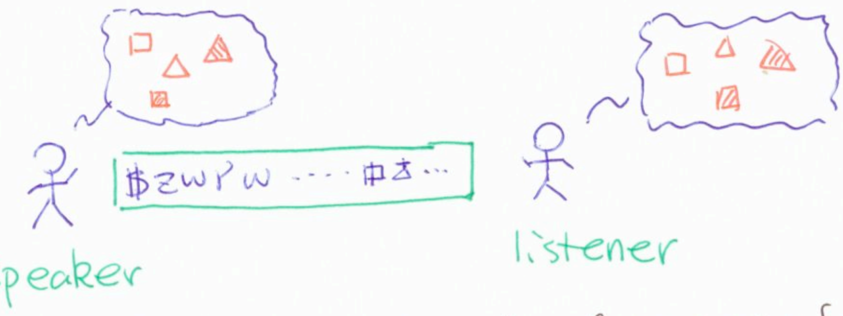
1. Natural Language processing Basics

1.1. what is language for? - Communication



A person has a representational space of semantics, by using sounds, words or even gestures to tell others about what you are thinking.

1.2 Two approaches to language

"Universal Grammar"

Rationalist (Poverty of the Stimulus, Chomsky)

Empiricist (emphasis on learning rather than hardwired principals of human brains)

1.3 Grammaticality and Conventionality

Colorless green ideas sleep furiously

grammatically correct!

Sheep year China Feburary begin

unconventional but Semantically understandable.

* A good novelist should be with an imaginable mind as well as a technique of writing or Communicating with words.

1.4 why Statistical?

- ☑ human cognition is probabilistic. So as the language.
- ☑ Computing probability of sentence from a corpus of utterances would assign the same low probability to all untested sentences, grammatical or ungramnatical, which is not true in linguistics (Chomsky)

1.5. Statistical analysis of language - Some examples.

- ☑ example of Tom Sawyer by Mark Twain.
  8018 words for the novel (11,100 for news)
- ☑ How to Count words - stemming
- ☑ Zipf's Law $f \propto \frac{1}{r}$

  frequency rank

- ☑ Mandelbrot's formule

  $f = P(r+\rho)^{-B}$ or by taking logarithm:

  $\log f = \underline{\log P} - \underline{B} \log(r + \underline{\rho})$

  parameters to make Zipf's law more practical based on the texts given.

1.6 Can you tell an author's style from statistical analysis of their works?

It will be great to try! our first homework!

1.7. Mathematical Foundation for NLP!

- ☑ Essential probability theory
- ☑ Conditional probability and Bayes theroem
- ☑ Statistical distributions.
- ☑ estimation
- ☑ Entropy (Information Theory)
- ☑ Noisy Channel Model

1.8 Zipf's law in Chinese language Research Question!

(Character-based, or word-based)

Zipf-Mandelbrot Law extension.

1. Language Model

For Some NLP tasks as the following

A. Machine Translation

$P($我是一个学生$\mid$ I am a Student$)$

B. Pinyin (spelling)

$P($我是一个学生$\mid$ wo shi yi ge xue sheng$)$

C. Speech Recognition

$P($我是一个学生$\mid$ ～～～～$)$

The Signals sounds like

D. Word segmentation

P(我|是|一个|学生)　　P(我|是|一个学|生)

P(我是|一个|学生)　　P(我|是一|个学|生)

which is bigger??

All these problems are related to one central question. what are the probabilities of particular Sentences?

$$P(e) = ?? \qquad e = \{\omega_1, \omega_2 \cdots \omega_N\}$$

E.g. in machine translation:

$$P(\underset{\sim}{c}\mid\underset{\sim}{e}) \propto P(e\mid\underset{\sim}{c})\,P(\underset{\sim}{c}) \quad \text{(Bayesian Rule)}$$

English　　Chinese

$P(e)$ ~ likelihood of an English sentence

$P(c\mid e)$ ~ probability of a Chinese sentence c given an English sentence e.

2. It is always too hard to estimate the probability of a sentence (why?). we then go for the probability of substrings of words. E.g.

$$P(e) = P(\omega_1, \omega_2, \omega_3 \cdots \omega_N) \underbrace{\approx \prod_{i=1}^{N} P(\omega_i)}_{\text{A good model??}}$$

2.1 $p(e) = p(w_1, w_2 \cdots w_N)$

Independence assumption of words (A Bag-of-words)

$$p(w_1, w_2 \cdots w_N) = p(w_2, w_3 \cdots w_N, w_1) \quad \text{unigram}$$
$$= p(w_3, \cdots w_{N-1}, w_N, w_1, w_2)$$
$$= p(\text{All possible permutations})$$

am
a Student
I of
Beihang

The word order doesnot matter (really?)
how about the letter order in a word?

