CS-579, EECS, KAIST, Dearjeon, Korea

Ch.2 Computer processing of lexicons

Prof. Feng Zhiwei

2.1 Statistics of words and word list

-- Some idea and research for word study by statistics

A. De Morgen, statistic study for word length as a stylistic feature of text, 1851

F. W. Kaeding, frequency dictionary of German language, 1898

G. K. Zipf, Zipf's law, relation between word frequency and its order number in frequ ency dictionary, 1935

E. Varder Beke, word range (distribution rate), 1935

G. U. Yule, <Statisric Analysis for literature words>, 1944

R. Michea, lexical statistics, 1949

P. Guiraud, richness of word, 1954

G. Herdan, <language as choice and chance>, 1956.

D. Keil, lexicometric, 1956

--Word list

. word list in Babylon cuneiform charactera (used in ancient Babylonia, Assyria, Persi a and some other areas of the Near East Asia). before 3000 years

. Aefrie, English-Latin classification word list in <Latin Grammar>, 9 century

. Abbe de l'Eppee (France), Word list of French, 5400 words, 3 stages, 1800 words p er stage, for language teaching to deaf-mute children

The criterions for selection of words in list:

. subjective criterions: selection of words is according to the subjective experience of scholars.

Basic English (C.K. Ogden, I.R. Richards, in 1930s): 800 words to give the defi nition of all other English words. It is similar as LONGMAN Dictionary (2000 words to explain all other English words).

. objective criterions for selection of words in list:

- frequency criterion:

<German Frequency Dictionary>, F. W. Kaeding, 1898, for German stenogra phy study (shorthand).

corpus (original materials of language): 110 texts (one text :100,000 words), total sum of words: 10,919,777 words

The number amount of frequently-used words (its absolute frequency > 4) : 79716 different words

The limit of reliability of frequency: If there are 10 papers, frequency of w ord A is 0.020, frequency of word B is 0.018, but A was distributed in 1 paper only,

and word B was distributed in 10 papers. which word is more important? -- obviously, word B.

Some times the frequency might conceal the actual state of affairs, and give a false impression.

- range criterion: The range is the distribution rate of word in the text.

<Handbook of French vocabulary>, E. Varder Beke (Canadian scientist), 1935.

corpus: 88 texts (one text: 13,000 words),
total sum of words: 1,100,000 words.
different words: 19,253.

range index: the distribution times in different texts. If a word was distribut ed 5 times in 88 texts, then its range index equals 5.

Amount of words which range index is greater than 5 (> 5): 6067, it is 31.5% of total sum of words.

C. Muller (French scientist) said: "If the conception of frequency can not be combined with the conception of range, then the frequency conception will be not so valuable."

- availability criterion: The availability is the association degree of a word in the human brain. So it is the act of joining or the state of being joined of a word with other words in the speech activity. In the speech activity, many words with low frequ ency can be often associated, these words are also important in the language. e.g. the words 'spoon, fork, face, neck, hand, arm, finger, bed, toilet, etc' are very easily associ ated in the speaker's brain, but their frequency in the text is not so higher.

<Francais fondamental>, G. Gougenheim, R. Michea, P. Rivenc, A. Sauvageot, 1954
corpus: 312135 words
different word: 7995
high frequency words: 1063 (absolute frequency > 20)
selection procedure:
- first step: 805 words (absolute frequency > 25, range index > 29) selected from 106 3 words,

701 words selected from 805 (104 vulgar words were deleted).

- second step: to add 774 words with good availability (in the field of body, cloth, ha use, furniture, food, drinking, etc)

The Total sum of basic vocabulary: 1475 words:

. notional words: 1222

noun: 692, 46.9% (in original list, the amount of noun is only 395, mo st of noun was selected by the availability criterion0.

verb: 339, 22.9%

. functional words: 253

<List of Grunddeutsch>, J. A. Pfeffer, 1960s.
Corpus: 833,000 words
- First step: 737 words (based on frequency criterion and range criterion, absolute fre quency > 40, range index > 25)

- Second step: 347 words ware selected by avilability criterion.
The amount of words: 737 + 347 = 1084 words
- Third step: 185 words were added according to the experience. e.g. if "sun' was sel ected in the list, then 'moon, star" can be added in the list.

the nunber of basic words in French and German is very similiar (=2000 words).

<A General Service List English Words>, M. West, 1953, London
the most frequently used English words: 2000 words
semantic frequency for polysemy:
e.g. 'game'

1. to laugh at, make fun of : frequency 9%
   It's not serious, it's just a game.
2. a form of play or sport: frequency 33%
   A game of football
   Indoor games
   outdoor games
3. A particular set of sport competition (in plural form games): frequency 8%
   Olympic Games
   ….

<French idiom list>, F.D. Cheydler, 1940, New York

idiom	absolute frequency
faire: Il faut (batir une maison)	1140
avoir: Il y a (des plumes sur la table)	1638
avoir: Il a peur de (tomber)	178
.....	

<The Teaching of English Suffixes>, E. L. Thorndike, 1941, New York
frequency of English suffixes

<Semantic Frequency List of English, French, German and Spanish>, E. I. Eaton, 19 40, Chicago
It is a multilingual frequency dictionary.

<Frequenct Dictionary of Modern Chinese>, Peking Language Institute, 1979.
Corpus: Newpaper: 440,000 Chinese characters, 24.4%
Scientific paper: 290,000 Chinese characters, 19.8%

Speech material: 200,000 Chinese characters, 11.1%
Literature paper: 890,000 Chinese characters, 48.7%

Total amount: 1,820,000 Chinese characters.

