Laboratorio di Teoria dell'Informazione

Incertezza e entropia di un dizionario

Obiettivo: valutazione di misure di informazione a partire da un dizionario di parole in un data lingua (esempio inglese o italiano).

Parte 1. Modello ad albero (Markov) delle parole del dizionario

Si consideri un dizionario *D* contenente un numero N di termini, ciascuno costituito da un numero variabile di caratteri. Data ogni sequenza di caratteri $c_1c_2c_3$ .... corrispondente a una parola del dizionario costruire un modello ad alberi multipli in cui:

- ogni radice (nodo di livello $l = 1$) rappresenta un possibile carattere iniziale c_1 di parola;
- ogni nodo g_l dell'albero a livello $l > 0$ rappresenta un possibile carattere c_l in posizione l-esima;
- ogni nodo g_l dell'albero a livello $l > 0$ è connesso da un arco verso un nodo g_{l+1} al livello successivo $l + 1$ se esiste almeno una parola che contiene la coppia di caratteri c_lc_{l+1} nelle rispettive posizioni;
- ogni percorso nell'albero da una radice verso una foglia rappresenta una parola del dizionario;
- ogni nodo dell'albero porta l'informazione relativa alla carattere che rappresenta e un contatore che rappresenta il numero di parole che passano o terminano in tale nodo.

Parte 2. Misure di informazione

A partire dai contatori delle occorrenze dei caratteri contenuti nei nodi dell'albero misurare le seguenti grandezze:

1. Entropia del primo carattere $H(c_1)$
2. Dato un percorso nell'albero, ovvero una parola, si misuri $H(c_{l+1}|c_l)$ per alcuni livelli dell'albero e se ne spieghi teoricamente l'andamento

$$H(c_{l+1}|c_l) = \sum_x P(c_l = x)H(c_{l+1}|x)$$

Parte 3. Tastiera predittiva

L'osservazione dell'andamento dell'entropia condizionata al precedente punto 2 permette di prevedere l'efficacia di tastiere di tipo predittivo. Si utilizzi la struttura ad albero per simulare una tastiera che in seguito all'inserimento di ogni carattere di un dato termine:

1. Valuta la probabilità di ogni carattere successivo a quello inserito dall'utente (percorrendo il relativo percorso nell'albero)

2. Mostra i k caratteri successivi più probabili (in ordine decrescente di probabilità)
3. L'utente viene così agevolato nella scelta del carattere successivo.

Usando i passi precedenti procedere iterativamente nel suggerimento del prossimo carattere durante l'inserimento di una parola: l'utente inserisce il primo carattere e si suggerisce di conseguenza il secondo, quindi sceglie il secondo carattere e il sistema suggerisce il terzo, e così via.

Parte 4. Calcolo dell'entropia del dizionario

Utilizzando il modello ad albero si calcoli l'entropia delle parole del dizionario, misurata in bit per parola. A tale scopo è conveniente scrivere l'entropia usando la cosiddetta chain rule:

$$H(D) = H(c_1 c_2 c_3 \dots) = \sum_{l=1} H(c_l | c_{l-1} c_{l-2} \dots c_1)$$

$$= H(c_1) + H(c_2|c_1) + H(c_3|c_2 c_1) + \dots$$

Si commenti il risultato ottenuto confrontandolo con il costo in bit di una codifica a lunghezza fissa di tutte le parole inserite nel dizionario.