Machine Translation

NILADRI CHATTERJEE

DEPARTMENT OF MATHEMATICS

IIT DELHI

November 14, 2013

Workshop on Multilingual Technologies

Q: What is the Major Use of Computers today?

Almost invariably it is something to do with texts:

- Social Networking
- Email
- Searching for documents
- Annotation of Data, Image
- Writing documents
- Knowledge Engineering.

Consequently mono-lingual and Multilingual technology development takes high priority in today's India – both technically and linguistically.

Perhaps the major application of NLP is:

Machine Translation

Which is inherently a bilingual activity.

What is Machine Translation

- Machine Translation (MT) pertains to automated translations of text or speech from one natural language to another.
- MT aims at providing a tool for breaking the language barrier.
- In a multilingual environment (like EU, India) there may be two types of translation system:
  - Between two languages local to the environment
  - Between a foreign language and a local language
- Our focus is on text translation.

Difficulties in Dealing with Natural Languages

- Expression is not unique. The same sense may be conveyed in many different ways.
- Construction of sentences is governed by a set of rules or grammar. But often there are exceptions.
- How this information is organized in our brains is not known. Consequently knowledge representation in NLP systems is a significant area of research.
- This is true for different NLP applications. In this talk we shall focus on Machine Translation.

We start with some Examples:

Difficulties Example

- The girl is beautiful
- The girl is pretty
- The girl looks beautiful
- She is good-looking
- I find the girl beautiful
- To me the girl is beautiful

All these convey the same meaning. But can a MT system understand it?

Google Translator has the following outputs:

Difficulties Example

The girl is beautiful -> लड़की खूबसूरत है

মেয়ে সুন্দর (Meye Sundor)

આ છોકરી સુંદર છે (Ā chōkarī sundara chē)

The girl is pretty -> लड़की सुंदर है

মেয়ে সুন্দর হয় (Meye Sundor hoy)

આ છોકરી ખૂબ સુંદર છે (Ā chōkarī khūba sundara chē)

The girl looks beautiful -> लड़की सुंदर लग रहा है

মেয়ে সুন্দর দেখায় (Meye sundor dekhaay)

આ છોકરી સુંદર દેખાય છે (Ā chōkarī sundara dēkhāya chē)

Difficulties Example

The girl is good-looking -> लड़की अच्छी लग रही है

মেয়ে সুদর্শন হয় (Meye sudarshan hoy)

આ છોકરી દેખાવડું છે (Ā chōkarī dēkhāvaḍu chē)

I find the girl beautiful -> मैं लड़की सुंदर मिल

আমি মেয়ে সুন্দর এটি (aami meye sundor eTi)

હું છોકરી સુંદર શોધવા (Huṁ chōkarī sundara śōdhavā)

To me the girl is beautiful -> मेरे लिए लड़की सुंदर है

আমার মেয়ে সুন্দর (Aamaar meye sundor)

મને છોકરી સુંદર છે Manē chōkarī sundara chē

Mane chokari sundara lagechhe **is correct one**

Between Two Indian Languages

लड़की सुंदर है	সুন্দর মেয়ে
लड़की खूबसूरत है	সুন্দর মেয়ে
लड़की सुंदर लग रही है	সুন্দর মেয়ে দেখে মনে হচ্ছে
लड़की अच्छी लग रही है	ভাল খুঁজছেন মেয়ে
मेरे लिए लड़की सुंदर है	আমার জন্য সুন্দর মেয়ে

Example

Punjabi University accredited with NAAC 'A' Grade is one of the premier Institutions of Higher Education in the Northern region. The University primarily offers Post-Graduate and Under-Graduate Programmes in diverse areas of academics and research. The core research areas of the Department are Digital Image Processing, Multilingual Processing and Computer Networks. The Department is actively involved in research and software development since its establishment and has been recognized by UGC under SAP I and II and also by DST-**FIST** Level **I.**

After the successful organization of one week national workshop on **DIP** using Matlab during 2012, In 2013, we are organizing a One week National workshop on "Multilingual Technologies" from 11th November to 17th November, 2013 at Punjabi University, Patiala. The workshop will include invited talks from eminent personalities in the field of Multilingual Technologies. In addition, this workshop will also benefit the participants in API score as per UGC guidelines.

पंजाबी विश्वविद्यालय एनएएसी के साथ मान्यता प्राप्त 'ए' ग्रेड उत्तरी क्षेत्र में उच्च शिक्षा के प्रमुख संस्थानों में से एक है . विश्वविद्यालय के मुख्य रूप से शिक्षा और अनुसंधान के विभिन्न क्षेत्रों में स्नातकोत्तर और अंडर ग्रेजुएट कार्यक्रम प्रदान करता है . विभाग के मुख्य अनुसंधान क्षेत्रों डिजिटल इमेज प्रोसेसिंग , बहुभाषी प्रसंस्करण और कंप्यूटर नेटवर्क हैं . विभाग सक्रिय रूप से अपनी स्थापना के बाद से अनुसंधान और सॉफ्टवेयर विकास में शामिल है और एसएपी में और द्वितीय के तहत और भी डीएसटी - मुद्दी स्तर आई. द्वारा यूजीसी द्वारा मान्यता दी गई है

डुबकी 2012 के दौरान Matlab का उपयोग कर , 2013 में, हम पंजाबी विश्वविद्यालय , पटियाला में 17 नवम्बर - 11 नवम्बर से " बहुभाषी टेक्नोलॉजीज " , 2013 पर एक एक सप्ताह राष्ट्रीय कार्यशाला का आयोजन कर रहे हैं पर एक सप्ताह के राष्ट्रीय कार्यशाला के सफल आयोजन के बाद . कार्यशाला बहुभाषी टेक्नोलॉजी के क्षेत्र में प्रतिष्ठित व्यक्तियों से आमंत्रित वार्ता में शामिल होंगे . इसके अतिरिक्त, इस कार्यशाला भी यूजीसी के दिशा निर्देशों के अनुसार एपीआई स्कोर में भाग लेने वालों को फायदा होगा

Some Interesting Translations

The Fan is on	*Pankhaa Upar hai*
It is raining	*Yah barashtaa hai*
My house faces east	*Meraa ghar purab chehraa hai*
They have a big fight	*Unke paas baraa yudh hai*
The shopkeeper ran out of vegetables	*Dukaandaar sabji ke bahar bhaagaa*

Translations by different MT systems

Most commonly used English to Hindi MT systems:

MT1 : Google
http://translate.google.com

MT2 : MANTRA
http://mantra-rajbhasha.cdac.in/mantrarajbhasha

MT3: MaTra2
http://www.cdacmumbai.in/matra

MT4: Anuvadaksh
http://tdil-dc.in/

Online Translator Ex -1

The headquarters of planning committee are located in New Delhi.