Word list: (arrange in the decreasing frequency order)

The relation between order rank (order degree) and coverage to corpus in the wo rd list:

order rank	coverage to corpus
<100	40%
<500	70%
<2562	85%
<31159	100%

2.2 Zipf's law

The relation between the frequency of word and the rank of word in the frequency word list.

2.2.1 J. Estoup (stenography scientist, France, 1916)

If the word list was arranged in the decreasing order of absolute frequency (expressed by n), the rank of word in the list (expressed by r) is from 1 to L ($1 \leqslant r \leqslant L$) ,

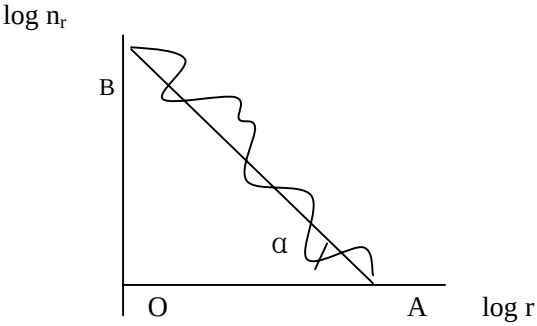
Word rank	1	2	……	r	……	L
Word frequency	n_1	n_2	……	n_r		n_L

then we may have

$$n_r \cdot r = K$$

K is a constant.

2.2.2 E. Condon (physician, Bell Telephone Company, USA, 1928)



Let x=logr, y=log n_r,

OB=log*k*

tg α = γ ,

then we have

$$OA = OB/tg\,\alpha = \log k/\gamma$$

and we have also

$$x/OA + y/OB = 1$$

$$\log r / \log k/\gamma + \log n_r/\log k = 1$$

$$\gamma \log r / \log k + \log n_r/\log k = 1$$

$$\gamma \log r + \log n_r = \log k$$

$$\log r^{\gamma} + \log n_r = \log k$$

so

$$r^{\gamma} \cdot n_r = k$$

$$n_r = k / r^{\gamma}$$

$$n_r = k\ r^{-\gamma}$$

he found that $\alpha = 45^o$

$$\gamma = tg\,\alpha = tg\ 45^o = 1$$

then above formula becomes

$$n_r = k\ r^{-1}$$

$$n_r /N = k/N \cdot r^{-1}$$

$n_r /N = f_r$,

let $k/N = c$,

So We have

$$f_r = c\ r^{-1}$$

In Condon's formula, c is considered as a constant.

2.2.3 G.K. Zipf (philologist, USA, 1935) checked the result of E. Condon.
<Word index to James Joyce's 'Ulysses'> (different words: 29,899 , total words in text: 260,432 words)

- When test times t -> ∞, frequency (f) becomes probability (p), we have

$$p_r = c\ r^{-1}$$

- When r=1,

$$p_r = c \times 1^{-1}$$

$$= c$$

It means that c is the word probability when its rank is 1.
Zipf believed: c= 0.1, so c is a constant.

- Afterward, Zipf found that c must be a parameter,

$$0 < c < 0.1$$

For r = 1,2, ,….., n, this parameter makes

$$\Sigma p = 1$$

Its rank distribution law was called as Zipf's law.

2.2.4 M. Joos formula (1936, two parameters)):

The Zipf' formula

$$p_r = c\, r^{-1}$$

was come from formula

$$p_r = c\, r^{-\gamma}$$

where γ is not always = 1. If the word number becomes large, γ becomes large also; if word number becomes smaller, the γ become smaller ($\gamma = tg\ \alpha$) also. Therefore γ is a parameter, $\gamma = b$, so we have formula

$$p_r = c\, r^{-b}$$

It is Joos's formula. b>0, c>0, for r = 1, 2, …., n, we have

$$\sum p = 1.$$

When b=1, Joos's formula becomes to Zips's formula.

2.2.5 B. Mandelbrot formula (1950s, three parameters):

$$p_r = c\,(r + a)^{-b}$$

It is Mandelbrot's fomula, $0 \leqslant a < 1$, b>0, c>0, for r-1,2, .., n, we have

$$\sum p = 1.$$

The meaning of parameters are as follows:

- The parameter c is related with probability of word which has the highest probability;
- The parameter b is related with the number of words which have the high probability.
- The parameter a is related with the number of words n, its value can be selected freely. The value of a depends on the concrete condition.

In the Mandelbrot formula,
When a=0, then formula has following form

$$p_r = c\, r^{-b}$$

It becomes Joos's formula.
When a=0, b=1, the formula has following form

$$p_r = c\, r^{-1}$$

It becomes Zipf's formula.

2.2.6 coverage rate of words with higher probability in text :

In Zipf's formula, if c=0,1 in a concrete language, then

$$p_r = 0.1 \times r^{-1}$$
$$= 0.1 / r$$

the sum of frequency of 1000 words with higher probability in the word list can be calculated.

$$\Sigma p = \Sigma (0.1 / r)$$
$$= 0.1 \times (1/1 + 1/2 + 1/3 + \cdots + 1/1000)$$
$$= 0.1 \times 0.748$$
$$= 74.8\%$$

It means that 1000 words with higher probability can cover the majority part of the text.

2.2.7 data dispersal phenomena: However, for the words with lower probability (rank > 1000), the data will become dispersal, the Zipf's law can not be true.