2.2 $P(\text{I am a student}) = p(\text{I} \mid \text{Start-of-sentence}) * P(\text{am} \mid \text{I}) * p(\text{a} \mid \text{am}) * p(\text{Student} \mid \text{a}) * p(\text{end-of-sentence} \mid \text{Student})$

$$P(w_1, w_2 \cdots w_N) = \prod_{i=1}^{N-1} P(w_{i+1} \mid w_i) \quad \text{Bi-gram}$$
$$P(w_1, w_2 \cdots w_N) = \prod_{i=1}^{N-2} p(w_{i+2} \mid w_{i+1}, w_i) \quad \text{tri-gram}$$

Markov property

2.3 Smoothing in N-gram Model

$$P(w_{i+2} \mid w_{i+1}, w_i) \rightarrow p(z \mid x, y) = \frac{\#(x, y, z)}{\#(x, y)}$$

$\#(\cdot)$: Counting occurrences

By smoothing:

$$P(z \mid x, y) = 0.95 * \frac{\#(x, y, z)}{\#(x, y)} + 0.04 * \frac{\#(y, z)}{\#(y)} + 0.008 * \frac{\#(z)}{\#(\text{all words})} + 0.002$$

$p(z \mid x, y) \geq 0.002$, these smoothing coefficients are manuely assigned.

- P(I want to eat British food) = P(I|<start>) P(want|I) P(to|want) P(eat|to) P(British|eat) P(food|British)
  = .25*.32*.65*.26*.001*.60 = .000080

Eat on	.16	Eat Thai	.03
Eat some	.06	Eat	.03
Eat lunch	.06	Eat in	.02
Eat dinner	.05	Eat	.02
Eat at	.04	Eat	.02
Eat a	.04	Eat	.01
Eat Indian	.04	Eat	.007
Eat today	.03	Eat British	.001

<start> I	.25	Want some	.04
<start> I'd	.06	Want Thai	.01
<start> Tell	.04	To eat	.26
<start> I'm	.02	To have	.14
I want	.32	To spend	.09
I would	.29	To be	.02
I don't	.08	British food	.60
I have	.04	British restaurant	.15
Want to	.65	British cuisine	.01
Want a	.05	British lunch	.01

2.4 Perplexity

Based on what we have learnt from last a few sections on language model. A model often consists of a generative "story" (people produce words based on the last one (two or more) words they said).

uni, bigram or trigram

$$P(\text{model} \mid \text{test-data}) \propto \underbrace{P(\text{model})}_{\text{uniform unless we have more information.}} * P(\text{test-data} \mid \text{model})$$

$$P(\text{test-data} \mid \text{model}) = P(e) = \underbrace{\prod_{i=1}^{N} P(w_{i+1} \mid w_i)}_{\text{product of small numbers.}}$$

$$\text{perplexity} = 2^{-\log(P(e))/N}$$

$\prod_j a_j \rightarrow \log \sum_j a_j$

In information theory, perplexity is used to measure how well a probability distribution or probability model predicts a sample, used to compare probability models.

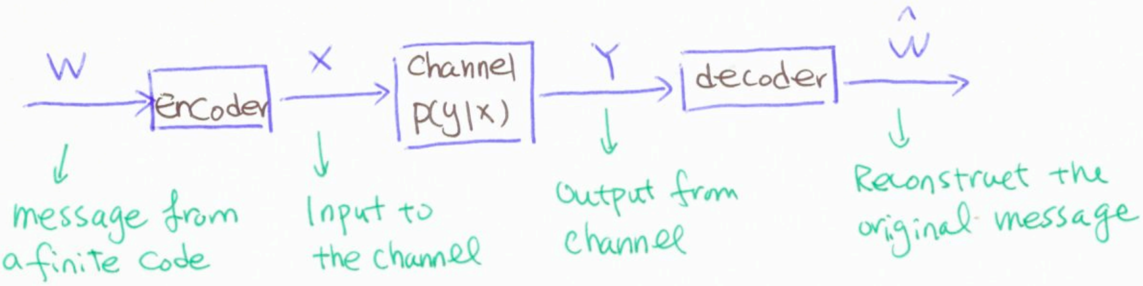
(First used by Frederick Jelinek, whose legacy in NLP can be remembered by his famous saying: "Everytime I fire a linguist, the performace of the speech recognizer goes up")