- मुख्यालय का/की/के नियोजना समिति हैं स्थित में नई दिल्ली

(Angla Bharti)

- समिति योजना के मुख्यालय नई दिल्ली में स्थित हैं.

(Google)

- योजना बनाने वाले समिति का केंद्रस्थान नई दिल्ली स्थित हैं

(Matra Rajbhasa)

- प्लैनीग समिति का मुख्यालय न्यू दिल्ली में पता लगाया जाता है

(Matra2)

Online Translator Ex -2

Where did you hide the can opener?

- आपने डिब्बा ओपनर को कहाँ छिपाया

(Angla Bharti)

- तुम खोल कर सकते हैं कहाँ छिपा था?

(Google)

- जहाँ किया हुआ आप प्रारंभ करने वाला छुपाते हैं

(Matra Rajbhasa)

- आप कैन खोलनेवाला छिपाते हो

(Matra2)

Difficulties of Machine Translation

Problems of Machine Translation

- Word level difficulties
- Syntactic ambiguity
- Referential ambiguity
- Semantic ambiguity
- Metaphors and symbols

Word Level difficulties

- **Polysemy:** Same word may have different meaning.
  - *I am going to the bank.*
  - *This is of high interest.*
- **Synonymy:** Synonymous words may not be recognized.
  - *He has a car.*
  - *He has an automobile.*

Word Level difficulties (2)

- **Hyponyms:** Class/subclass identification may be a difficulty.
  - *He has a car.*
  - *He has a sedan.*
  - *He has a SX4*
  - *He has a Maruti*
- **Homograph:** Same word may be used as different part of speech.
  - *Drinking more water is good for health.*
  - *Please water the saplings carefully.*

Word Level difficulties (3)

- **Idiomatic expressions:** Idioms often do not have any correspondence with the constituent words.
  - *My mother gave me a piece of cake.*
  - *The test was a piece of cake for me.*

Syntactic Ambiguity

Structure of sentence does not clearly convey the sense.

- *Flying planes can be dangerous.*
- *I saw the man with a telescope.*
- *This my favourite chef 's dish.*
- *My father loves me more than my brother.*

Referential Ambiguity

- Pronouns refer to certain words but it is often not obvious to which noun it is referring to. References might even cross sentence boundaries
  - *The **computer** is printing data. **It** is fast.*
  - *The computer is printing **data**. **It** is numeric.*

Semantic Ambiguity

Sentences may have the same syntactic structure, but their meaning changes with constituent words.

- He took the rice *with* a curry
- He took the rice *with* a spoon
- He took the rice *with* difficulty
- He took the rice *with* a bad smell
- He took the rice *with* a sad face
- He took the rice *with* a friend.

Langauge Specific Features

- Metaphors
- Idioms
- Proverbs
- Symbols

Are often difficult to translate.

Selection of Right Word

What is the Right Word?

Target language may have many words
Corresponding to one source-language word:

Uncle -> चाचा मामा ताऊ मौसा
Bird -> पक्षी पंछी चिड़िया
Pumpkin -> कुम्हड़ा, कद्दू, काशीफल, कोहँड़ा, कोहड़ा, कुष्मांड, कुष्मांड, सीताफल, मीठा-कद्दू, पिंडफल, पिण्डफल, पुष्पफल, वृहत्फल, वेष्टक, आमक

Neela (Hindi) -> Blue, Indigo, Azure etc.
Ice -> 32 varieties in Eskimo language

Pattern Ambiguity

This is another difficulty observed with respect to English to Hindi MT [Chatterjee et. al. 2005]

This happens when the same verb is used in different senses.

E.g *Run* has 41 different senses. *Have* has 19 different senses.

They need to be translated differently:

English Sentence	Hindi Verb
The river ran into the sea.	*milnaa*
The army runs from one end to another.	*failnaa*
They run an N.G.O	*chalaanaa*
He runs for treasurer.	*khadaa honaa*
Wax runs in sun.	*galnaa*
We ran the ad three times	*prakaashit*

Noun Compounds

A compound noun is a noun that is made up of two or more words. Most compound nouns in English are formed by nouns modified by other nouns or adjectives or verb.

The joined words may form a single word also.

Examples:

tooth + paste = toothpaste. (N + N)

black + board = blackboard. (Adj + N)

baking + soda = baking soda

pressure cooker

prize distribution ceremony

} Separated Words

Noun Compound Translation

The primary difficulty for an MT system is to understand the semantics.

If semantics can be known and conveyed properly then translation becomes easier:

Often this boils down to grouping the words:

Sometimes it is easy from the adjective:

E.g.

Deep Blue Shirt -> ((Deep Blue) Shirt)

Round Copper Lid -> (Round (Copper Lid))

Noun Compound Translation

Sometimes very similar structure has different semantics:

Solid State Physics -> ((solid state) physics)

Solid Iron Pillar -> (solid (iron pillar))

OR

Public Service Commission ->
((Public Service) Commission)

Public Underground Transport ->
(Public (Underground Transport))

Noun Compound Translation

But sometimes it is confusing:

What about

- Brown Chocolate Muffin
- White Milk Powder

etc.

Identification of Noun Compounds and their Grouping is itself an important research work For many languages.

May be less for Sanskrit based languages:

