# Machine Discovering Assignment 1-1

Tung-Chun, Chiang R05922027

October 3, 2016

## 1 Problem description

- Given:
  - character-level bigram language model
  - the probabilities of encoding a character to another
  - encoded text
- Goal:
  - the original text before encoded (decoding problem)

## 2 Assumptions

- Assumptions:
  - In the original text, a character only depends on its last character.
  - An encoded character only depends on its original character.
  - Whitespaces are always encoded to whitespaces.

## 3 Graphical model

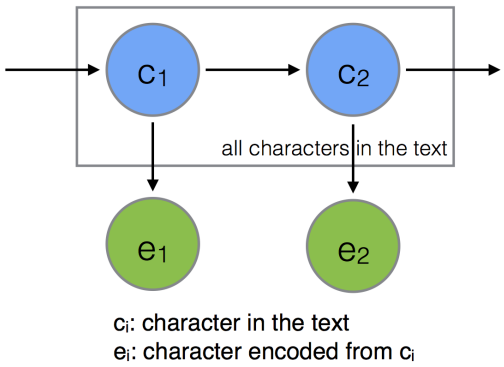


Figure 1: Graphical model

# 4 Inference

Based on our graphical model, a word only depends on its last word. Also, whitespaces are always encoded to whitespaces. We can split the whole encoded text into segments(words) with whitespaces. So, we can decode each segment separately.

## 4.1 Viterbi algorithm

According to our graphical model, our goal is to find a sequence $C$ to maximize the posterior given the evidence sequence $E$:

$$\begin{aligned} MPA(C|E) &= \underset{C}{\operatorname{argmax}} P(C|E) \\ &= \underset{C}{\operatorname{argmax}} \frac{P(E|C)P(C)}{P(E)} = \underset{C}{\operatorname{argmax}} P(E|C)P(C) \\ &= \underset{C}{\operatorname{argmax}} P(c_1|whitespace)P(e_1|c_1) \sum_{i=2}^{N} P(c_i|c_{i-1})P(e_i|c_i) \end{aligned}$$

For each segment, we assume it start from a whitespace (including the first segment) and use Viterbi algorithm to decide the decoded text. Viterbi algorithm is a kind of dynamic processing algorithm, and it works efficiently.

## 4.2 n-best Viterbi algorithm

Different from original Viterbi algorithm, we save top $k$ candidate paths. But we can't use English wordlist, so we only choose the top one as result. If we can use wordlist, we can find the top English word as result. If there is no English word in top $k$ candidates, we choose the top segment as answer.

# 5 References

**Viterbi Algorithm:** `https://en.wikipedia.org/wiki/Viterbi_algorithm`