- ☑ lower perplexity implies better language models.
- ☑ perplexity of a probability distribution is:

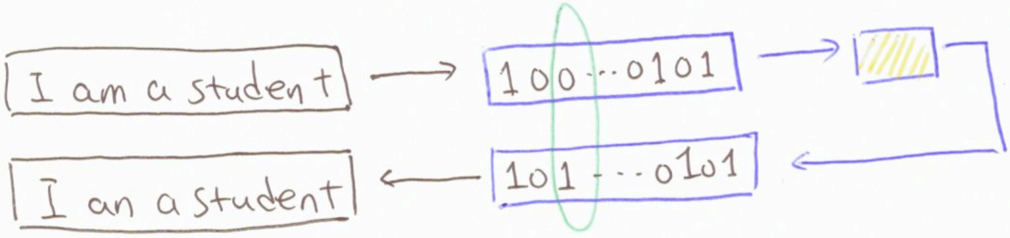
$$2^{H(P)} = 2^{-\sum_x P(x)\log_2 P(x)} \qquad H(P) \sim \text{entropy}$$

- ☑ How to estimate the entropy of a language? Obviously, the entropy is related to the language model.
- ☑ Based on reseach of Brown et.al (1992), the upper bound of entropy of characters in printed English is 1.75 bits per character. How?

2. Information Theory for NLP.



Classical noisy channel model was proposed by Shannon as the foundation of the Information Theory.



An example of Communication.

2.1 why the noisy Channel Model is important for NLP?

☑ Machine Translation

Input :	L_1 word sequence	我是学生
Output :	L_2 word sequence	I am a student
$p(i)$:	$p(L_1)$ language model	p(我)p(是)p(学)p(生)
$p(o\|i)$:	translation model	p(I|我) ⋯ p(is|是)

☑ Speech recognition

Input : word sequence
Output : speech signal.
$p(i)$: probability of word sequence
$p(o|i)$: acoustic model (HMM)

From the above, we can easily see how the information theory is related to machine learning. Most of the learning cases, we just need the right model of $p(i)$ and $p(o|i)$, where "i" is examples and "o" is classes.

Any disussions or Comments??

2.2. Entropy

Information measure $I(x)$

Def: $$H(X) = E\left(\boxed{\log \frac{1}{P(X)}}\right) = \sum_i P(x_i) \log \frac{1}{P(x_i)}$$

$$= -\sum_i P(x_i) \log P(x_i)$$

$$= -\sum_i P_i \log P_i$$

where $I(x)$ satisfy the following:

$$\begin{cases} I(x_1, x_2) = I(x_1) + I(x_2) \\ I(x) \propto \frac{1}{P(X)} \end{cases}$$

{ that is why we use this equation to define the content of information

2.3 Joint entropy

$$H(X,Y) = \sum_x \sum_y P(x,y) \log \frac{1}{P(x,y)}$$

$$= -\sum_x \sum_y P(x,y) \log P(x,y)$$

2.4. Conditional entropy

$$H(Y|X) = \sum_{x_i} P(x_i) H(Y|x_i)$$

$$= \sum_x P(x) \left[-\sum_y P(y|x) \log P(y|x)\right]$$

$$= -\sum_x \sum_y P(x,y) \log P(y|x)$$

So that: $H(X,Y) = H(X) + H(Y|X)$

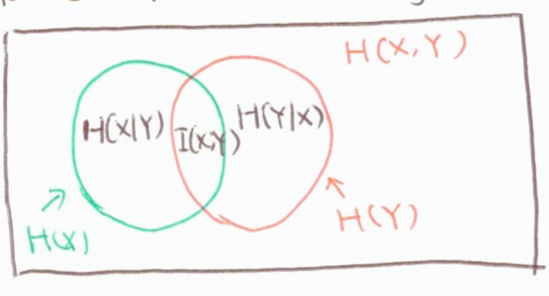
2.5 Mutual information

Since $H(X,Y) = \underline{H(X) + H(Y|X)} = \underline{H(Y) + H(X|Y)}$

(remember $p(x,y) = p(x)\,p(y|x) = p(y)\,p(x|y)$?)

then: $H(X) - H(X|Y) = H(Y) - H(Y|X) = I(X;Y)$

is named the mutual information.



Information Relations

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

According to a researcher at Cambridge University, it doesn't matter in what order the letters in a word are, the only important thing is that the first and last letter be at the right place. The rest can be a total mess and you can still read it without problem. This is because the human mind does not read every letter by itself but the word as a whole.