Statistical Machine Translation

LECTURE – 2

PRELIMINARIES

April 13 , 2010

Topics of Discussion

- Aproaches to MT
-Word, Sentence, Corpora
-Probability
-Bayes Rule
- Entropy
-Regression

Approaches to MT

Direct Model

No linguistic representation is needed.

Words or sequence of words (n-grams) are translated. (Basically word-for-word translation/ or literal translation)

Transfer Model

Here knowledge about the difference between the source and target language is used

Basic steps:

- Perform analysis of the source sentence – both lexical & syntactical

-Transfer (map) the structure into target language

- Generate corresponding sentence in TL

Interlingua Model

Here representation is based on semantics – which is in principle language-independent

Basic Steps:

- Analyze SL sentences lexically, syntactically and semantically.

- Interpret the meaning in Interlingua

- Transfer the meaning from Interlingua to TL

(Advantage: O(n) growth NOT O(n^2))

Example

Gallia est omnis divisa in partes tres. <-- Latin text

Gallia is all divided in parts three. <-- literal translation

Gaul is divided into three parts. <-- Syntactic translation

Gaul has three divisions. <-- More semantic translation

Example

The lady is looking beautiful
The woman is looking good.

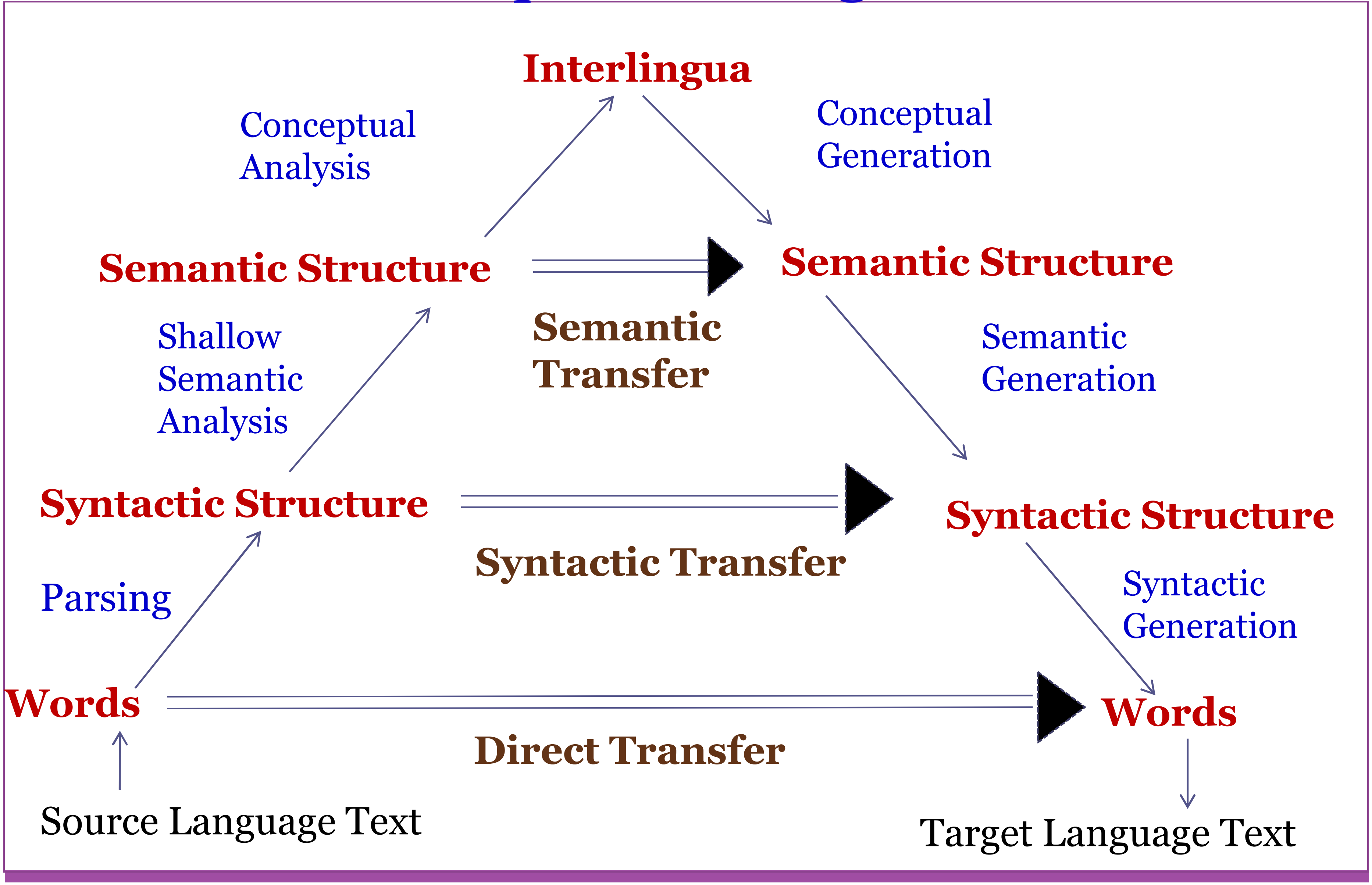
} Different lexicons
Same syntax

The lady is looking beautiful
The lady is beautiful looking

} Same lexicons
Different syntax

But Knowledge representation Techniques e.g. Conceptual Dependency (Schank 1972) allows us to represent these facts independent of language

Vauquois’ Traingle



Interlingua

Conceptual Analysis

Conceptual Generation

Semantic Structure

Semantic Structure

Semantic Transfer

Shallow Semantic Analysis

Semantic Generation

Syntactic Structure

Syntactic Structure

Syntactic Transfer

Parsing

Syntactic Generation

Words

Words

Direct Transfer

Source Language Text

Target Language Text

Word - Sentence - Corpora

Word

Word - Basic atomic unit of meaning.

[

Sub-word level entities: letter, syllable are typically not much useful from translation perspective.

Although for some languages letters used as *inflections* are very important.

In written text words are identifiable because of space.
(For Chinese, for example, there is no such inter-word space.)

For speech it is very difficult to identify.

Not as simple as it appears!!

Word

1) Even for texts it is sometimes difficult: - e.g. you're, it's, hyphenated words.

2) Even sometimes it is ambiguous : - she's
Used for "she has" and "she is"

3) For some languages it is not easy even for texts: German, Sanskrit.
