Information Theory

CSCI 394: Computational Linguistics

CSCI 394

October 17, 2013

Entish

'Hoom, hmm! Come now! Not so hasty! You call yourselves hobbits? But you should not go telling just anybody. You'll be letting out your own right names if you're not careful.'
'We aren't careful about that,' said Merry. 'As a matter of fact I'm a Brandybuck, Meriadoc Brandybuck, though most people call me just Merry.'
'And I'm a Took, Peregrin Took, but I'm generally called Pippin, or even Pip.'
'Hm, but you are hasty folk, I see,' said Treebeard. 'I am honoured by your confidence; but you should not be too free all at once. There are Ents and Ents, you know; or there are Ents and things that look like Ents but ain't, as you might say. I'll call you Merry and Pippin if you please— nice names. For I am not going to tell you my name, not yet at any rate.' A queer half-knowing, half-humorous look came with a green flicker into his eyes. 'For one thing it would take a long while: my name is growing all the time, and I've lived a very long, long time; so my name is like a story. Real names tell you the story of the things they belong to in my language, in the Old Entish as you might say. It is a lovely language, but it takes a very long time to say anything in it, because we do not say anything in it, unless it is worth taking a long time to say, and to listen to.

Tolkien, *TLotR III.4*

Encoding

ASCII/Unicode (last four bits):

A	0001	F	0110	K	1011
B	0010	G	0111	L	1100
C	0011	H	1000	M	1101
D	0100	I	1001	N	1110
E	0101	J	1010	O	1111

Sample encoding:

0001	1110	1110	1001	1011	0001
A	N	N	I	K	A

Message size: $4 \times 6 = 24$ bits.

Encoding

Variable-length codes (frequent letters are shorter):

A	0	F	100	K	01
B	10	G	101	L	0000
C	001	H	110	M	11
D	010	I	00	N	1
E	011	J	111	O	0001

Sample encoding:

0	1	1	00	01	0
A	N	N	I	K	A

Message size: $1 + 1 + 1 + 2 + 2 + 1 = 8$ bits.

Encoding

A	0	F	100	K	01
B	10	G	101	L	0000
C	001	H	110	M	11
D	010	I	00	N	1
E	011	J	111	O	0001

0	1	1	00	01	0
A	N	N	I	K	A

Or did you mean

011	0001	0
E	O	A

Encoding

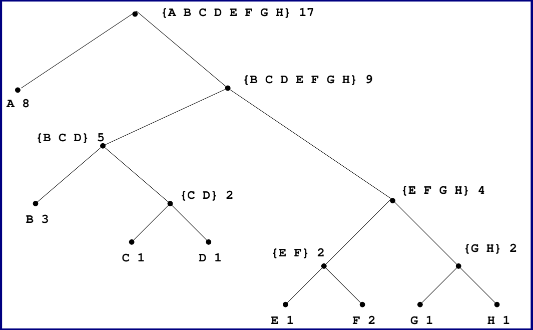
Prefix code

A	0	F	...	K	111
B	...	G	...	L	...
C	...	H	...	M	...
D	...	I	110	N	10
E	...	J	...	O	...

0	10	10	110	111	0
A	N	N	I	K	A

Message size: $1 + 2 + 2 + 3 + 3 + 1 = 12$ bits.

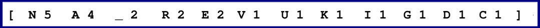
Trees



BACADAEAFABBAAAGAH

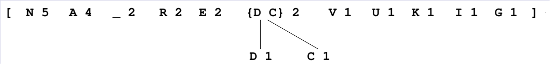
A	0
B	100
C	1010
D	1011
E	1100
F	1101
G	1110
H	1111

Building the tree



```
[ N 5  A 4  _ 2  R 2 E 2 V 1  U 1  K 1  I 1  G 1  D 1 C 1 ]
```

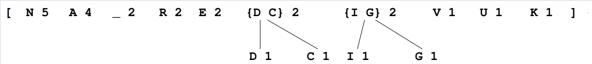
Building the tree



[N 5 A 4 _ 2 R 2 E 2 {D C} 2 V 1 U 1 K 1 I 1 G 1]

