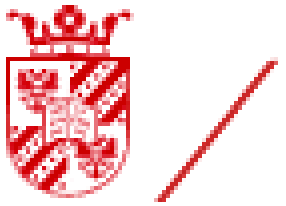
Corpus Cross Entropy

Peter Nabende

Rahmad Mahendra



university of groningen

Outline

- Entropy
- Cross Entropy
- Machine Transliteration
- Algorithms and Models
- Corpus Cross Entropy
- References

university of groningen

Entropy (1)

- Measurement in information theory
- Randomness, uncertainty
- Expected value function of information content in random variable
- Based on [Shannon,1948]

university of groningen

Entropy (2)

- Suppose a set of events whose probabilities of occurence $p_1, p_2, p_3, \ldots, p_n$
- $H(p_1, p_2, p_3, \ldots, p_n)$ that satisfying following properties
  1. H should be continuous in the p_i
  2. If all the p_i are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n
  3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

university of groningen

Entropy (3)

$$H(x) = -K\sum_{i=1}^{n} p(x_i).\log\, p(x_i)$$

K is positive constant

university of groningen

Entropy: Language Model Case (1)

- Suppose sequence of tokens, i.e. words (`w1, w2, w3, ..., wn`) we have entropy

$$H(w_1, w_2, ..., w_n) = -\sum_{W_i^n \in L} p\left(W_i^n\right).\log p\left(W_i^n\right)$$

- For each word, entropy rate

$$\frac{1}{n}.H(w_1, w_2, ..., w_n) = -\frac{1}{n}.\sum_{W_i^n \in L} p\left(W_i^n\right).\log p\left(W_i^n\right)$$

university of groningen

Entropy: Language Model Case (2)

- If a language is stationary and ergodic, the Shannon-McMillan-Breiman theorem

$$\frac{1}{n}.H(w_1, w_2, ..., w_n) = -\frac{1}{n}.\sum_{W_i^n \in L} \log p\left(W_i^n\right)$$

- A language is stationary if the probability distribution of the words do not change with time.
- A language is ergodic if its statistical properties can be deduced from a single, sufficiently long sequence of words

university of groningen

Cross Entropy (1)

- Comparing probability distribution.
- Kullback-Leibler information measure, Relative entropy
- Cross entropy of two probability distribution p and m for a random variable X

$$H(p,m) = -K.\sum_i p(x_i).\log m(x_i)$$

university of groningen

Cross Entropy (2)

- Not symmetric function

  $H(p,m) \neq H(m,p)$

- The cross entropy $H(p,m)$ is a upper bound on true entropy p

  $H(p,m) \geq H(p)$

- Used to compare approximate model
  - Between two model and , one whose lower cross entropy value considered as more accurate model

university of groningen

Machine Transliteration

- Translating names and/or technical term across languages with different alphabets and sound inventories
- P.Nabende examine data Russian – English
  - барбадос barbados
  - луксор luxor
  - линкольн lincoln