In Sanskrit: words are joined based on syntactic rules (*sandhi*) and semantics (*samasa*)
Even in English : knoweldge base, but database

4) Mostly an MT works on words, although sometimes need to process *sub-word* level

Tokenisation

Basic processing step to break up text into words

Often challenging for languages like:

- Chinese : no spacing
- Sanskrit/German: Compound words

Even for English:

- Hyphenated words (e.g *co-supervisor*)
- Merged words (e.g. *They've*)
- Possessive case (e.g. *John's*)
- Abbreviations: UCLA, SUNY, UFO

Tokenization of English text also requires:
truecasing/decasing, detokenization for punctuation

Distribution of Words

Statistical analysis of texts needs to find distribution of word:

Content words - Objects, Actions, Properties

vs.

Functional words – Relation between Content Words

Typically top frequency words are Functional

Zipf's Law: rank x frequency = constant

(linearity of logs: log f = log c – log r)

Content vs. Function Words

Both pose problem for Machine Translation

Content Words: Too many
Numbers are increasing
Difficult to capture all in texts
WSD may needed
(e.g. bank, plant)

Typically: Noun, Verb, Adjective and Adverb

Content vs. Function Words

Function Words: Limited in Number for any language.
Typically: Article, Preposition, Conjunction etc.

May have specific role in a language, but NOT
in the other.

e.g. Hindi, Chinese - no equivalent for *the*

Some languages have many definite articles depending
Upon number and gender:

Italian has: *il, lo, i, gli, la, le*, l‘*,* lo stesso

Some languages (e.g. German, Bengali) article depends
on the case of the corresponding noun.

Content vs. Function Words

Roles of prepositions is difficult to understand:

e.g. Meat with pasta Vs. Meat with fork
Run into trouble Vs. Run into a room

Some words may have to be replaced with more than one word:

What will be will be -> **jaa** hobar **taa** hobe (Bengali)

Content vs. Function Words

Should a translator rely only on key words?

Consider: She looks beautiful

She is beautiful to look at

She is beautiful looking.

Vs.

The horse runs good

This is a good run by the horse

The horse is good to run on

Hence content words alone do not serve MT.

Some Other Words

A tokenizer also needs to identify some Special purpose words and needs to interpret them properly:

- Interjections
- Foreign words
- List/item Numbers
- Headings
- Numerals
- Abbreviations
- Spl. Symbols (e.g. currency.)
- Emoticons

Detecting POS is a well-known NLP problem: POS-tagging
Typically a large amount of tagged text is used to train a tagger.