D 1 C 1

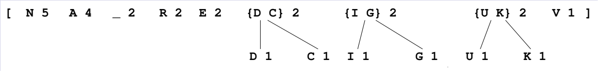
Building the tree



[N 5 A 4 _ 2 R 2 E 2 {D C} 2 {I G} 2 V 1 U 1 K 1]

D 1 C 1 I 1 G 1

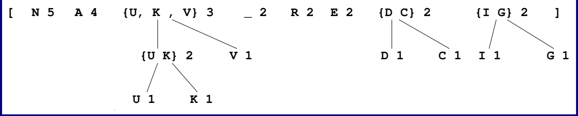
Building the tree



[N 5 A 4 _ 2 R 2 E 2 {D C} 2 {I G} 2 {U K} 2 V 1]

D 1 C 1 I 1 G 1 U 1 K 1

Building the tree

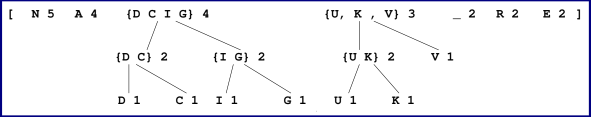


[N 5 A 4 {U, K , V} 3 _ 2 R 2 E 2 {D C} 2 {I G} 2]

{U K} 2 V 1 D 1 C 1 I 1 G 1

U 1 K 1

Building the tree

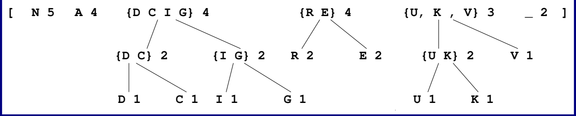


[N 5 A 4 {D C I G} 4 {U, K , V} 3 _ 2 R 2 E 2]

{D C} 2 {I G} 2 {U K} 2 V 1

D 1 C 1 I 1 G 1 U 1 K 1

Building the tree

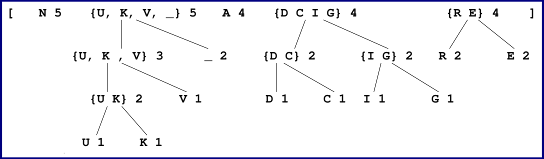


[N 5 A 4 {D C I G} 4 {R E} 4 {U, K , V} 3 _ 2]

{D C} 2 {I G} 2 R 2 E 2 {U K} 2 V 1

D 1 C 1 I 1 G 1 U 1 K 1

Building the tree



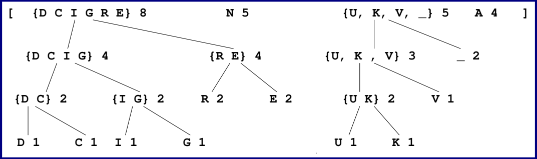
[N 5 {U, K, V, _} 5 A 4 {D C I G} 4 {R E} 4]