2.3 Entropy and redundancy of language

2.3.1 language action as a probabilistic process:

sender of message → telecommunication medium → receiver of message

- The words (event) appeared in language action is a function of time, it can be changed along with the change of time;
- The function value of words (event) in every moment is non-deterministic & stochastic (probabilistic).The function value of word (event) in every moment is distributed according to the probability of words (event).

The text can be regarded as the alphabet (event) chain of probabilistic test result.

- independent chain with equal probability: the event (element of language) is independent and its probability is equal.

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQ SDHYD QPAAMKBZAACIBZLHJQD

- Independent chain with non-equal probability: the event (element of language) is independent and its probability is not always equal

OCRO HLIRGWR NMIELWIS EU LLNBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRI

- One grade Markov chain: The probability of an event (element of language) is dependent only on the event (element of language) that directly precedes it. It means that each event depends upon one directly previous event.

ON IE ANTSOUTINYS ARE TINCTORE BE S DEAMY ACHIND ILONASINE TUCDOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTIBE

- Two grade Markov chain: The probability of an event (language element) depends upon two previous events (language elements)

IN NO IST LAT WHEY CRATICT FROUREBIRS CROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOAQCTIONA OF CRE

We can have three grade Markov chain, four grade Markov chain, five grade Markov chain, etc.

If the event is the word of language, then the result of probabilistic test are as follows:

−− Independent chain with non-equal probability:

REPRESENTING AND SPEEDILY IS AN GOOD APT OR CAME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TOOF TO EXPERT GRAY COME TO FURNISHES THE MESSAGE HAD BE THESE

−− One grade Markov chain

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

−− Two grade Markov chain

FAMILY WAS LARGE DARK ANIMAL CAME ROARING DOWN THE MIDDLE OF MY FRIENDS LOVE BOOKS PASSIONATELY EVERY KISS IS FINE

−− Four grade Markov chain

ROAD IN THE COUNTRY WAS INSANE ESPECCIALLY IN DREARY ROOMS WHERE THEY HAVE SOME BOOKS TO BUY FOR STUDYIBF GREEK

The limit of grade (15 grades) of Markov chain may be the normal sentence.

E.g. The **people** who called and wanted to rent your house when you go away next year **are** from California.

2.3.2 Entropy of language

In information theory, the non-definiteness of event in the probabilistic test is called as entropy. The entropy is the mathematical measure of the non-definiteness. It equal to the information content that was received in the language activity.

The information content before communication	after communication
No information	new information
Non-definitiveness is big	non-definitiveness was decreased to zero

The entropy equals to the decreased amount of non-definitiveness.

In information theory, the formula of entropy (H) is as following:

−− For the independent chain with equal probability

$$H_0 = \log_2 n$$

H_0 is the entropy, n is the number of language element.

If n =2, then

$$H_0 = \log_2 2$$
$$= 1 \text{ (bit)}$$

The unit of entropy is bit.

-- For the independent chain with non-equal probability

$$H_1 = \Sigma P_i \log_2 P_i$$

H1 is entropy, p is the probability of the language element in the text.

$$\text{Log}_2 n \geqslant \Sigma P_i \log_2 P_i$$

$$H_0 \geqslant H_1$$

-- The conditional entropy:

General formula: $H_n = - \Sigma P_{[bi(n-1),j]} \log_2 P_{bi(n-1)}(j)$

-- For the one grade Markov chain:

$$H_2 = - \Sigma P_{ij} \log_2 P_i(j)$$

P_{ij} is the probability of i and j in the text, $P_i(j)$ is probability of j when its previous element is i.

-- For two grades Markov chain:

$$H_3 = - \Sigma P_{ijk} \log_2 P_{ij}(k)$$

P_{ijk} is the probability of i, j and k in the text, $P_{ij}(k)$ is the probability of k when its previous elements is i and j.

The series H_k is non-increasing. So we have

$$H0 \geqslant H1 \geqslant H2 \geqslant H3 \geqslant \cdots \geqslant H \ldots \geqslant \ldots \rightarrow H_\infty$$

H_∞ is limit entropy when the grade of Markov chain became to infinitively greet.. It is the information content including the text.

The structure of language makes the decreasing of entropy of language.

The entropy (for independent chain with non-equal probability) some languages:

French:	3.98 bits
Italian:	4.oo bit
Spanish:	4.01 bits
English:	4.03 bits
German:	4.10 bits
Russian:	4.35 bits
Chinese:	9.65 bits

2.3.3 Redundancy of language

The redundancy of language is the proportion of redundant elements in a language.

The formula of redundancy:

$$R = 1 - H_\infty / H_0$$

In Russian, $H_\infty = 1$, $H_0 = 1$,

so $R = 1 - 1/5 = 0.80 = 80\%$

The redundancy of some languages:

	Russian	English	French	Germany
Roman	0.812	0.818	0.773	0.743
Science paper	0.868	0.875	0.872	0.835

Application of redundancy: -- Compression of text: Optimal encoding of text.
-- Repair the damaged documents.

2.4 Machine Readable Dictionary (MRD)

−− Dictionary of French verbs (Morris Gross, LADL, Laboratoire Automatique de Documentation Linguistique, France,1990): 6000 verbs, 81 matrix list.

−− Electronic Dictionary project (EDR, Japan Electronic Dictionary Research Institute, Japan, 1986) :

- Main Dictionary: morphological, Syntactic and semantic information
- Concept Dictionary: classification system of conception, description of conception.

−− Contemporary Chinese Dictionary of Grammar information (ICL, Institute of Computational Linguistics, Peking University, 1996): 50,000 words

2.5 Word-Net and lexical knowledge base

2.5.1 WordNet (G. A. Miller, R. C. Beckwick, C. Fellbaum, Princeton University, USA, 19 85)

2.5.1.1 Presuppositions of WordNet:

-- separability hypothesis: The lexical component of language can be isolated and studied in its own right.