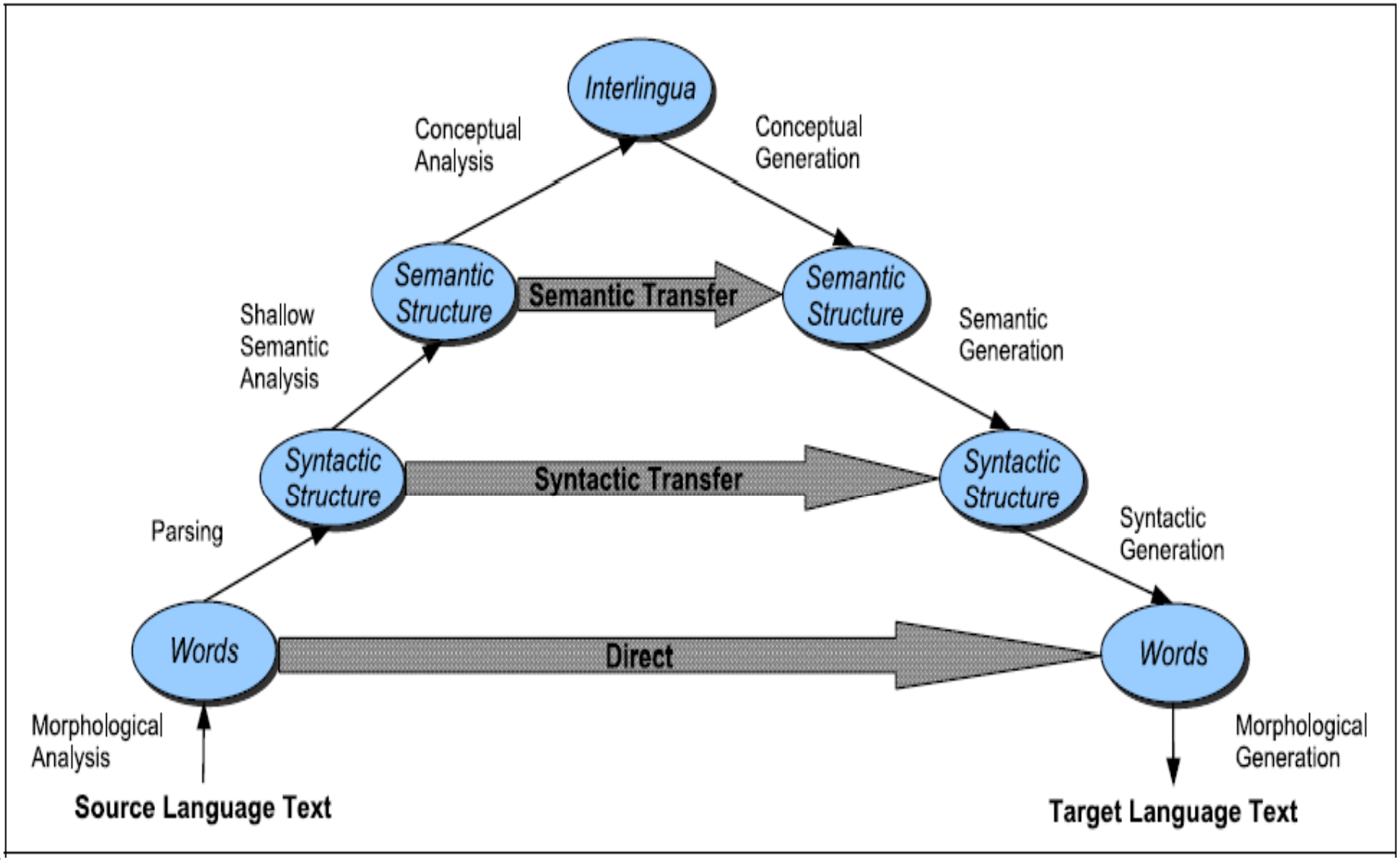
E.G: kerosene oil -> *mitti kaa tel*

A General Overview of MT Techniques

The VAUQUOIS Triangle

- The Vauquois triangle was used in the linguistic rule-based era of machine translation to describe the complexity/ sophistication of approaches to machine translation, and also the evolution of those approaches.

The VAUQUOIS Triangle

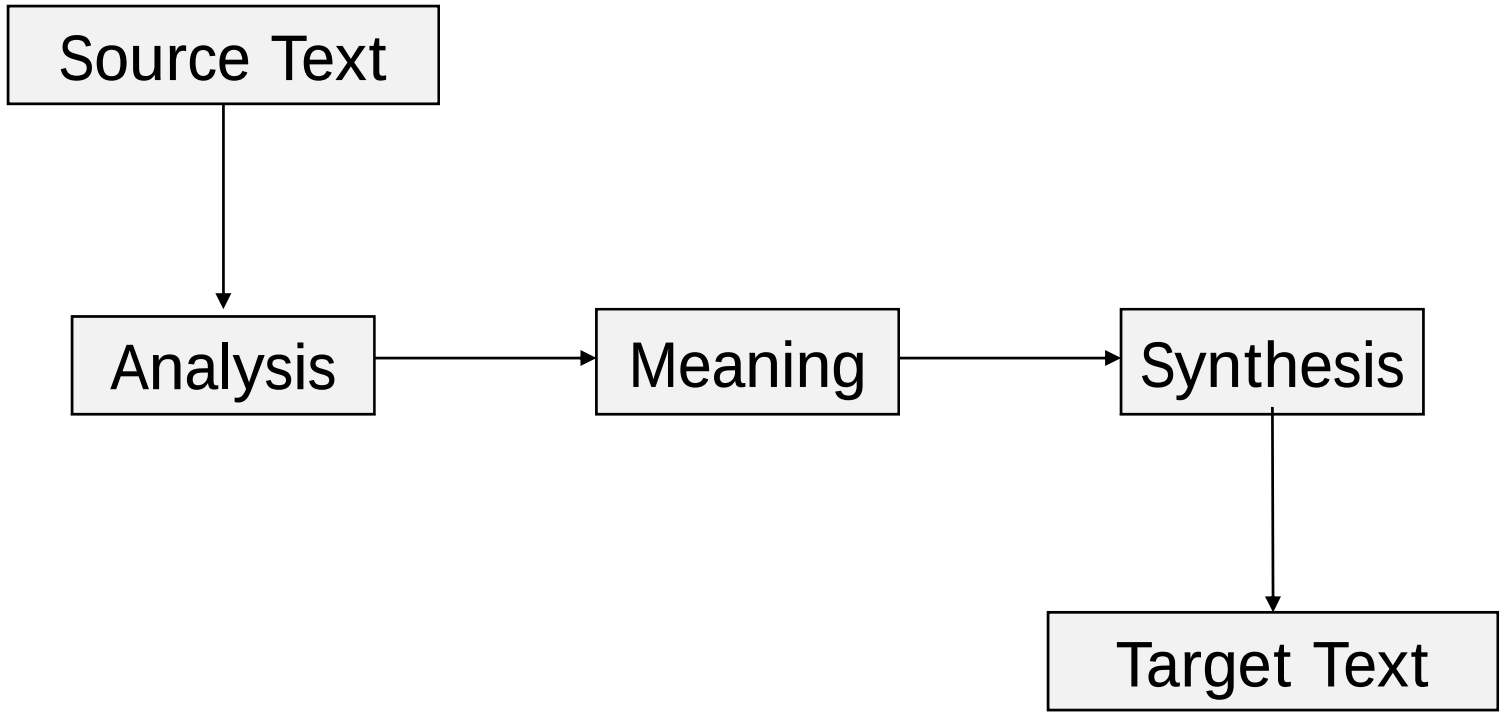


The VAUQUOIS Triangle

- The first approach used was a direct lexical conversion between languages.