POS Tagging

English language has many POS tagger.
Here is an example from Stanford POS tagger

<l>As/IN slow/JJ our/PRP$ ship/NN her/PRP$ foamy/JJ track/NN ,/,</l>   <l>Against/IN the/DT wind/NN was/VBD cleaving/VBG ,/,</l>

Some standard tags are:

NN	Noun	VB	Verb base	JJ	Adjective
NNS	Noun pl	VBD	Verb past	JJR	comparative
NNP	P. Nn s.	VBG	Verb gerund	JJS	superlative
NNPS	P. Nn p	VBZ	Verb 3rd sing	PRP	Personal Pronoun
DT	Determiner	RB	Adverb	PRP$	Possessive Pronoun
PDT	Predeterminer	IN	Preposition	SYM	Symbol

For resource-poor languages it is a necessity.

Morphology

Morphology is the identification, analysis and description of the structure of words. (Wikipedia).

While words are generally accepted as being the smallest units of syntax, it is clear that in most (if not all) languages, words can be related to other words by rules.

E.g: go (root word) has several variations:
going, goes, gone, went

Morphology

A translation system needs to take care of the Morphological variations to understand the roles of each word in the sentence.

Different languages vary in their number, and Usage of their morphological forms.

E.g the same verb *go ~ jaanaa* has so many more variations in Hindi: jaa, jaataa, jaayegaa, jaaoge, jaaoongii, jaayenge, gayaa, gayii, gaye, ……

And many more based on Number, Gender, Tense

In Sanskrit each verb has 45 morphological variations!!

Morphology

Let us consider an example:

Lion eats deer ->

Der Löwe frißt das Reh
Das Reh frißt der Löwe

Changing the relative position does not change
The semantics: But changing the morphology does:

Eg. *Den Löwen frißt das Reh*
Das Reh frißt den Löwen

[Taken from: Philip Koehn]

Compare it with English: John loves Mary
vs.
Mary loves John

Morphology

Very similar thing happens in Hindi also.

Paris killed Achilles :

paris ne achiles ko maaraa
achiles ko paris ne maaraa
maaraa achiles ko paris ne
achiles ko maaraa paris ne

Etc. Are all permissible.

However the problem is the Morphological suffix is not always attached with the words.

Sentence

Sentence

Defined to indicate a grammatical and lexical unit consisting of one or more words that represent distinct concepts. Sentence can include words grouped meaningfully to express a statement, question, exclamation, request or command.

(Wikipedia)

Typically in MT we look at sentences at the next level. Although in modern approaches people are looking at phrases also.

Sentence

Semantically the central element of a sentence is its verb.

Other words attached to a verb can be:

- Subject
- Object (direct / indirect / none
  based on the *valency* of the verb)
- Adjuncts (providing additional information)
  (e.g time, adverb, other comments)

MT needs proper identification of each component.

Sentence

The standard problems are:

- recursive nature of language Eg. Phrases

(a **phrase** is a group of words functioning as a single unit in the syntax of a sentence)

(John bought the house at the end of this road)

- Clauses

(A clause will have a verb that can act as   argument or adjunct

(John bought the house which is at the end of the road)

-Prepositional Phrase attachment:

Not only structural – there is problem with scope also

>> I'll go to the market and then play cricket with my brother.

- Adjective attachment (problem with scope)

>> This is my favourite friend's picture.

Sentence

Their translations may differ across languages:

The girl in white dress (Eng)

→ *das Mädchen im weißen Kleid* (German)