-- patterning hypothesis: the people just to take advantage of systematic pattern and re lation among the meanings that words can be used to express,

-- comprehensiveness hypothesis: the system would need to have available a store of lexical knowledge as extensive as people have.

2.5.1.2 Contents of WordNet:

Although WordNet contains compounds, phrasal verbs, collocations, and idiomatic phra ses, the word is the basic unit. (Noun, Verb, Adjective, and Adverbs).

2.5.1.3 Nouns in WordNet

80,000 noun word form organized into some 60,000 lexicalized concepts.

The basic semantic relation in WordNet is synonymy.

Sets of synonyms (SYNSET) form the basic building blocks.

2.5.1.3.1 Hyponymy (subordination relation)

hypernym : hyponym

bird : robin (the bird with a brown back and wing and a red breast)

the noun robin is a hyponym (subordinate) of the noun bird, or, conversely, bird is a hypernym (superordinate) of robin.

It is the semantic relation that organizes nouns into a lexical hierarchy.

robin: "a migratory bird that has a clear melodious song and a reddish breast wi th gray or black upper plumage". hypernym (genus term) of robin: bird.

bird: "a warm-blooded egg-laying animal might having feathers and forelimbs mo dified as wing."

animal: "an organism capable of voluntary movement and possessing sense organs and cells with non-cellulose walls."

organism: " a living entity."

Each hypernym leads on to a more generic hypernym.

Hypernymy cannot be represented as a simple relation between word forms.

E.g. when we say that 'tree' is a kind of 'plant', we are not talking about 'tree gra phs' or 'manufacturing plants'.

Hypernymy is a relation between particular senses of words.

Hypernymy is a relation between lexicalized concepts, a relation that is represente d in WordNet by a pointer between the appropriate synsets.

{robin, redbreast} @-> {bird} @-> {animal, animate_being} @-> organism, life_fo rm, living_thing}

@ is the transitive, asymmetric.

Semantic relation that can be read 'IS-A' or 'IS-A-KIND-OF'

@-> is said to point upward

generalization: @-> goes from specific to generic. Ss @-> Sg

specification: ~-> goes from generic to specific Sg ~-> Ss

Inheritance: If Rex is a collie (a kind of dog much used for tending sheep), Rex is a dog; and if Rex is a dog, then Rex is an animal; and if Rex is an animal, then Rex is capable of voluntary movement.

Unique Beginners (Primitive semantic component):

WordNet use the set of 25 unique beginners for noun source file:

{act, activity}
{animal, fauna}
{artifact}
{attribute}
{body}
{cognition, knowledge}
{communication}
{event, happening}
{feeling, emotion}
{food}
{group, grouping}
{location}
{motivation, motive}
{natural object}
{natural phenomenon}
{person, human being}
{plant flora}
{possession}
{process}
{quantity, amount}
{relation}
{shape}
{substance}

{time}

Reduce the 25 noun source file to 11 unique beginner (The unique beginners are italicize d)

		Animal
	Organism	Person
Entity		Plant
		Artifact
	Object	Natural object
body		
		Substance
food		
		Attribute
		Quantity
	Abstraction	Relation
communication		
		Time
		Cognition
	Psychol. Feature	Feeling
		Motivation
		Nat. phenomenon
process		
		Activity
		Event
		Group
		Location
		Possession
		Shape
		State

2.5.1.3.2 Meronymy (relation between parts and whole)

-- The part-whole relation between nouns in WordNet is generally considered to b e a semantic relation -- Meronymy (meros in Greek : part)

meronym : holonym -- "is a part" or 'has a"

If Sm is a meronym of Sh, then Sh is said to be a holonym of Sm.

'*If Wm is a part of Wh* ' is acceptable, then 'W_m *is a meronym of* W_h *'* ; if 'W_h *has a* W_m *(as a part)* ' is acceptable, then ' W_h *is a holonym of* W_m' .

-- Meronymy is often compared to hyponymy: both are asymmetric and (with rese rvation) transitive. "is a part of"

e.g. a finger is a part of a hand, a hand is a part of a arm, an arm is a part of a body.

-- Six types of meronyms (Winston, Chaffin, 1987):

component-object (branch/tree)

member-collection (tree/forest)

portion-mass (slice/cake)
stuff-object (aluminum/airplane)
feature-activity (paying/shopping)
place-area (Princeton/ New Jersey)

-- Only three types of meronyms are coded in WordNet:

'Wm #p-> Wh' indicates that 'Wm is a component part of Wh'
'Wm #m-> Wh' indicates that 'Wm is a member of Wh'
'Wm #s-> Wh' indicates that 'Wm is the stuff that Wh is made from'

2.5.1.3.3 Antonymy

Semantic opposition is not a fundamental organizing between nouns, but it does exist and so merits its own representation in Wordnet:

e.g. [{man} !-> {woman}]
[{woman} !-> {man}]

The noun antonyms nearly always have the same hypernym, often the same immediate hypernym.