- Later efforts moved up the pyramid and introduced more complex processing, and also a modularization of the process into steps:
  - analysis of the source language
  - transfer of information between the languages,
  - generation of target language output.

Translation Process



However we shall see that even without such explicit knowledge huge success can be achieved

Knowledge Used in Translation

One would expect use of different types of knowledge:

- Knowledge of the source language
- Knowledge of the target language
- Knowledge of correspondences between the source and target languages
- Knowledge of the subject matter and general knowledge used to understand what the text means
- Knowledge of the culture, social conventions, customs, expectations of speakers of the source and target languages

The VAUQUOIS Triangle

- Each step up the triangle required greater effort in
  - source language analysis
  - target language generation
  - reduction of effort involved in conversion

- The ideal of the field is a complete analysis of each sentence into an "interlingua" - a schema capable of representing all meaning expressable in any language in language-independent form.

Different Paradigms

The Major Paradigms

Example- Based	(EBMT)	– Nagao, Somers
Knowledge-Based	(KBMT)	– Carbonell, Nirenburg
Lexical-Based	(LBMT)	-- Dorr, Tsujii & Fujita
Neural-Net Based	(NBMT)	-- McLean
Rule-Based	(RBMT)	– Kaplan, Okumura
Statistics Based	(SMT)	-- Brown, Koehn

Different Paradigms

Lexicon-based MT—Based on relating the lexicon entries of one language to the lexicon entries of the other language e.g. Anusaarka (IIT-K, IIIT Hyderabad) late1990s.

Knowledge-based MT– concentrates on development of knowledge intensive morphological, syntactic and semantic information for the lexicon e.g. Pangloss [CMU,1980], GAZELLE [USC, 1990].

EBMT Proposed as an MT technique by Nagao in 1984. Based on the idea of performing translation by imitating examples of translations of sentences of similar structure.

Different Paradigms

- Rule-based MT– relies on different linguistic levels of rules for translation between two languages.

- Statistical MT--based on n-gram modeling, and probability distribution of the occurrence of a source-target language pair in a very large corpus. e.g.

  IBM model, Matador (Univ. of Maryland)

  - Started in the ‘90s,
  - Became more popular after 2000
  - Modeling Translation Task as optimization

Example Based Machine Translation

Example Based Machine Translation

- Proposed as an MT technique by Nagao in 1984.
- Based on the idea of performing translation by imitating examples of translations of sentences of similar structure.
- A large number of translation examples between the source language (SL) and target language (TL) are stored in a system's knowledge base.
- These examples are subsequently used as guidance for future translation tasks.
- In order to translate a new input sentence in SL, one (or more) SL sentence (s) are retrieved from the example base.
- Their translations in TL are also retrieved from a parallel database.
- This example is adapted suitably to generate a translation of the given input.

Issues Related to EBMT

- There are many issues pertaining to EBMT.
- We focus on:
  - Retrieval
  - Adaptation
- The more *similar* is the retrieved sentence to the input one, the easier will be its *adaptation.*
- However, there is no straightforward way to measure *similarity* between sentences.
- Semantic similarity does not always work

Examples

She is good looking

She looks good

She is good to look at

BUT

This horse runs good

This horse is good to run on

It was a good run by this horse

So word based Similarity is not the key to the Success of EBMT!!

Syntactic Similarity

- Syntactic similarity refers to the structure of the sentence with the help different constituent words of a sentence.

- Although two sentences may not be similar apparently, they are often useful from adaptation point of view.

  John is playing football.

  Mary is eating bread.

- We have designed a systematic adaptation scheme on the basis of some *elementary operations* involving words and suffixes.

Adaptation Procedure using Word and Suffix Operations

Word Replacement:

Input sentence : Ram is eating rice

Database sentence : Ram is eating bread

*~ram **rotii** khaa rahaa hai*

bread (*rotii*) *replace with* rice (*chawal*)

Word Deletion and Addition

Input sentence: Sita is eating rice.

Database sentence:

Sita **and Gita** are eating rice.

~ *sita* ***aur gita*** *chawal khaa rahii hai*

Word Addition (WA): Addition of a new word to the retrieved translation example works in the opposite way.

Advantages of Suffix operations

- Number of suffixes is fixed.
- Fixed costs (K) for all the suffix operations.
- No dictionary search required for suffix operation.

Suffix Addition

A suffix is added to some word in the retrieved example. The word here is in its root form.

- **Singular – plural of nouns**

  girl (*ladki*) → girls (*ladki**yaan**)*

- **Morphological changes in verb**

  He is reading (*wah pad rahaa hai*)

  →He reads (*wah pad**taa** hai*)

Suffix Replacement

- Singular – plural of nouns

  Boy (*ladkaa*) → Boys (*ladke*)

- Change of Adjectives

Bad boy (*buraa ladkaa*) → Bad girl (*burii ladkii*)

- Morphological changes in verb

  He reads (*wah padtaa hai*)

  → She reads (*wah padtii hai*)

Suffix Deletion

Morphological changes in verb

He reads (*wah pad**taa** hai*)

→She is reading (*wah pad rahii hai*)