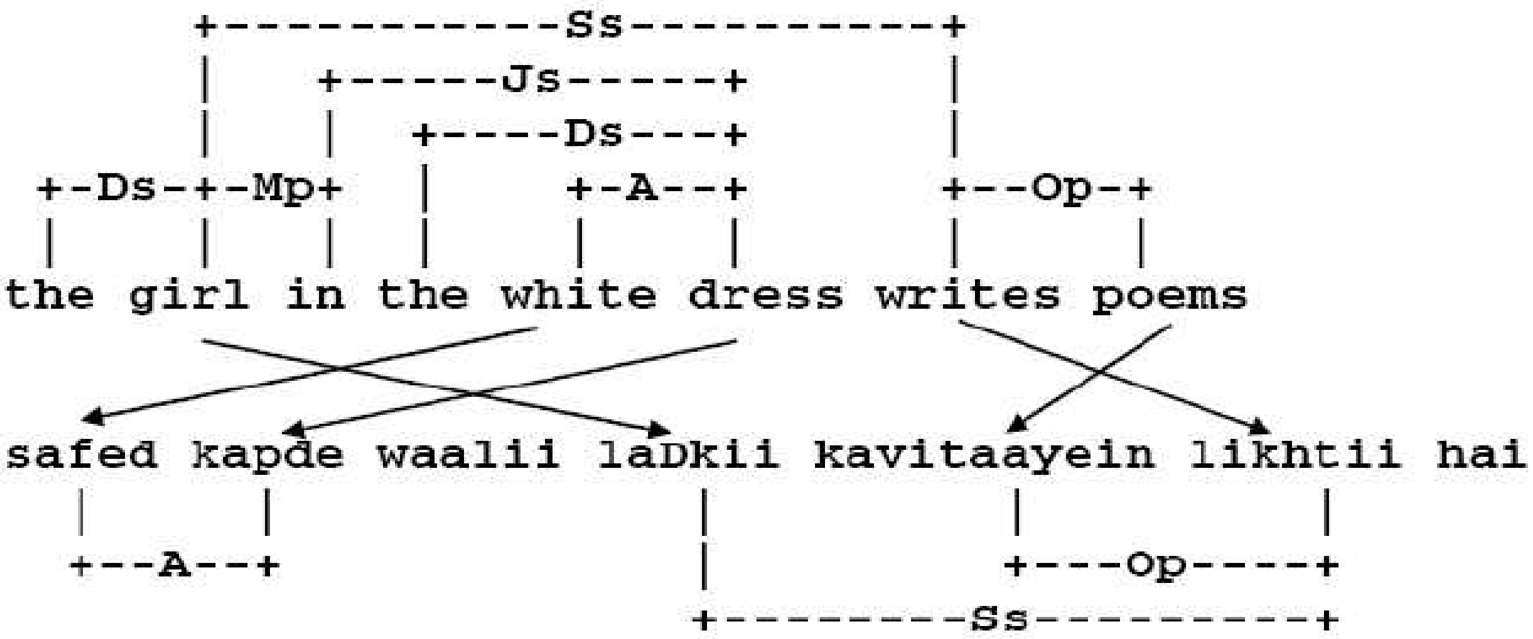
→ *la ragazza in abito bianco* (Italian)

BUT

→ *safed kapde wali ladki* (Hindi)
white dress having girl

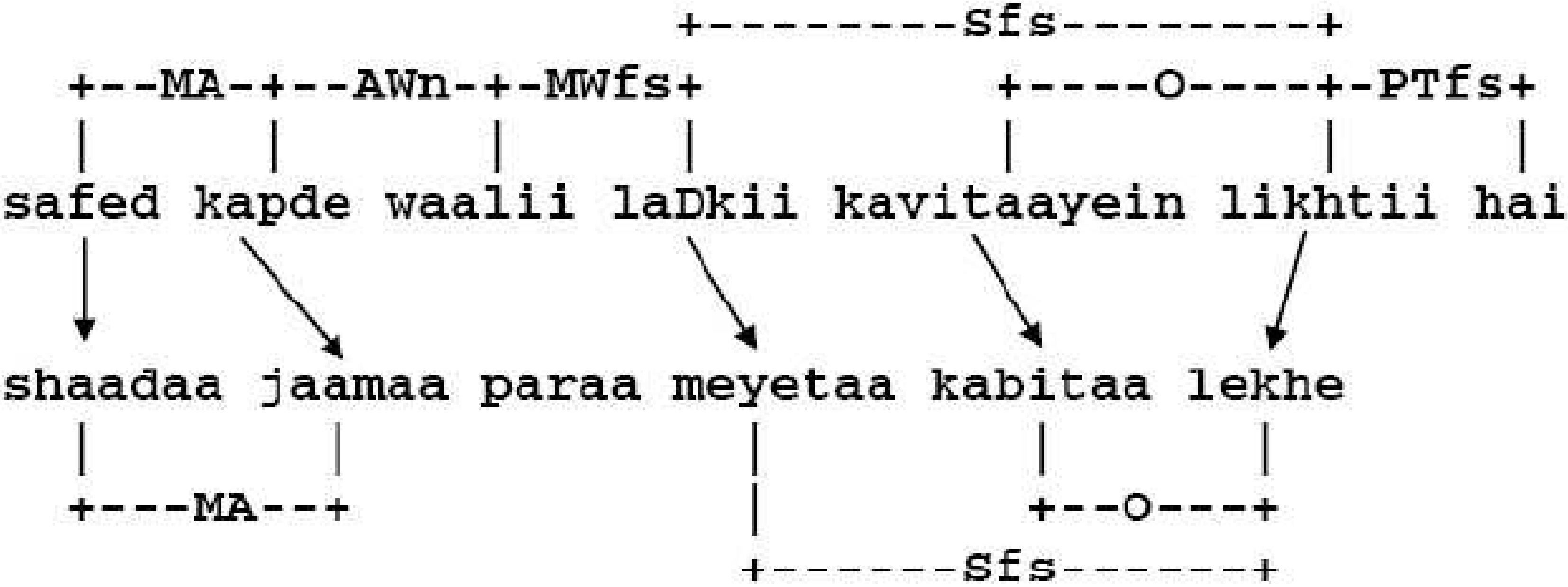
→ *shada jaamaa paraa meyeta* (Bengali)
white dress wearing the girl