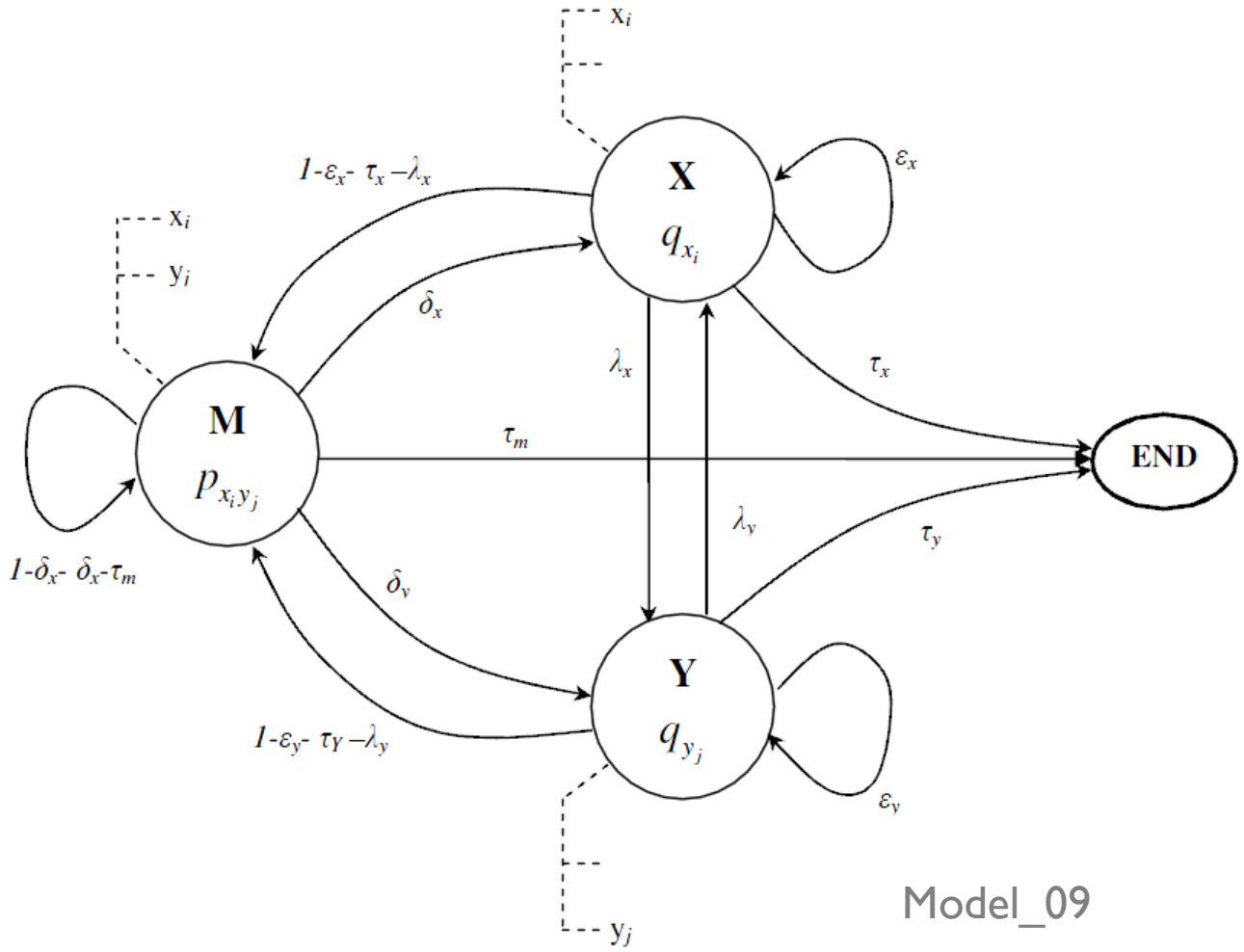
university of groningen

Algorithms and Models (1)

- Hidden Markov Model
  - Forward
  - Viterbi
- Model_06 and Model_09

university of groningen

Algorithms and Models (2)



Model_09

university of groningen

Algorithms and Models (3)

- Input of algorithm: two observation sequences (`s,t`)
- Output of algorithm: similarity score

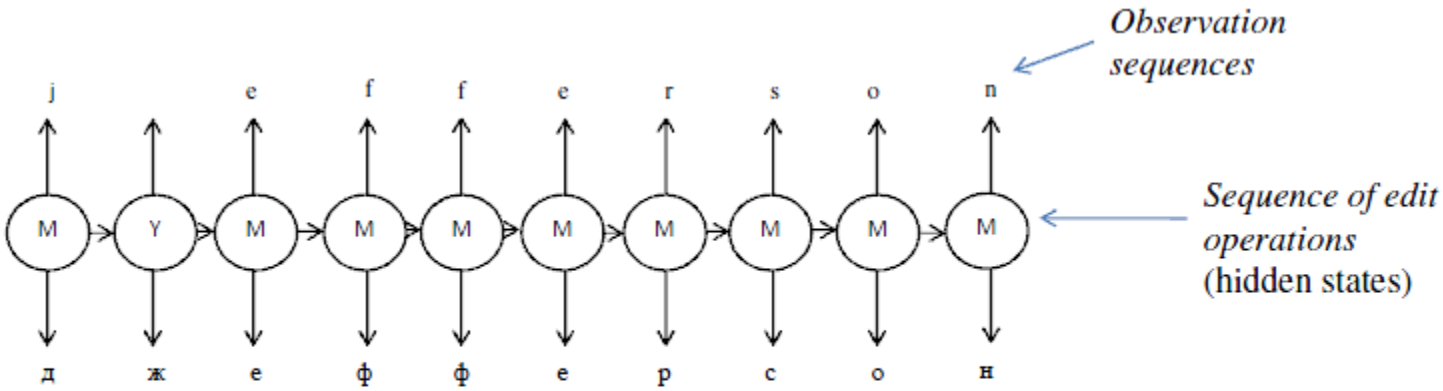


Fig.1: Illustration of alignment for an english name "jefferson" and its Russian transliteration following the pair-HMM concept

university of groningen

Corpus Cross Entropy (1)

- Given a corpus consisting *n* pair English-Russian data with similarity score, suppose as token `ci`

  (`c1, c2, c3, ..., cn`)

- computing cross entropy across the algorithm and models

$$H_c(m) = -\frac{1}{n}.\sum_i \log m(c_i)$$

Corpus Cross Entropy (2)

Result of Computing Corpus Cross Entropy for two algorithms, two models of 743 pairs English - Russian

	Forward	Viterbi
Model_06	6.04955661824	21.73845715804
Model_09	12.89473170688	19.65702544689

university of groningen

Corpus Cross Entropy (3)

- Interpretation
  1. Forward algorithm is more appropriate to be used
  2. Model_06 is considered more accurate (in case of Forward algorithm)

university of groningen

References

- P.Nabende (2009) *Cross Entropy for Measuring Quality in Models* http://www.let.rug.nl/~nerbonne/teach/rema-stats-meth-seminar/presentations/Nabende_x_entropy_model_accuracy_2009.pdf
- C.E. Shannon (1948) "A Mathematical Theory of Communication". *Bell System Technical Journal*
- D.Jurafsky & J.Martin (2009) *Speech and Language Processing*.

university of groningen