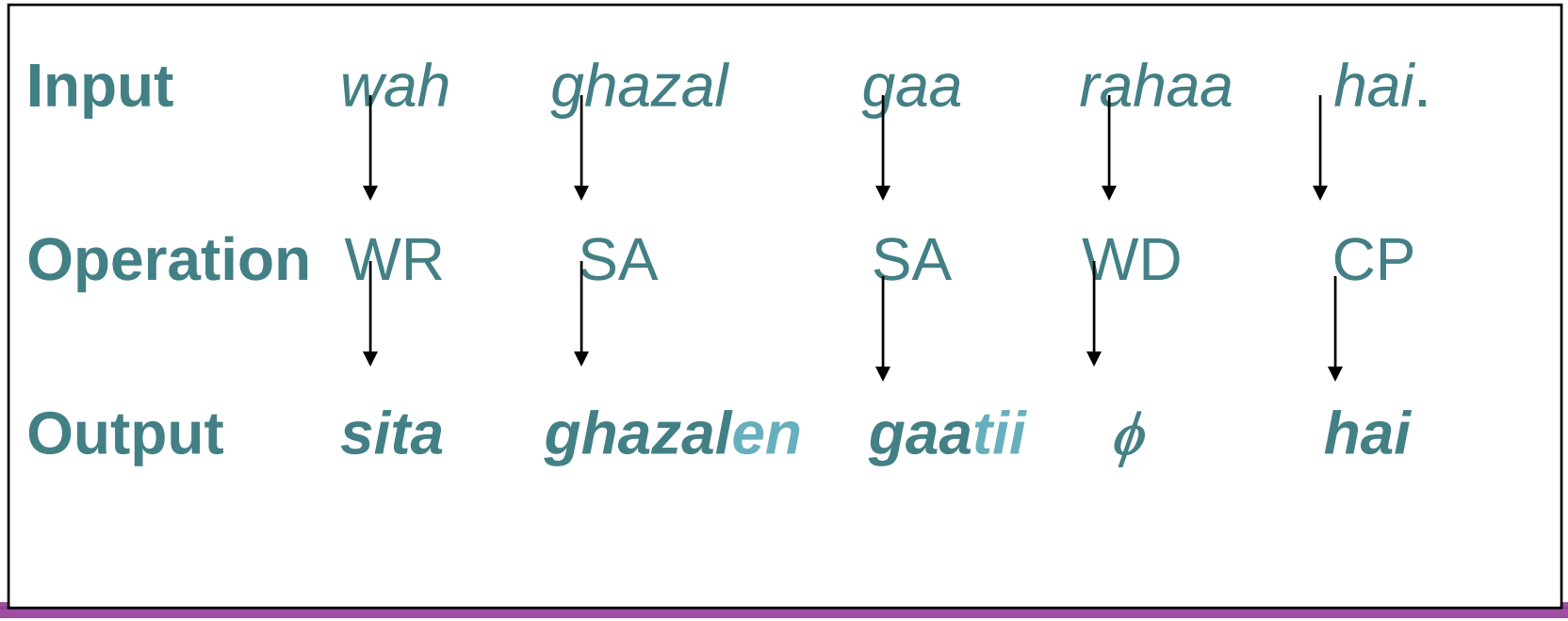
Example of Adaptation

Input sentence: Sita sings ghazals

Retrieved example

He is singing ghazal ~ *wah ghazal gaa rahaa hai*

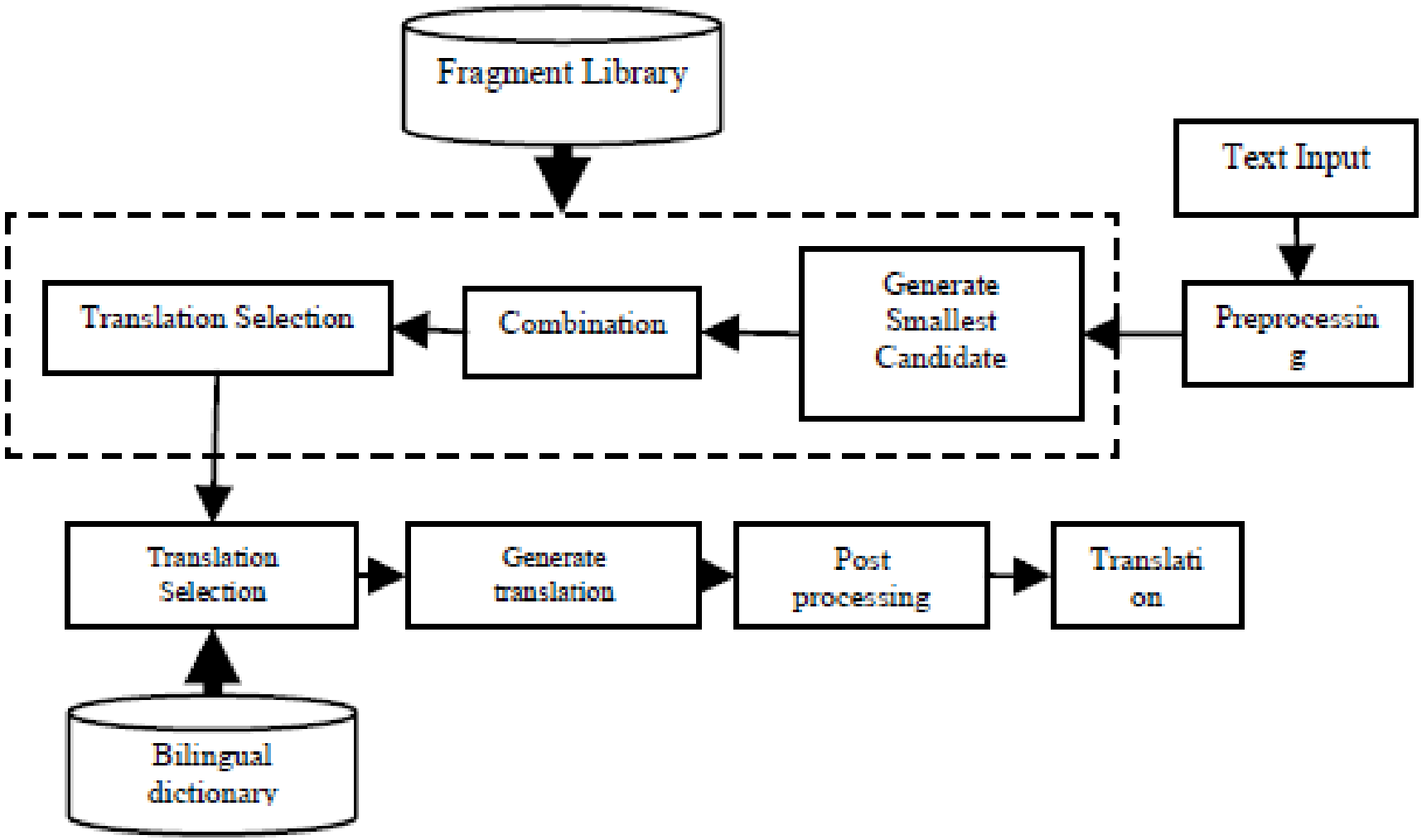
.



EBMT – What it needs

- From a corpus of source-target sentence pairs, EBMT models of translation perform three distinct phases in order to transform a new input string into a target language translation:
  - **Matching Phase:** Searching the source side of the parallel corpus for ‘close’ matches and their translations.
  - **Alignment Phase:** Determining the sub sentential translation links in those retrieved examples.
  - **Recombination Phase:** Recombining relevant parts of the target translation links to derive the translation

Architecture of an EBMT System



Where Does Adaptation fail?