Difficult to identify word - to- word correspondence



```
    +------------Ss-----------+
    |    +-----Js-----+       |
    |    |   +----Ds---+      |
+-Ds-+-Mp+   |    +-A--+      +--Op-+
|    |   |   |    |    |      |     |
the girl in the white dress writes poems

safed kapde waalii laDkii kavitaayein likhtii hai
  |      |             |          |          |
  +--A---+             |          +---Op-----+
                       +--------Ss-----------+
```



```
                     +--------Sfs--------+
  +--MA-+--AWn-+-MWfs+         +----O----+-PTfs+
  |     |      |     |         |         |     |
safed kapde waalii laDkii kavitaayein likhtii hai
```

```
shaadaa jaamaa paraa meyetaa kabitaa lekhe
  |       |            |        |      |
  +---MA--+            |        +--O---+
                       +------Sfs------+
```

Sentence

A parser can be used to understand the roles of different words in a sentence.

Different types of grammar:

- Phrase Structure
- Dependency structure
- Context Free grammar

We shall concentrate on the first two.

Discourse

Discourse

Although we often look at single sentence translation. Often we like to translate long passages

Sentences are correlated – having certain relationships.

- co-reference

  President Pratibha Patil of India ....

  ....

  The *president* said ....

While translating "president" may be considered in a different way (e.g. club president, company president)

- anaphora

E.g John came to his brother's house. *He* was in hurry.

Discourse

Topic of a discourse may sometimes be useful in getting right translation:

- I am going to the bank. I have to withdraw money.

  >> Sto andando in banca. Devo ritirare del denaro. (sense 1)

- I am going to the bank. I love watching sailing *boats*.

  >> Sto andando in banca. Mi piace guardare le barche a vela.

  (sense 1)

  (wrong!!! Should be Sto andando alla riva)

- The green plant looks beautiful. But it emits dark *smoke*.

  >> L'impianto di verde sembra bello. Ma emette fumo nero

  (sense 1)

- The green plant gives red fruit >>

  La pianta verde dà frutti rossi. (sense 2)

Corpora

Corpora

Developing SMT system needs analysis of data – corpora.

In Linguistics it is: a body of utterances, as words or sentences, assumed to be representative of and used for lexical, grammatical, or other linguistic analysis.

Corpora can be text / speech.

However, speech has several difficulties:

- incomplete sentence, fluff, filler words

Hence corpora needs to be cleaned.

Speech translation has certain advantages: *modality, utterance*

Parallel Corpora

A parallel corpus is a collection of text paired with its translation in another language.

For European languages:

- Europarl (*www.statmt.org/europarl/*)
- Acquis (*wt.jrc.it/lt/ www.ldc.upenn.edu/*)

For English-Asian

- Emillee

Are well-known parallel corpus.

Parallel Corpora

Most MT related research work e.g.

- statistical analysis
- knowledge extraction

Are based on parallel corpora.

Basic preparations:

- Corpus cleaning , EOS marker
- POS tagging
- Chunking
- Sentence alignment

(Translation is not always 1 – 1).

Parallel Corpora

Chunks are often identified as a translation Unit:

(a method for parsing natural language sentences into partial syntactic structures)

Example of chunking:

United	→	united (I) (wrong!!)	*sammilito* (B)
States	→	Gli Stati (I)	*raajyoguli* (B)
Of America	→	d’America (I)	*america-r* (B)

United States of America →

Stati Uniti d'America (I)

america yuktarashtra (B)

Parallel Corpora

Alignment examples:

(A) Shylock the jew was a usurer →

1 -2 *shylock ek yuhudi thaa.*
woh sudkhor thaa.

(B) Shylock the jew was a usurer.
He used to live in Venice. →

2-1 *yuhudi shylock ek sudkhor thaa,*
jo venice mein rahtaa thaa.