2.5.1.4 Adjective in WordNet

2.5.1.4.1 The adjective are words whose sole function is to modify nouns.

e.g. a *large* chair, a *comfortable* chair

The nouns, present and past participles of verbs, prepositional phrase, even entire clauses are frequently used as adjectives:

e.g. *kitchen* chair, *barber* chair (noun).
The *creaking* chair (present participle),
The *overstuffed* chair (past participle)
Chair *by the window* (prepositional phrase)
The chair that you bought at the auction (entire clause)

WordNet contains 16,428 adjective SYNSET synonym set including many nouns, participles, prepositional phrase

2.5.1.4.2 Two categories of adjectives

- descriptive adjective: big, beautiful, interesting, possible, married.
  The descriptive adjective typically describes to a noun a value of an attribute
  'X is Adj' presupposes that there is an attribute A such that A(x) = Adj.
  e.g 'the package is heavy' → there is an attribute WEIGHT such that WEIGHT (package) = heavy
  heavy or light is the value of attribute WEIGHT.
  WordNet contains pointer between descriptive adjectives and the nouns by which appropriate attribute are lexicalized.
- relational adjective (they are related by derivation to nouns): 'electrical' is related to the noun 'electricity'.

2.5.1.4.3 Antonymy

The basic semantic relation among descriptive adjective is antonymy.

good -- bad

The mutuality of association is a salient feature of the data for descriptive adjective.

-- bipolar: The attributes of descriptive adjective tend to be bipolar.

Antonymous adjective expresses opposing values of an attribute.

e.g. the antonym of ‘heavy’ is ‘light’, which expresses a value at the opposite pole of the W EIGHT attribuye.

In WordNet, the binary opposition is represented by reciprocal labeled pointers meaning ‘IS-A NTONYMOUS-TO’ and is displayed as ‘*heavy (vs. light)*’ and ‘*light (vs. heavy)*’.

heavy !-> light

light !-> heavy

-- word form representation: A word form with two different meanings is two different word form. In WordNet, the difference is represented by digits: if ‘hard’ meaning ‘unyielding’ is ‘hard1’ and ‘hard’ meaning ‘difficult’ is ‘hard2’, then the antonym of ‘hard1’ is ‘soft’ and t he antonym of ‘hard2’ is ‘easy’.

-- direct antonym and indirect antonym:

‘heavy/light’, ‘weighty/weightless’ is direct antonymy,

‘ponderous’ (large and heavy) is the adjective lacking antonyms, but it is similar in me aning to adjectives ‘heavy’ that to have antonym. The term ‘similar’ is to say: the class of n ouns that can be modified by ‘ponderous’ is including in the class of nouns that can be modif ied by ‘heavy’. The ‘ponderous’ is similar to ‘heavy’, and ‘heavy’ is the antonym of ‘light’, s o a conceptual opposition of ‘ponderous/light’ is mediated by ‘heavy’. ‘ponderous/light’ is indir ect antonym, but it is not lexically paired.

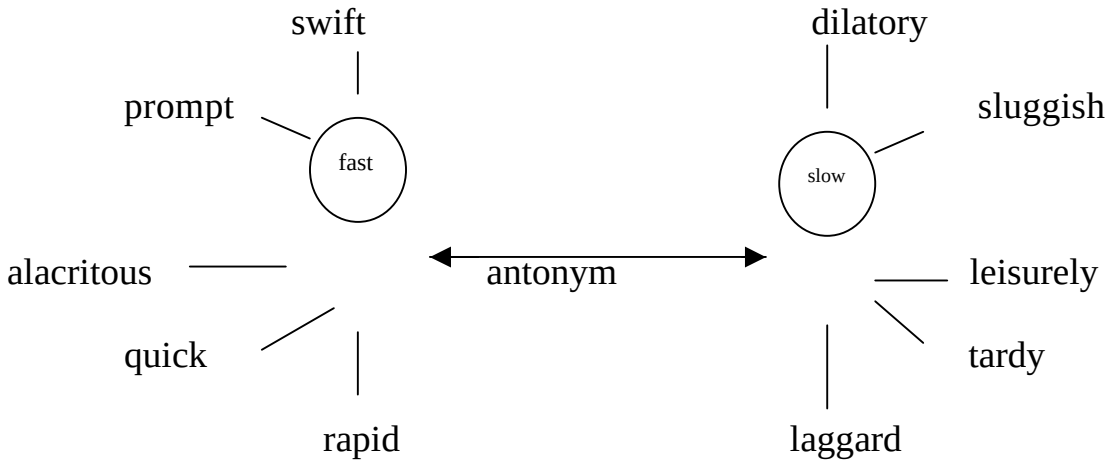
Form heavy !-> light and ponderous &-> heavy.

We have ponderous !-> light

Under this formulation, all descriptive adjectives have antonyms, those lacking direct ant onyms have undirect antonyms (i.e., are similar in meaning to adjectives that have direct anton yms).

In WordNet, direct antonym are represented by ‘!’. Indirect antonyms are inherited throu gh the similarity, which is indicated by the similar pointer ‘&’ meaning is ‘IS SIMILIAR TO’.

--bipolar cluster of adjectives: We organize adjectives into bipolar clusters.



WordNet contains over 1,732 of these clusters, one for each pair of antonyms; thus ther e are 3,464 clusters of closely similar senses.

--head synset and satellite synset: The cluster for ‘fast/slow’ which defines the attribute ‘SPEED’, consists of two half clusters, one for senses of ‘fast’, one for senses of ‘slow’. Each

half cluster is headed by ‘head synset’ (fast/slow)

Following the head synset is ‘satellite synset’, which represents senses that are similar t o the sense of the head adjective.

--Antonymous pairs expressing the same sense or closely related senses and representing values of the same attribute: e.g. ‘big/little’ and ‘large/small’ are equally salient as antonyms defining the attribute ‘SIZE’,

In WordNet, a single cluster was created headed by both pairs and displayed as *large (vs. small), big (vs. little)*
for one half cluster and *small (vs. large), little (vs. big)* for the other half cluster.