Translation Divergence

What is Divergence?

Divergence occurs "*when structurally similar sentences of the source language do not translate into sentences that are similar in structures in the target language.*" [Dorr, 1993].

Can often be found in translations between languages of same origin,
(e.g. English- German, English-Spanish, Bengali - Hindi)

We shall illustrate with examples from English-Hindi

Structural Divergence

Verbal Object: : Noun Phrase (NP) in SL

→ Prepositional Phrase (PP) in TL

John will read this book

→ *John yah kitaab padhegaa*

Vs.

John will attend this meeting

→ *John iss sabhaa mein jaayegaa*

Categorial Divergence

Predicative Adjunct → Verb

She is in trouble →

wah musiibat mein hai.

BUT

She is in tears → *wah ro rahii hai*

Demotional Divergence

Main verb → the subjective complement upon translation

It suffices ~ *yah paryaapt hai*

These two sofas face each other

~ *yeh do sofa ek dusre ke saamne hain*

The soup lacks salt

~ *soup mein namak kam hai*

Pronominal Divergence

Focus is on sentences with "it" as the subject.

It is running

~ *wah bhaag rahaa hai*

(यह चल रहा है (google))

BUT

It is raining

~ *barsaat ho rahii hai*

(यह बारिश हो रही है (Google))

Issues between Hindi & Bengali

The girl **in white dress** sings well:

safed kapdi**walli** ladkii achhi gaatii hai

shada kapor **pora** mayeta bhalo gaay.

Consider translation of this sentence:

The boy who comes here everyday riding a blue bicycle sings well.

Issues between Hindi & Bengali

Reduplication.

I don't like taking fish daily ->

ami **roj roj** machh khete bhalobasi naa.

mujhe **har roz** machhli khaana pasand nahin hai.

He comes here secretly ->

se **chupi chupi** ekhane ase.

who **chhup chhup ke** idhar aataa hai

Statistical Machine Translation

Prologue

- Gained tremendous momentum in recent years

- Generally languages are so rich, complex, different that it is difficult to distil knowledge to frame exhaustive set of rules or examples, which can be encoded into program

- Can then the rules be discovered automatically?

  (perhaps from a pair of corpus, and analyzing the data statistically)

This begins a new line of research and gives rise to

Statistical Machine Translation

Observation

- The rules cannot be very *deterministic*.
- An *element of probability* is associated with translation.
- Perhaps one can estimate the probabilities from a huge set of *parallel corpus.*
- One needs to consider the *associations between different pairs of words*.
- Hence arises the question of modeling – *statistical modeling.*

Statistical Modelling

Statistical MT

A true translation which is both Faithful and Fluent is often impossible.

A translation is said to be faithful if it conveys the full sense of the source sentence.

E.g. *wo ladkaa kal sham ko yahaan aayaa thaa* >>
The boy came here yesterday (NOT Faithful)

A translation is said to be fluent if its construction correctly follows the grammar of the target language.

E.g. *wo ladkaa kal sham ko yahaan aayaa thaa* >>
The boy came yesterday evening here
(NOT Fluent)

Statistical MT

A compromise is often tried for.

We want a model that maximizes a value Function.

SMT is about building a probabilistic model to combine faithfulness and fluency:

$$\text{Best translation } \hat{T} = \underset{T,\,S}{\text{argmax}} \; \text{faithful}(T,S) * \text{fluency}(T)$$

Consider that a source language sentence S may translate into any target language sentence T.

Some translations are just more likely than others.

How do we formalize "more likely"?

Statistical MT

P(**s**) -- a priori probability. The chance that **s** happens.

For example, If **s** = "May I know your name"

Then P(**s**) is the chance that a certain person at a certain time will say "May I know your name" as opposed to saying something else.

P(**t** | **s**) -- conditional probability. The chance of **t** given **s**.

For example, Let **s** = May I know your name

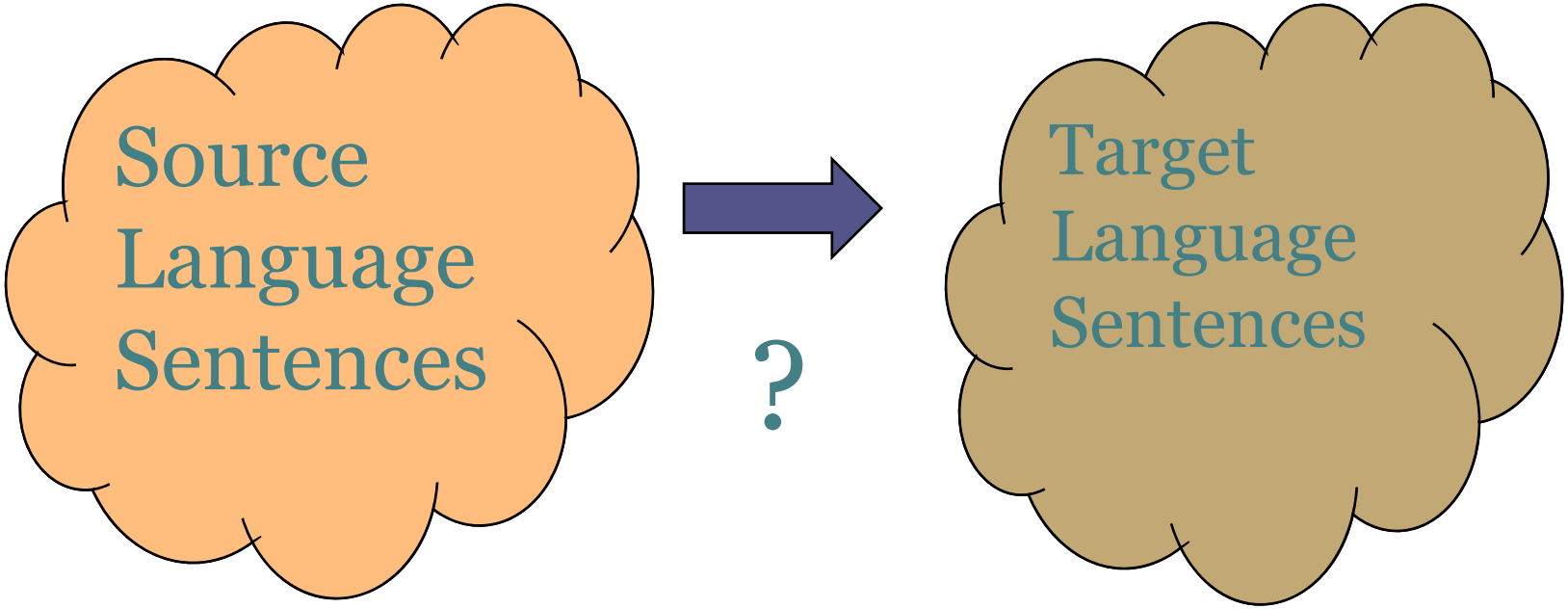
and

t = *kyaa main aapkaa naam puchh saktaa hoon*

then P(**t** | **s**) is the chance that upon seeing **s**, a translator will produce **t**.