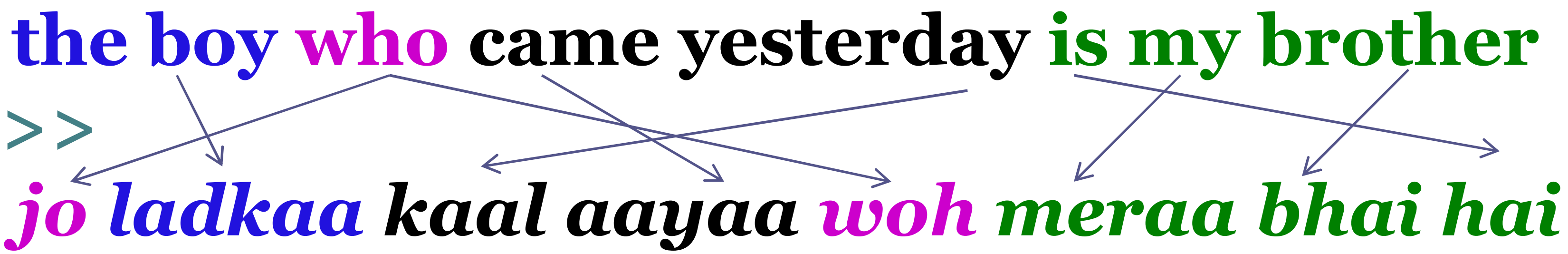
2-3 *shylock a yuhudi thaa.*
who sudkhor thaa..
woh venice kaa rahanewale thaa

Parallel Corpora

Just like sentence alignment one may think of:

- Phrase alignment
- Paragraph alignment

E.g.



the boy who came yesterday is my brother

>>

jo ladkaa kaal aayaa woh meraa bhai hai

(H)

- Word alignment

often bi-lingual dictionaries are useful.

Speech

Basics

Speech is quite different from text.

Typically speech a continuous flow of waves, digitized For various speech processing.

Has the advantage of intonations.

Lots of disadvantages: lack of inter-word gap, slips, homonym/homophones, incomplete sentences, monotonicity, articulation variations.

Still lots of speech processing activities are going on.

Hidden Markov Model is used.

Basic statistics and Probability

Preliminary Statistics

Aim: To automatically analyze existing human sentence translations, in order to gather general translation rules.

We will use these rules to translate new texts Automatically.

Basic Probability

Probability

Probability Mass Function
Probability Density Function
Probability Distribution Function
Standard Probability Distributions:
Binomial, Poisson, Gaussian Distribution

Concept of a Random Variable

Joint Probability Distribution and Independence

Conditional Probability Distribution (Marginal Distributions)

Bayes Rule

$$P(X|Y) = \frac{P(Y|X)\ P(X)}{P(Y)}$$

This rule expresses a conditional density P(X | Y) in terms of :

- Its inverse P(Y | X) (called posterior density)
- P(X) (Prior Density)

and

P(Y) (which is less significance)

Bayes Rule

One use can be in Bayesian Model Estimation.

Suppose we have two random variables - data sample D and a model M.

We are interested in the most probable model that fits the data.

$$\arg\max_M P(M \mid D) =$$

$$\arg\max_M \frac{P(D \mid M)\, P(M)}{P(D)} =$$

$$\arg\max_M P(D \mid M)\, P(M)$$

Bayes Rule - Example

Suppose a coin is tossed 100 times and 40 of them were heads. What is the Model?

Model can be characterized by *p* - probability of Head..

We shall choose the p that maximizes P(40 Heads)

Assume wolg that all *p*'s 0.1 0.2 .. 0.9 are equally likely. So P(M) is constant.

Essentially we need to find p that maximizes P(40 H).

Given p, P(42 H) = $^{100}C_{42}\, p^{42}(1-p)^{58}$

The p that maximizes this value is???

Bayes Rule - Example

One simple way is to tabulate and check:
We can as well compute logs

p	42 log p	58 log (1-p)	Sum
0.1			
0.2	-29.36	-5.62	-34.98
0.3	-21.96	-8.98	-30.94
0.4	-16.71	-12.87	-29.58
0.5	-12.64	-17.46	-30.10
0.6	-9.31	-23.08	-32.39
0.7			
0.8			
0.9			

Note that you cannot get it using Differentiation !!

Lagrange Multiplier

But a more mathematical way of doing this is using Lagrange multiplier.