--The adjectives can be graded in WordNet in different attribute.

SIZE	LIGHTNESS	QUALITY	BODY-WEIGHT	TEMPERATURE
Astronomical	snowy	superb	obese	torrid
huge	white	great	fat	hot
large	ash-gray	good	plump	warm
…	gray	mediocre	…	tepid
small	charcoal	bad	slim	cool
tiny	black	awful	thin	cold
infinitesimal	pitch-black	atrocious	gaunt	frigid

Most attributes have an orientation. It is natural to think of them as dimensions in a h yperspace, where one end of each dimension is anchored at the point of origin of the space.

2.5.1.4.4 Relational adjective

The relational adjective is related semantically and morphologically to noun, though the morph ological relation is not always direct.

e.g ‘musical’ (in ‘musical instrument’) is related to ‘music’;
‘dental’ (in ‘dental hygiene’) is related to ‘tooth’.

Frequently, the noun can be modified by both the relational adjectuve and the noun from whic h it is derived.

e.g. ‘atomic bomb’ : ‘atom bomb’
‘dental hygiene : ‘tooth hygiene’.

The different of relational adjective and descriptive adjective:

--relational adjectives do not refer to a property of the nouns they modify and so do not rel ate to an attribute.

--relational adjective are not gradable (*’the very atomic bomb’);

--most of relational adjectives lack direct antonyms.

So the relational adjective can not be included the cluster, they can not have the bipolar.

In WordNet, the relational adjectives file contains 2823 adjective synset. Every relational adject ive has a pointer to the corresponding noun.

e.g. stellar, astral

⇨ star

⇨ Celestial body, heavenly body – (natural objects visible in the sk y)

2.5.1.4.5 Adverb:

--Most of adverbs derived from adjectives by suffixation>

e.g. ‘beautifully, oddly, quickly, interestingly, hurriedly’.

--Other adverbs are derived by adding any of a number of other suffixes: -ward (northward, forward), -wise (crosswise), -ways (sideways).

--In WordNet derived adverbs are linked to adjective sense by means of a pointer meaning ‘DERIVED-FROM”.

2.5.1.4.6 Verb:

2.5.1.4.6.1 Unique beginner of verbs:

14 semantic domains: motion, perception, contact, communication, competition, change, cognition, consumption, creation, emotion, possession, body care and function, social behavi or and interaction.

Pulman (1983) suggests just ‘be’ and ‘do’ as the root of all verbs: verb ‘be’ (static ver b), verb ‘do’ (activity). But WordNet believed iyt is not entirely appropriate due to the pol ysemy. WordNet distinguishes 12 senses each for ‘do’ and ‘be’. E.g. do: ‘do my hair’, ‘do my room in blue’.

Be: ‘to be or not be, that is the question’, ‘Let him be, I tell you’.

There are 11,500 verb synset in Wordnet 1.5 version)

Within a single semantic field, it is frequently the case that not all verbs can be group ed under a single unique beginner. Some semantic domains can be represented only by sev eral independent trees.

E.g motion verb: move1 (translational movement), move2 (movement without displacem ent),

Possession verb goes upwards to three concepts expressed by three synset: {give, t ransfer}, {take, receive}, {have, hold}.

Communication verb: verbal communication, nonverbal (gestural) communication.

motion,

2.5.1.4.6.2 Entailment

Entailment refers to the relation between two verbs V1 and V2 that holds when the sentence ‘Someone V1’ logically entail the sentence ‘Someone V2’.

e.g. ‘nore’ lexically entails ‘sleep’ because the sentence ‘He is snoring’ entails ‘He is sleeping’. The second sentence necessarily holds if the first sentence does.

- Lexical entailment is a unilateral relation: If a verb V1 entails another verb V2, then it cannot be the case that V2 entails V1.
- When two verbs can be said to be mutually entailing, they must be synonyms; t hat is they must have the same sense.
- Negation reverses the direction of entailment: ‘not sleeping’ entails ‘not snoring’, but ‘not snoring’ does not entail ‘not sleeping’.
- Converse of entailment is contradiction: If the sentence ‘He is snoring’ entails ‘H e is sleeping’, then ‘He is snoring also contradicts the sentence ‘He is not sleepi ng’.
- temporal relation in the verbs which have entailment relation: e.g. ‘drive’ and ‘r ide’ are connected in that when you drive a vehicle, you necessary also ride in i t (simultaneously).
- Temporal inclusion of entailment relation: ‘Snoring’ and ‘sleeping’ is also te

mporally coextensive, the time you spend snoring is a part of the time you s pend sleeping (the snoring time is including in the sleeping time, not always simultaneous). When you stop sleeping, you also necessarily stop snoring (b ut you may continue sleeping without snoring). That is to say, the set of ver bs related by entailment have in common that one member temporally includ es the other. A verb V1 will be said to include a verb V2 if there is some stretch of time during which the activity denoted by the two verbs to occur, but no time during which V2 occurs and v1 does noy. If there is a time dur ing which V1 occurs but V2 does not, V1 will be said to properly include V2.

2.5.2 HowNet

HowNet is an on-line common-sense knowledge base describing inter-conceptual relations and inter-attribute relations of concepts. These concepts are expressed in lexicons of the Chinese and their English equivalents.

2.5.2.1 Motivation

Dong ZhenDong (CIP Association, Beijing) brought to light the following viewpoints in a series of papers published in 1988.

