be determined in different ways, e.g using a development set, or some explicit knowledge like average morph length. Higher α reduces segmentation, lower α increases segmentation.

Semi-supervised (Kohonen et al., 2010)

$$L(\theta, D) = -\log p(\theta) - \alpha \log p(D \mid \theta) - \beta \log p(\underset{annotated\ data}{D_A} \mid \theta).$$

For semi-supervised learning another term is added to the cost, the likelihood of a set of annotations coming from the model. Also here a weight β is introduced to control the effect.

References

Creutz, M. and Lagus, K. (2002). Unsupervised discovery of morphemes. In Maxwell, M., editor, *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30, Philadelphia, PA, USA. Association for Computational Linguistics.

Creutz, M. and Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1):3:1–3:34.

Kohonen, O., Virpioja, S., and Lagus, K. (2010). Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78–86, Uppsala, Sweden. Association for Computational Linguistics.

Virpioja, S. (2012). *Learning Constructions of Natural Language: Statistical Models and Evaluations*. PhD thesis, Aalto University.

Virpioja, S., Kohonen, O., and Lagus, K. (2011). Evaluating the effect of word frequencies in a probabilistic generative model of morphology. In Pedersen, B. S., Nešpore, G., and Skadiņa, I., editors, *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, volume 11 of *NEALT Proceedings Series*, pages 230–237. Northern European Association for Language Technology, Riga, Latvia.

Virpioja, S., Smit, P., Grönroos, S., and Kurimo, M. (2013). Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, Aalto University, Finland.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos and Mikko Kurimo
Aalto University