- It is technique to find maxima or minima (in general, "extrema") of a function.
- It is often difficult to find a closed form for the function being extremized. when one wishes to maximize or minimize a function subject to fixed outside conditions or constraints.
- The method of Lagrange multipliers is a powerful tool for solving this class of problems

The theory is as follows:

Lagrange Multiplier

Suppose we want to maximize a function *f(x,y)*.
Subject to the condition *g(x,y)* = *c*

We consider a new function $F(x,y,\lambda) = f(x,y) + \lambda (g(x,y) - c)$

We then take partial derivatives with respect to all the variables
Including λ. Thus we have:

$$\frac{\delta f}{\delta x} + \lambda \frac{\delta g}{\delta x} = 0 \qquad (1)$$

$$\frac{\delta f}{\delta y} + \lambda \frac{\delta g}{\delta y} = 0 \qquad (2)$$

$$g(x) - c = 0 \qquad (3)$$

The number of equations
Will depend upon

- The no. of variables
- The no. of constraints

Lagrange Multiplier

With respect to the above problems we have A single variable p. Let us assume that we can put the constraint as

$(p - 0.4) * (p - 0.5) = 0$

Thus we have to maximize: $f(p) = 42 \log p + 58 \log (1 - p)$

subject to: $(p - 0.4) * (p - 0.5) = 0$

Thus our function is:

$$42 \log p + 58 \log (1 - p) + \lambda (p^2 - 0.9 p + 0.2) = 0$$

After differentiation we have:

$$\frac{42}{p} - \frac{58}{1-p} + \lambda(2p - 0.9) = 0 \qquad (1)$$

$$(p^2 - 0.9 p + 0.2) = 0 \qquad (2)$$

Concept of Entropy

Entropy

Roughly – it measures the disorder.

With respect to Probability Theory - measure of uncertainty of outcomes.

The higher the entropy – the more is the Uncertainty.

In SMT Entropy based models are used to measure Perplexity of a sentence based on probability.

If X is a r.v. Then its entropy is:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Entropy

Examples:

$S = (x_1/1) \rightarrow \quad H(X) = 1.\log 1 = 0$

$S = (x_1/½, \ x_2/½) \rightarrow \quad H(X) = 1$

$S = (x_1/1/8, \ x_2/\ 7/8) \rightarrow H(X) \cong -\ 1/8 * -3 - 7/8 * (2.8 - 3)$

$= 3/8 + \ 1.4/8 < 1$

$S = (x_1/_{1/4}, \ x_2/_{1/4}, x_3/_{1/4} \ , \ x_4/_{1/4}) \rightarrow \quad H(X) = 2$

With each doubling of equally likely values, entropy Increases by 1.

Entropy

In information Theory Entropy is associated with the Minimum no. of bits required to encode the information.

E..g. 8 options. If all are equally probable: Entropy = 3

And we know that to encode we need 3 bits.

Now suppose the following probability distribution:

A1/½ A2/¼ A3/ $^1/_8$ A4/$^1/_{16}$ A5/ $^1/_{64}$ A6/ $^1/_{64}$ A7/ $^1/_{64}$ A8/ $^1/_{64}$

Entropy = ½ + ¼*2 + $^1/_8$*3 + $^1/_{16}$*4 + 4 ($^1/_{64}$*6) = 2 bits

And we can easily find the coding whose average cost is 2 bits.

Entropy Why Logarithm?

It gives additivity to uncertainty.

Consider two dice: {x1 .. x6} and {y1 .. y6}.

What is the uncertainty related with joint outcome?

Note there are 36 equally likely outcomes.

Case 1: Independent.

$$H(X,Y) = -\sum_{x\in X, y\in Y} p(x,y)\log_2 p(x,y) = -\sum_{x\in X, y\in Y} p(x,y)\ \log_2(p(x).p(y))$$

$$= -\sum_{x\in X, y\in Y} p(x,y)\ (\log_2 p(x)+\log_2 p(y)$$

$$= -\log_2 p(x) - \log_2 p(y) \quad = H(X) + H(Y)$$

In general H(X,Y) is called joint entropy.

Entropy : Why Logarithm?

Case 2: X and Y are not independent.

In this case we can think of conditional entropy.

$$H(Y \mid X) = -\sum_{x \in X} p(x) H(Y \mid X = x)$$

$$= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 (p(y \mid x))$$

$$= -\sum_{x \in X} \sum_{y \in Y} p(x) p(y|x) \log_2 (p(y \mid x))$$

$$= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 (p(y \mid x))$$

Now we show that: H(X,Y) = H(X) + H(Y|X)

Entropy : Why Logarithm?

To show : $H(X,Y) = H(X) + H(Y|X)$

$$H(X,Y) = -\sum_{x\in X, y\in Y} p(x,y)\log_2 p(x,y)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x)\,p(y|x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x) \; -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y|x)$$

$$= -\sum_{x\in X} p(x)\log_2 p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(y|x)$$

$$= H(X) + H(Y|X)$$

Thus uncertainty can be split into sum of two uncertainties

Mutual Information

For symmetricity one uses mutual information:

How much information is shared between two variables

Defined as: $I(X, Y) = \sum_{y}\sum_{x} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$

Note: - If X and y are independent: $I(X,Y) = 0$

- If X completely predicts Y: then p(x,y) = p(x). Hence $I(X,Y) = H(Y)$ – i.e. the shared information contains all the uncertainty of Y

Show that $I(X, Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
$= H(X) + H(Y) - H(X,Y)$

Linear Regression

Problem of Classification

Often we are challenged with the task of classify an Observation into a set of discrete classes:

- good translation or bad translation
- Word sense Disambiguation
- word / sentence alignment

In sentiment analysis – positive or negative sense.