- In the final analysis, natural language processing ultimately requires the support of a powerful knowledge base.
- Knowledge is computer-operable, it is a system surrounded with the varied relations amongst concepts as well as those amongst the attributes of concepts.
- On the creation of a knowledge base, a common-sense knowledge base constitut ing a knowledge system should first be constructed.
- The knowledge is owned by all. By this reason, the knowledge engineers first design the framework and suggest a common-sense knowledge base prototype. Upon this foundation, work can be extended to develop a specialized knowledge bas The idea is analogous to the edition of a dictionary for general use and an enc yclopedia.

Research and construction of HowNet is a manifestation of the above-mentioned viewpoints.

2.5.2.2 Philosophy of HowNet

The philosophy behind HowNet lay ground on its understanding and interpretation of the objective world.

- All matters (physical and metaphysical) are in constant motion and are ever changing in a given time and space.
- Things evolve from one state to another. The change of things will lead to the corresponding change in their attributes.

In the case of "human", it is characterized by the four main states of living: at birth, aging (becoming old), fall sick and dead. Age (an attribute) catches up in a person, giving the attribute "age" a value, i.e. "old". As a person grows, his/he r hair color (an attribute) turns grey (the attribute-value). On the other hand, as a person grows, the character (metaphysical) gradually matures (attribute-value), so is the knowledge (metaphysical product) that will develop wider and deeper (the

attribute-values).

The above depicts the units for manipulation and description in HowNet: being thing (sub-divided into physical and mental), Part, Attribute, Time, Space, Attribute-value and Event.

The significance of Part and Attribute in the philosophy of HowNet must be emphasized.

-- **Part**: all objects are probably part of something else while at the same time, all objects are also the whole of something else.

e.g. Doors and windows are parts of buildings;
The limbs are parts of animals;
The buildings form parts of a community;
The individual is part of the family or society he/she belongs to.

All things can be divided into their respective components.

e.g. Space can be segmented into "up", "down", "left", "right".
Time can be seen from "the past", "the present" and "the future".

Depending on the system of reference, the same point of reference can either be regarded as a whole or a part.

In HowNet, Part is taken as a constituent in a larger whole.

The role and function of Part in whole is analogous to the human body, for instance, "hilltop", "hillside", "mountain foot", "table leg", "back of chair".

"door" and "window" of buildings are analogous to the relevant parts of the human body such as the eyes, the mouth etc.

It is interesting to note that the same analogy applies to different languages. This shows how similar the mankind shares their views on the relations between part and whole.

-- **Attribute**: Any one object necessarily carries a set of attributes. Similarities and differences between the objects are determined by the attributes they each carry.

There will be no object without attributes.

e.g. Human beings are attached with natural attributes such as race, color, gender, age, ability to think, ability to use language as well as social attributes such as nationality, job, wealth etc.

Under specific conditions, it is true to say that the attached attributes are even more important than the host itself.

For instance, if we want to clamp a nail on the wall but does not have a hammer, what would be the best alternative tool? Obviously, it would be something that carries attributes close to a hammer, where in this case, weight and hardness would be the key attributes. The relationship between the attributes (e.g. weight and hardness, etc.) and their host (a hammer) is aesy to understand.

The attributes simply come with the host and vice versa.

The attribute-host relation differs from the part-whole relation. HowNet reflects this difference by way of coding specifications such that attributes are necessarily defined in terms of the possible classes of host. In this connection, HowNet also requires pointers to indicate the relevant attributes when defining attribute-value

s.

2.5.2.3 Characteristic of HowNet

- Fully computational: HowNet is a system by the computer, for the computer, and expectantly, of the computer.
- As a knowledge base, the knowledge structured by HowNet is a graph rather than a tree.
- HowNet is devoted to demonstrate the general and specific properties of concepts.

For instance, "human being" is the general property of "doctor" and "patient". The general properties of "human being" are documented in Main Features of Concepts.

“doctor” and “patient” have respectively their specific properties:
Being the agent of cure is the specific attribute of "doctor".
Being the experiencer of unwell is the specific attribute of "patient".

Be it the millionaire or the poor; the beauty or the ugly, being a human being is the general property they all share, though each take a distinct attribute-value, namely, rich, poor, beautiful and ugly.

HowNet teaches the following knowledge graph to the computer so that they are computer-operable.

- hypernym-hyponym (implied by main features of concepts)
- synonym (by means of "SACR" [Synonymous, Antonymous and Converse Relations])
- antonym (by means of "SACR")
- converse (by means of "SACR")
- part-whole (coded with pointer %, e.g. "heart", "CPU", etc)
- attribute-host (coded with pointer &, e.g. "color", "speed", etc)
- material-product (coded with pointer ?, e.g. "cloth", "flour" [powder made from grain], etc)
- agent-event (coded with pointer *, e.g. "doctor", "employer", etc) (may also be "experiencer" or "relevant", depending on the type of event)
- patient-event (coded with pointer $, e.g. "patient", "employee", etc) (may also be "content" or "possession", etc. depending on the type of event)
- instrument-event (coded with pointer *, e.g. "watch", "computer", etc)
- location-event (coded with pointer @, e.g. "bank", "hospital", "shop", etc)
- time-event (coded with pointer @, e.g. "holiday", "semester", etc)
- value-attribute (coded without pointer, e.g. "blue", "slow", etc)
- entity-value (coded without pointer, e.g. "fat", "fool", etc)
- event-role (coded with role-name, e.g. "crying", "shopping", "swelling", etc)
- concepts co-relation (coded with pointer #, e.g. "cereal", "coalfield", etc)

A notable characteristic of HowNet is that synonyms, antonyms and converse relation

s can be generated by the users themselves based on the rules for synonym relation.

