Морфемная сегменатция

Гусев Илья

Московский физико-технический институт

Москва, 2018

Содержание

Задача морфемной сегментации

- По слову нужно получить его разбиение на морфемы.
- Мотивация - обработка редких и не встретившихся в обучющей выборке слов для раличных задач NLP.
- Разбиение на морфемы можно использовать вместо BPE (byte pair encoding), морфемы обычно имеют какой-то смысл.

Отступление: byte pair encoding

- Считаем самую частотную пару символов одним символом
- Повторяем, пока все не будут встречаться по одному разу или пока за каждым символом $<$ N терминалов
- Мотивация - сильно уменьшаем размер словаря по сравнению с word-level, но при этом это лучше, чем char-level в некоторых случаях

Пример:

1. aaabdaaabac
2. ZabdZabac
3. ZYdZYac
4. XdXac

Задача морфемной сегментации

Примеры

забытье	забы*ть*е
статичный	стат*ич*н*ый
учитель	уч*и*тель
скрыться	скры*ть*ся
тысячами	тысяч*ами
подержаться	по*держ*а*ть*ся

Morfessor

Data likelihood 1

W - слова, $w \in W$
A - анализы слов, $a \in A$, $a = \phi(w; \theta)$, $a = (m_1, ..., m_n)$
D_W - обучающая выборка, $|W| = N$, $\#_w$ - границы между словами (для анлиза сложных слов)
θ - параметры модели
$\Phi(w) = \{a : \phi^{-1}(a) = w\}$

$$\theta_{map} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|D_W) = \underset{\theta}{\operatorname{argmax}}\, p(\theta)p(D_W|\theta)$$

$$L(\theta, D_W) = \log p(\theta) - \log p(D_W|\theta)$$

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

Morfessor

Data likelihood 2

Y - скрытая переменная, сопоставляющая $\forall w_j \to \Phi(w_j), Y = (y_1, \dots, y_N)$

$$\log p(D_W|\theta, Y) = \sum_{j=1}^{N} \log p(y_j|\theta) = \sum_{j=1}^{N} \log p(m_{j_1} \dots m_{j_{|y_j|}}, \#_w|\theta)$$

$$\log p(D_W|\theta, Y) = \sum_{j=1}^{N} (\log p(\#_w|\theta) + \sum_{i=1}^{|y_j|} \log p(m_{ji}|\theta))$$

Morfessor

Prior

$p(\theta) = p(L)$
Рассматриваем только такие m_i: $p(m_i|\theta) > 0$

$$p(L) = p(\mu) * p(properties(m_1), \ldots, properties(m_\mu)) * \mu!$$

Ниже m_i рассматривается как последовательность неделимых элементов (букв, спец. символов), поэтому мы обозначим её как σ_i

$$p(\sigma_i) = p(L = |\sigma_i|) \prod_{j=1}^{|\sigma_i|} p(C = \sigma_j)$$

Кроме того, можно включить prior по количеству использований различных сегментов: $p(m_i|\theta) = \tau_i/(N + \nu)$.

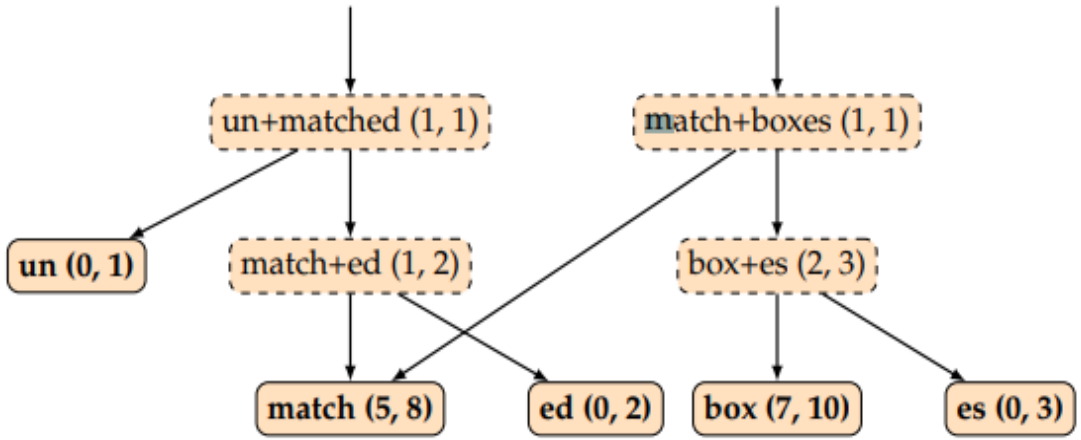
Morfessor

Обучение

- Foward-backward аналогично HMM (вариация EM-алгоритма)
- Global Viterbi аналогично HMM
- Local Viterbi
- Recursive Baseline (жадный поиск)

$$y_j^{(t)} = \underset{y_j \in Y_j}{\operatorname{argmin}} \{ \min_{\theta} L(\theta, Y^{t-1}, D_w) \}$$

$$\theta^{(t)} = \underset{\theta}{\operatorname{argmin}}\, L(\theta, Y^t, D_w)$$



Morfessor

Метрика

$$precision = \frac{number\ of\ correct\ boundaries\ found}{total\ number\ of\ boundaries\ found}$$

$$recall = \frac{number\ of\ correct\ boundaries\ found}{total\ number\ of\ correct\ boundaries}$$

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Для английского примерно 0.77 в unsupervised и 0.86 в semi-supervised

Другие подходы

- MORSE
- Seq2seq

Полезные ссылки I


Morfessor2.0:Python Implementation and Extensions for Morfessor Baseline
https://bit.ly/2QAzWtW


Morfessor 2.0: Toolkit for statistical morphological segmentation
https://www.aclweb.org/anthology/E14-2006


MORSE: Semantic-ally Drive-n MORpheme SEgment-er
https://arxiv.org/abs/1702.02212


Morphological Segmentation with Sequence to Sequence Neural Network
http://www.dialog-21.ru/media/4287/arefyevnv.pdf


Use of morphology in distributional word embedding models: Russian language case
http://www.dialog-21.ru/media/4260/sadov_kutuzov.pdf

Полезные ссылки II

- Morphessor 2.0 demo
  https://asr.aalto.fi/morfessordemo/
- Morphessor 2.0
  https://github.com/aalto-speech/morfessor
- morpheme_seq2seq
  https://github.com/kpopov94/morpheme_seq2seq
- XMorphy
  https://github.com/alesapin/XMorphy