Statistical MT

Of course the problem is how to estimate the probabilities:



Both are of infinite size.
Hence individual probabilities are zero.
Hence estimation is done in an indirect way.

But we look at the modeling aspect first.

Statistical MT

P(**s**, **t**) -- joint probability. The chance of **s** and **t** both happening. If **s** and **t** don't influence each other, then we can write P(**s**, **t**) = P(**s**) * P(**t**).

If **s** and **t** do influence each other, then we had better write P(**s**, **t**) = P(**s**) * P(**t** | **s**) (using Bayes thm).

That means: the chance that "**s** happens" times the chance that "if **s** happens, then **t** happens." If **s** and **t** are strings that are mutual translations, then there's definitely some influence.

This is known as Noisy-Channel Model, where three modelings are involved:

- Source model to compute P(s)
- Channel model to compute P(**t** | **s**)
- Decoder to produce **t** given **s**

The question is how do we go about this?

Language Modeling

Language Model

Tells how likely it is that a sequence of words will be uttered/ written in the language.

Helps in : fluency, word order etc., which vary across languages

E.g. <Noun> <Adj>$_{It}$ >> <Adj> <Noun>$_{En}$

Formally, LM is a function that gives the probability of a given sentence.

E.g. P(What is your name?) > P (What your name is?)

The obvious difficulty is:

- there are infinitely many sentences in
- any language.

So how to obtain??

Language Model

We try to compute probabilities from language corpus.

Computing probabilities of sentences are meaningless
Hence n-gram modeling is used.

For different values of n (e.g. 1, 2, 3, ..) the probability of the sequence of n words $w_1\ w_2\ ..\ w_n$.

Using Bayes' Theorem:

$$P(w_1\ w_2\ ..\ w_n) = P(w_1)\ P(w_2 \mid w_1)\ P(w_3 \mid w_1\ w_2)\ .....\ P(w_n \mid w_1\ w_2\ ..\ w_{n-1})$$

The size of n depends on language, corpus etc.

Estimation

The question is how to estimate the probabilities

1. Unigram (w) = $$\frac{count\ (w)}{Total\ \ no.of\ \ Words}$$

2. Bigram (w_1, w_2) = $$\frac{count\ \ (w_1, w_2)}{\sum_w count\ \ (w_1, w)}$$

3. Trigram (w_1, w_2, w_3) = $$\frac{count(w_1, w_2, w_3)}{\sum_w count(w_1, w_2, w)}$$

The bigger the corpus, the better the estimate!

N-grams

N-gram probabilities helps in translation:

e.g. *aami bhaat khai* $_{(B)}$ >> I eat rice
aami jal khai $_{(B)}$ >> I drink water
aami churut khai $_{(B)}$ >> I smoke cigar
aami osudh khai $_{(B)}$ >> I take medicine

N-grams allow us to choose the right translation.

N-grams

Very similar things can happen with other Language pairs also:

The boy >> Le garçon $_{FR}$
>> Il ragazzo $_{IT}$

The boys >> Les garçons $_{FR}$
>> I ragazzi $_{IT}$

The girl >> La fille $_{FR}$
>> La ragazza $_{IT}$

The girls >> Les filles $_{FR}$
>> Le ragazze $_{IT}$

WORD BASED MODELS

Introduction

Word models come from the original work at IBM.

The MT technologies have advanced since then.

These works helps us to understand the foundations of SMT and its techniques.

IBM has proposed five models - with gradually Improving versions.

Ref: The Mathematics of Statistical Machine Translation: Parameter Estimation - Peter F Brown et.al - Computational Linguistics, Vol 19, No. 2, 1993

MT by Translating Words

In the simplest form: It is *Lexical Translation*

A string can be translated by translating each word of the *source text* to the *target text*.

However, there is a difficulty:

A source language word may have more than one translation in the target language:

Haus (G) → House, Home, Household, Building (Eng)
→ *ghar, bhavan, mahal, prasad* … (Hindi)

How to choose the best one?

MT by Translating Words

How about computing the statistics?

After scanning a large number of documents we can estimate probability of each of the translations!!

How does it solve our purpose?

We can use the probabilities of the individual words of a foreign language text **f** to determine the most probable translation in the language **e**.

MT by Translating Words

A foreign language sentence may have words:

f1 f2 …fn

Each has its own choice of alternatives, and corresponding translation probabilities :

t(***e*** | ***f***) - Prob. that word ***f*** translates into word **e** where ***e*** is a word in the target language

These t ′s are called Translation Probabilities

MT by Translating Words

For example consider the following tables of translation probabilities (hypothetical):

yah	
this	0.5
the	0.3
that	0.1
_	0.1

makaan	
house	0.4
bungalow	0.3
residence	0.15
flat	0.15

sundar	
beautiful	0.45
nice	0.3
pretty	0.15
cute	0.1

hai	
is	0.75
exists	0.15
remains	0.1

What is the most likely translation of :

yah makaan sundar hai?

Word Alignment

A word-by-word translation gives us :

this house beautiful is

Thus implicitly we are using a mapping from the foreign words to the English words.