HowNet is a knowledge system, not a semantic dictionary.

All documentation on HowNet, including the Knowledge Dictionary forms an knowledge system. The main documentations of HowNet:

- Main Features of Concepts,
- Secondary Features of Concepts,
- Synonymous, Antonymous and Converse Relations (SACR),
- Event Relatedness and Role-Shifting (ERRS).

They can be used in conjunction with the Knowledge Dictionary.

2.5.2.4 Methodology

- As a knowledge system that describes relations between concepts, HowNet is not a thesaurus.
- HowNet attempts to construct a graph structure of its knowledge base from the inter-concept relations and inter-attribute relations.

This is the fundamental distinction between HowNet and other tree-structure lexical databases.

The philosophy of HowNet and its very nature underlined its unique method of building.

2.5.2.4.1. Extraction of Sememe

The principles for extraction of sememe:

- A sememe refers to the smallest basic semantic unit that cannot be reduced further. Take for instance "human being", despite being a most complex concept encompassing a set of attributes, it can be regarded as a sememe. All concepts can be reduced to the relevant sememes.
- There exist a close set of sememes, from which, composes an open set of concepts. If we can manage the close set of sememes to describe inter-concept relations as well as inter-attribute relations, an ideal knowledge base would be conceivable.
- Use the Chinese language to search for this close set of sememes. It is really trying a short cut. The Chinese characters (including simple word) is a close set that can be exploited to express both simple and complex concepts, as well as the inter-concept and inter-attribute connections.

The process of extraction of sememe:

The set of sememe is established on meticulous examination of about 6000 Chinese characters.

- ever extracted as much as 3200 sememes from Chinese characters (simple morpheme).
- After the necessary merger, 1700 sememes are derived for further classificatio

n that finally resulted in about 700 sememes. Note that up till this point, no polysyllabic words (in Chinese) are involved.

- These 700-odd sememes then served as a tagging set to tag polysyllabic words,
- In the process, the necessary adjustment and extension were made when the set cannot satisfy the requirements.
- Finally the process arrived at a set of over 800 sememes now using in HowNet.

HowNet also use some English words as sememe. E.g. HowNet would extract a common event sememe, "treat1" (provide medical treatment for) from the following English word: doctor, patient, hospital, medicine, therapy…

In sum, the building of HowNet is a bottom-up grouping approach. The first step is to form a tagging set of sememe through detail studying of all fundamental sememes and then apply tests to perfect the sememe list.

2.5.2.4.2. Examination and Confirmation of Sememes

At the formation of an initial list of sememes grouped to serve as a basic tagging set, the issues of examination and confirmation arise.

- check the coverage of the list of sememes against an extended scope of corpus annotation.
- examine the status of specific sememes in the concept network. If a sememe stands out among other concepts in either the same or a different category, then, it is a stable sememe that must be kept.

2.5.2.5. Preview to HowNet Knowledge System.

2.5.2.5.1 Database and Documentation of HowNet knowledge system

The HowNet knowledge system includes the following database and documentation:

- HowNet Management System
- Chinese-English Bilingual Knowledge Dictionary

The scale of HowNet depends on the size of its Chinese-English Bilingual Knowledge Dictionary. It includes 50,000 Chinese words or phrases corresponding 60,000 concepts, 55,000 English words or phrases corresponding 70,000 concepts.

2.5.2.5.2 Record Format in HowNet Knowledge Dictionary

The HowNet Knowledge Dictionary is the heart of the whole system. In this Dictionary, every concept of a word or phrase and its description form one entry. Regardless of the language types, an entry will comprise four items. Every item is made up of two portions joined by the "=" sign. To the left of the "=" sign is the data field, while that on the right is the data value. The items are arranged in the following sequence:

NO.
W_X= word / phrase form
G_X = word / phrase syntactic class

E_X = example of usage
DEF = concept definition

Here “X” means Chinese or English.
e.g.

NO.=005756
W_C=病 (bing)
G_C=N
E_C=
W_E=disease
G_E=N
E_E=
DEF=disease

NO.=092273
W_C=医生 (yisheng)
G_C=N
E_C=
W_E=doctor
G_E=N
E_E=
DEF=human, *cure, medical

NO.=034930
W_C=患者 (huanzhe)
G_C=N
E_C=
W_E=patient
G_E=N
E_E=
DEF=human, *SufferFrom, $cure, #medical

NO.=102368
W_C=治病 (zhibing)
G_C=V
E_C=
W_E=treat a disease
G_E=V
E_E=
DEF=cure, content=disease, medical

2.5.2.5.3 Selection of Words and Phrases and their Concepts

As it is known that the knowledge dictionary of HowNet is based on Words and Phrases and their concepts.

The process for selection of words and phrases and their concepts:

- Firstly, Chinese language doesn’t have the words in as strict sense as that in European languages. The selection of words and phrases mainly from 80,000 words and phrases with usage frequency out of a very large corpus with 400 million Chinese characters, rather than from any current Chinese dictionary. Much attention has been paid to those currently popular in usage, such as "Internet", "Euro", "dioxin", and "download", "click" or " hacker " in computer subject.
- Secondly, for the selection of concepts or meanings, the careful attention has

been paid to the popularity of any meaning of a word or phrase, usually only to choose those meanings that are still in use. and those obsolete ones will be discarded.

- Thirdly, to check if the description of meanings in the Chinese-English bilingual dictionary will fit both languages.