The task of classification is to take an observation X and then put it in the right class.

Typically done by extracting some relevant features of X and then computing some function of the featural values.

Problem of Classification

For example, consider the quality of a translation:

Let f be a sentence translated into e.
One can think that the quality of translation depends
On the relative lengths. i.e. $length\ (e) = g\ (length(f\ ))$

Suppose g is a linear function. Hence we get:

$$m = w.n + c$$

In a more general case: $m = w_o + w_1.n + w_2.k$
where k is the number of words
translated non-monotonically

How to estimate w_is ?

Linear Regression

We train the system by N known values

We then have the following:

$$e_j = w_0 + w_1 n_j + w_2 k_j \forall j = 1, .., N$$

To solve this we minimize the SSE:

$$\sum_{j=1}^{N} (e_j - w_0 - w_1 n_j - w_2 k_j)^2$$

By differentiating w.r.t the w's get 3 Normal Equations.

Solve them to get the w_0, w_1, w_2

Logistic regression

Linear regression is useful for real-valued features.

What happens for discrete values?

binary classification: Good or bad.

WSD: A word may have many meanings:
which one to take?

Becomes more realistic to ask the model to give
Probabilities for different classes.

Our aim is to use Linear Regression Techniques here.

Logistic regression

Suppose we train our model to compute probabilities:

$$P(y = true \mid x) = \sum_{i=1}^{k} w_i f_i \quad k \text{ is the no. of features}$$

This can be solved by using regression – giving the training samples 0 or 1 value.

But we are not sure that for a new sample the predicted Value will remain in [0,1]. Rather it may lie (-∞, ∞)

Suppose we try to predict not Probability –Rather the *odds* ratio:

$$\frac{P(y = True \mid x)}{1 - P(y = True \mid x)}$$

This is close, but not OK !!!

Logistic regression

In Logistic regression we try to predict:

$$\ln\left(\frac{P(y=True \mid x)}{1-P(y=True \mid x)}\right)$$

This is called the logit function of p(x).

$$logit\ (p(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right)$$

A linear function is used to estimate the above.

$$\ln\left(\frac{P(y=True \mid x)}{1-P(y=True \mid x)}\right) = \sum_i w_i f_i \quad = w.f\ \ (say)$$

Logistic regression

Once the logit value is estimated, we can compute the probabilities:

$P(y = True \mid x)$ and $P(y = False \mid x)$

$$P(y=True \mid x) = \frac{e^{w.f}}{1+e^{w.f}} \qquad P(y = False \mid x) = \frac{1}{1+e^{w.f}}$$

Show that for Classification the Hyperplane equation is $\sum_i w_i f_i > 0$

Learning in Logistic Regression

Linear regression - learn the weights by Minimizing the SSE.

Here Conditional Maximum Likelihood Estimation is used.

- i.e. the weights are so chosen that the probability of Observed y values in the training data are highest.

$$\hat{w} = \underset{w}{\operatorname{argmax}} \prod_{i=1}^{N} P(y^{(i)} \mid x^{(i)})$$

Under log-likelihood the above equation becomes:

$$\hat{w} = \underset{w}{\operatorname{arg\,max}} \sum_{i=1}^{N} \log P(y^{(i)} \mid x^{(i)})$$

Learning in Logistic Regression

i.e.

$$\hat{w} = \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} \log \begin{bmatrix} P(y^{(i)}=1|\, x^{(i)}) \; if \;\; y^{(i)}=1 \\ P(y^{(i)}=0|\, x^{(i)}) \; if \;\; y^{(i)}=0 \end{bmatrix}$$

Or,

$$\hat{w} = \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} y^{(i)} \log P(y^{(i)}=1|\, x^{(i)}) + (1-y^{(i)}) \log P(y^{(i)}=0|\, x^{(i)})$$

Putting values for the probabilities

$$\hat{w} = \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} y^{(i)} \log \frac{1}{1+e^{-w.f}} + (1-y^{(i)}) \log \frac{e^{-w.f}}{1+e^{-w.f}}$$

A problem of Convex Optimization – many Algorithms are used.

Thank you