Università di Padova
Dipartimenti di Studi Linguistici e Letterari

Tecnologie per la Traduzione 2020/2021

Codifica dell’informazione

Giorgio Maria Di Nunzio

Obiettivi

- Comprendere le idee alla base della codifica dei caratteri.
- Capire l’importanza della definizione di uno standard per la codifica dell’informazione.
- Conoscere i principali standard per la codifica dei caratteri.

Informazioni nel mondo reale

- L’informazione può presentarsi in forme molto diverse tra loro
  - Il valore numerico di una grandezza fisica
  - Il testo di un articolo di giornale
  - Il suono prodotto da uno strumento musicale
  - L’immagine di una fotografia
  - Le sequenze video di una ripresa televisiva

Codifica dell’informazione

- Per tutte queste forme è possibile trovare una rappresentazione adeguata per elaborare automaticamente l’informazione.

- La rappresentazione viene comunemente definita codifica.

- In informatica, la codifica viene effettuata applicando lo stesso principio di rappresentazione alle diverse forme.

Tipologie di informazione

- Informazione analogica
  - Rappresentazione di una grandezza (fisica) tramite una sua analogia.
  - Insieme infinito di variazioni.
- Informazione digitale
  - Rappresentazione attraverso un insieme finito di elementi.

Unità di misura dell'informazione

- Definizione di bit
  - Il bit, da binary digit, è l'unità di informazione.
  - Il bit è la minima informazione che può essere rappresentata ed elaborata da un computer.
- Definizione di byte
  - Un byte è una sequenza di 8 bit (256 valori diversi).
  - Il byte è l'unità di misura più frequentemente utilizzata in informatica.

Codifica dei numeri

- Si utilizza il sistema binario.
- Esempio di codifica dei numeri "con la virgola”
  - http://it.wikipedia.org/wiki/IEEE_754

Codifica del testo 1/3

- La codifica di informazione testuale ha seguito rapidamente la codifica dei numeri.
  - Il linguaggio testuale consente di interagire con i calcolatori in modo più naturale per l'utente.
  - La gran parte dei documenti prodotti dalla società, attuale e del passato, sono in forma testuale.
- La codifica dell'informazione testuale distingue la rappresentazione del contenuto con quella del formato.
  - Il contenuto è dato dalla successione di parole che costituiscono il documento e che ne rappresentano la semantica.
  - Il formato è legato al modo in cui le parole sono organizzate e rappresentate, fornendo informazioni accessorie che evidenziano alcuni concetti senza alterarne la semantica.

Codifica del testo 2/3

- Il contenuto di un testo è formato da una successione di parole, delimitate da spazi e da segni di interpunzione.
    - Il testo scritto è organizzato in forma lineare.
- Nelle lingue basate su di un alfabeto, le parole sono formate da sequenze di simboli presi da un insieme noto a priori.
    - I simboli sono normalmente le lettere dell'alfabeto, i numeri e altri caratteri a stampa di uso comune.
    - Gli alfabeti delle diverse lingue presentano delle differenze.
    - L'inglese non ha le vocali accentate (sono molto rare).
    - L'italiano e l'inglese non hanno le lettere: ñ - ç - ß
    - Altre lingue, ad esempio il russo e il greco hanno alfabeti totalmente diversi dall'alfabeto latino.

Codifica del testo 3/3

- Le prime codifiche del testo sono state fatte negli Stati Uniti, per questo spesso si fa implicitamente riferimento all'alfabeto inglese.

- I possibili elementi di un testo vengono divisi in:
  - Caratteri alfanumerici
  - I segni di interpunzione e di delimitazione (lo spazio è un carattere?)
  - Alcuni altri segni grafici (&, $, @, #, ~)
  - Caratteri speciali

- I caratteri alfanumerici sono quelli condivisi dalla maggior parte delle lingue (del mondo occidentale).

Codifica dei caratteri 1/2

- Poiché il contenuto di un testo può essere rappresentato da una sequenza di caratteri, per la sua rappresentazione è sufficiente codificare i caratteri con opportune sequenze di bit.

- Il concetto alla base della codifica di caratteri è che questi

    - Appartengono ad un insieme predeterminato di simboli, denominato alfabeto

    - L’alfabeto ha dimensione finita

- Ad ogni carattere quindi può essere univocamente associato un numero intero, che ne rappresenta il codice

    - L’associazione deve essere fatta per tutti i possibili simboli

    - I simboli che non vengono associati ad un numero (una sequenza di bit) non possono essere rappresentati

    - E’ necessario porre particolare attenzione alla scelta dell’alfabeto

Codifica dei caratteri 2/2

- Una volta fatta l'associazione, arbitraria, tra simboli dell'alfabeto e numeri è sufficiente utilizzare la codifica dei numeri interi.

- Un testo, che è una sequenza di caratteri, può quindi venire rappresentato da una sequenza di numeri interi.

- La scelta del numero di bit influenza la dimensione dell'alfabeto

  - Standard ASCII, 7bit = 128 simboli

  - Extended ASCII, 8 bit = 256 simboli

  - Unicode, 16 bit = 65.536 simboli

1963 - Codifica ASCII

- Uno degli alfabeti di codifica di caratteri più diffusi:
  - ASCII - American Standard Code for Information Interchange
- ASCII è uno schema di codifica di caratteri introdotto nel 1963 (X3.4-1963 di American Standards Association (ASA)) e in grado di rappresentare 128 diversi caratteri utilizzando 7 bit per rappresentarli e memorizzarli (2^7= 128).
- Questo schema è poi stato recepito dall'ISO (International Organization for Standardization) e identificato con la sigla ufficiale ISO 646-1972 o in breve ISO-7.

Tabella ASCII

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

1987 - ASCII Esteso

- La codifica ASCII negli anni è stata estesa per incorporare anche altri simboli, quindi si è passati da una rappresentazione a 7 bit ad una a 8 bit, con la possibilità di rappresentare 256 diversi caratteri

- I primi 128 caratteri rimangono immutati per consentire la compatibilità con la codifica ASCII a 7 bit. I restanti 128 caratteri permettono di rappresentare i caratteri particolari di altre lingue europee: ad esempio caratteri specifici dell'italiano o del francese, del tedesco, e l'alfabeto greco

- I diversi schemi estesi di codifica ASCII fanno parte della serie degli standard ISO/IEC 8859 del 1987 (IEC = International Electrotechnical Commission), chiamata anche Extended ASCII

1991 - Unicode

- Lo schema di codifica Unicode utilizza 16 bit (2^{16} = 65.536) per rappresentare ciascun carattere

- Questo schema è stato introdotto dal Consorzio Unicode a partire dal 1991 per superare le limitazioni dello schema ISO-7 e successivi

- Il sottoinsieme di caratteri Unicode rappresentati con 8 bit sono compatibili con le precedenti codifiche ISO

- Le versioni più recenti di Unicode estendono lo schema e fanno uso di 32 bit (232 = 4.294.967.296) per rappresentare ciascun carattere

- Al fine di risparmiare memoria, sono stati definiti diversi schemi di codifica, chiamati Unicode Transformation Format (UTF), per rappresentare i caratteri Unicode in modo compatto. Fra questi, il più utilizzato è UTF-8 (quando è possibile si utilizzano 8 bit per risparimare spazio).