Depending on the grammar we can have different mappings. In this case we have:

$1 \rightarrow 1$, $2 \rightarrow 2$, $3 \rightarrow 4$, and $4 \rightarrow 3$

The correct one is : this house is beautiful

Word Alignment

Other than a permutation, word alignment may suffer from Alignment pattern:

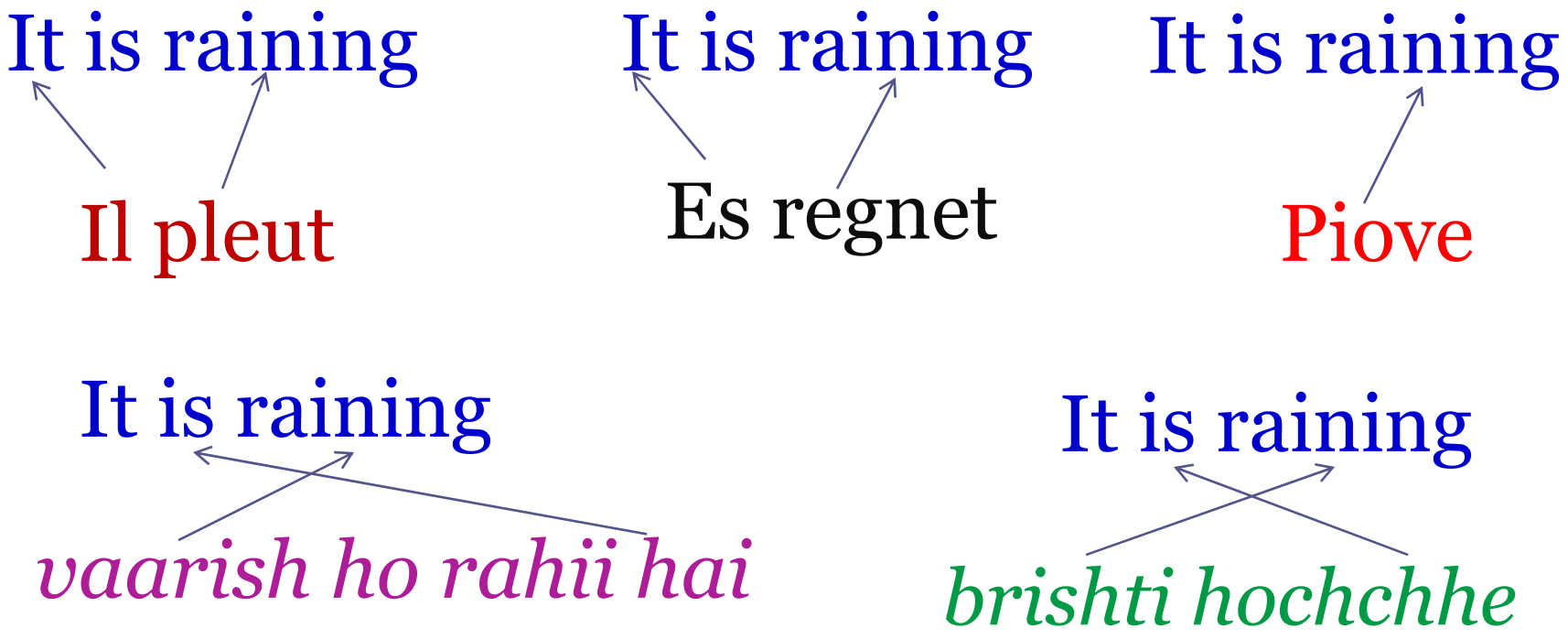
1 – 0 : yah makkan bahoot **hi** chhota hai→
the house is very small

3 - 2: yah makkan **sab se chhota** hai→
this house is **the smallest**

Etc.

Word Alignment

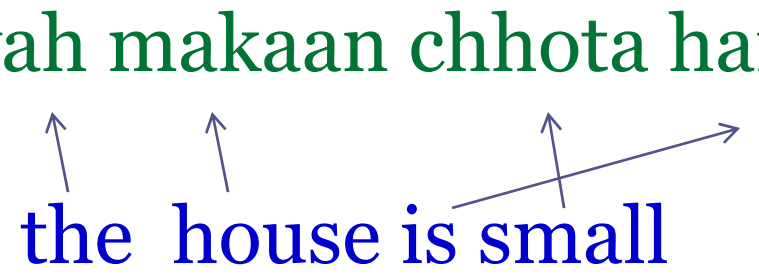
And it varies with language pairs:



Word Alignment function

- An alignment is best represented using an alignment function.
- It maps for each word of the Target Language to a word of the Source language

E.G



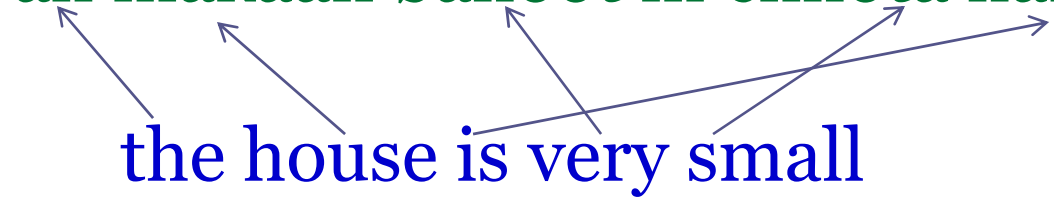
$$a: \{1 \rightarrow 1,\ 2 \rightarrow 2,\ 3 \rightarrow 4,\ 4 \rightarrow 3\}$$

Note: Alignment function is from target to source.

Word Alignment function

E.G 2.

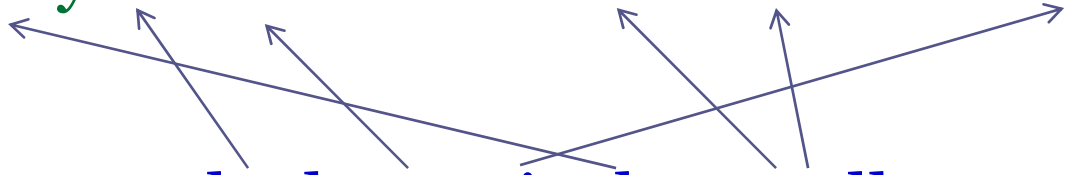
yah makaan bahoot hi chhota hai



the house is very small

a: { 1 → 1, 2 → 2, 3 → 6, 4 → 3, 5 → 5}

E.G 3. φ yah makaan sabse chhota hai



the house is the smallest

a: { 1 → 1, 2 → 2, 3 → 5, 4 → 0, 5 → {3,4}}

IBM Models

- IBM Model 1 : Based on Lexical Model
- IBM Model 2 : Adds Alignment Model
- IBM Model 3 : Adds Fertility Model - How many output words an input word produces.
- IBM Model 4 : Adds *Relative Alignment* Model Positions of any other *e* words that are connected to the same *f* word
- IBM Model 5 : Takes care of *deficiency*

Problem of models 3 and 4 is that they allow multiple output words to be placed at the same position. Model 5 keeps track of vacant positions, and allows new words to be inserted only in these positions.

Epilogue

More recently other models are coming up:

-Phrase Based models

- Treelet Based models

-FactorBased Models

Still in Indian scenario the effort is less.

I expect more research will come up in near future With Indian languages.

Thank You