{U, K , V} 3 _ 2 {D C} 2 {I G} 2 R 2 E 2

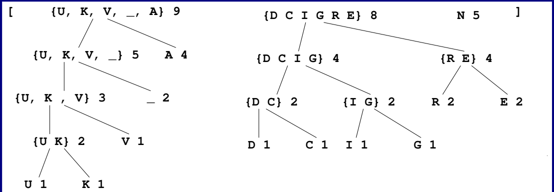
{U K} 2 V 1 D 1 C 1 I 1 G 1

U 1 K 1

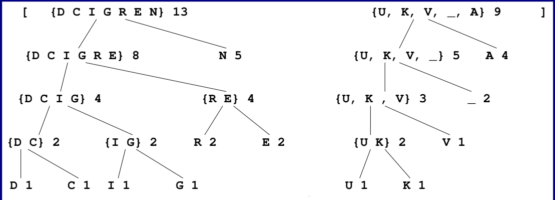
Building the tree



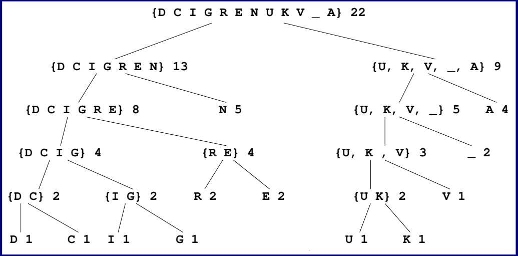
Building the tree



Building the tree



Building the tree



{D C I G R E N U K V _ A} 22

{D C I G R E N} 13

{U, K, V, _, A} 9

{D C I G R E} 8

N 5

{U, K, V, _} 5

A 4

{D C I G} 4

{R E} 4

{U, K , V} 3

_ 2

{D C} 2

{I G} 2

R 2

E 2

{U K} 2

V 1

D 1

C 1

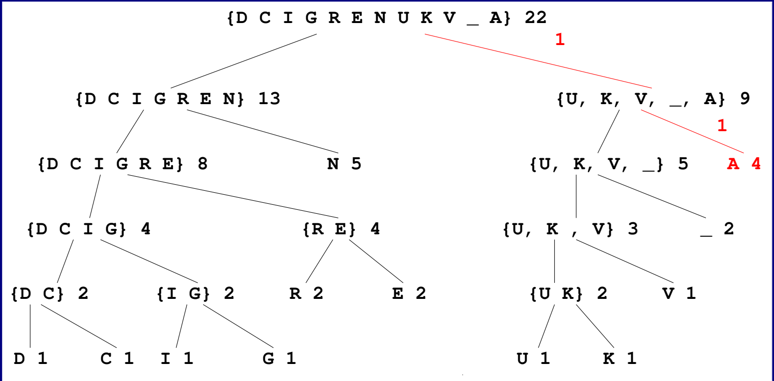
I 1

G 1

U 1

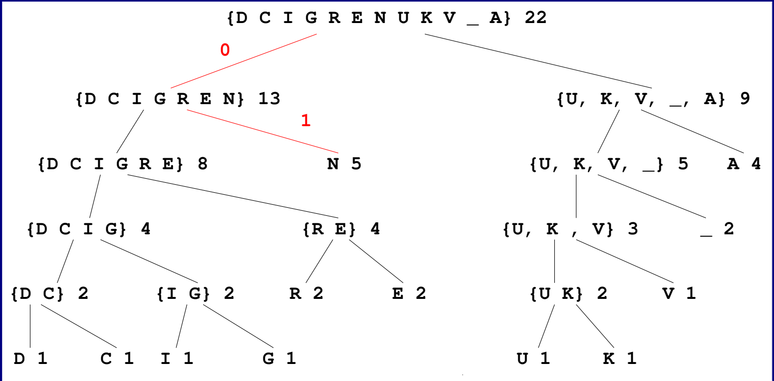
K 1

Encoding the message



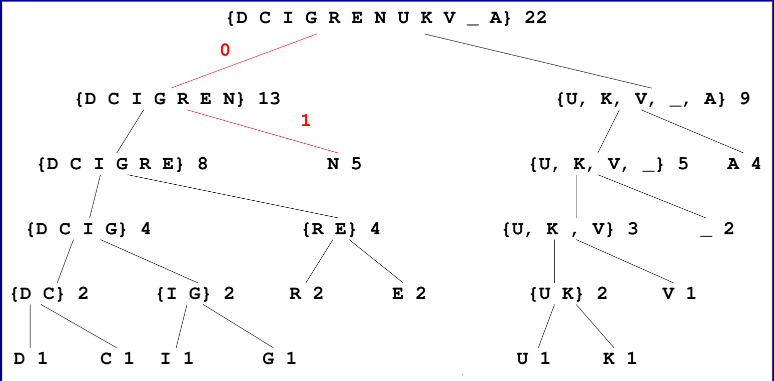
11
A

Encoding the message



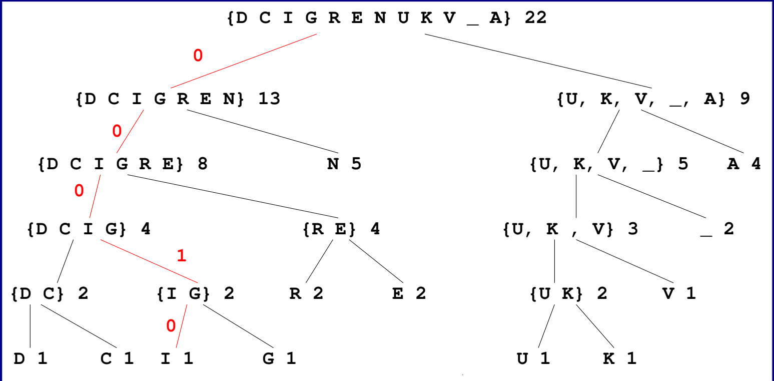
11	01
A	N

Encoding the message



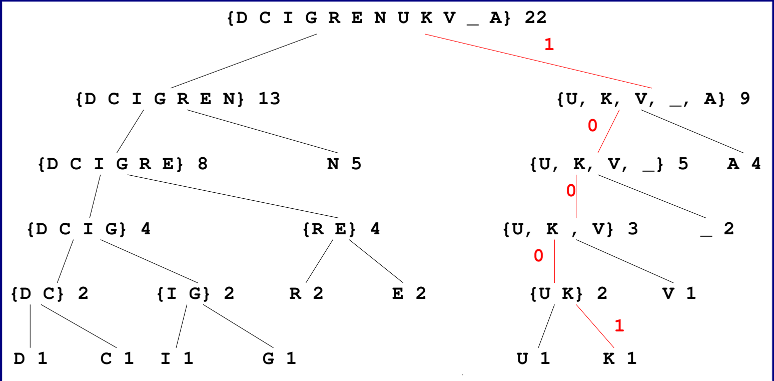
11	01	01
A	N	N

Encoding the message



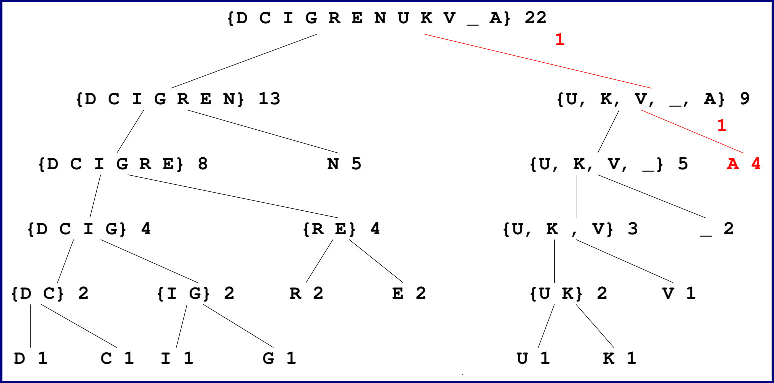
11	01	01	00010
A	N	N	I

Encoding the message



11	01	01	00010	10001
A	N	N	I	K

Encoding the message



11	01	01	00010	10001	11
A	N	N	I	K	A

The meaning of entropy

The word *entropy* had of course been used before Shannon. In 1864 Rudolf Clausius introduced the term. . . to represent a “transformation” that always accompanies a conversion between thermal and mechanical energy. . . .
[One of the authors] asked Shannon what he had thought about when he had finally confirmed his famous measure. Shannon replied: “My greatest concern was what to call it. I thought of calling it ‘information,’ but that word was overly used, so I decided to call it ‘uncertainty.’ When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, ‘You should call it *entropy*, for two reasons. In first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy is, so in a debate you will always have the advantage.’ ”
Tribus and McIrvine, “Energy and Information”, *Scientific American* # 224, Sept 1971, pg 178–184

Perplexity and cross entropy

We suspect that speech recognition people prefer to report on the larger non-logarithmic numbers given by perplexity mainly because it is much easier to impress funding bodies by saying that "we've managed to reduce preplexity from 950 to only 540" than by saying that "we've reduced cross entropy from 9.9 to 9.1 bits." However, perplexity does also have an intuitive reading: a perplexity of k means that you are as surprised on average as you would have been if you had had to guess between k equiprobable choices at each step.

Manning and Schütze, *Foundations of Statistical Natural Language Processing*, pg 78.