Statistical Machine Translation

LECTURE – 1

INTRODUCTION

April 12, 2010

Brief Outline

- General Introduction
- Machine Translation vis-à-vis NLP
- Role of Knowledge in MT
- History of MT
- Difficulties of MT
- Ambiguities
- Study of Divergence
- Motivation to SMT

What is Machine Translation

- Machine Translation (MT) pertains to automated translations of text 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
- It is a subfield of - Natural Language Processing/
  - Computational Linguistics

Natural Language Processing

- We expect computers to perform useful and interesting tasks involving human languages.

- To gain insights regarding human languages and human processing of language through computational work.

- Has become more meaningful in the era of internet where

  - billions of documents are available for one to use

  - more and more novel applications are being considered

Natural Language Processing

NLP tasks can be classified into three categories:

* Developing Basic Linguistic Tools:

E.g. Parser, Word-net, On-line Dictionary

* Fundamental Applications:

- Word Sense Disambiguator
- Text Summarizer
- Machine Translation System

* Innovative Applications

- On-line shopping, Sentiment Analysis

Natural language: A Means of Communication

- Language is one of the major means of communication for human beings.
- Each medium of communication has its own advantages and disadvantages.
- With respect to languages it is observed that individuals often do not know how they use languages to understand the content.

Complexity of Connected Text

Neil was returning from school dejected today was the mathematics test.

Complexity of Connected Text

Neil was returning from school dejected today was the mathematics test.

He couldn’t control the class – the boys were very noisy.

Complexity of Connected Text

Neil was returning from school dejected
today was the mathematics test.

He couldn't control the class –
the boys were very noisy.

The teacher shouldn't have made him responsible.

Complexity of Connected Text

Neil was returning from school dejected

today was the mathematics test.

He couldn't control the class –

the boys were very noisy.

The teacher shouldn't have made him responsible.

After all he is just a janitor !!!

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 course we shall focus on Machine Translation.

We start with some Example:

Example

- **Articolo 1**

  E' indetto un concorso, per titoli ed eventuale colloquio per l'attribuzione di n. 1 borsa di studio di **6 mesi**, dell'importo complessivo di **12.000 Euro** per lo svolgimento presso il Dipartimento di Informatica di una ricerca dal titolo **"Techniques of statistical machine translation based on syntax analysis".**

Translation by On-line Translator

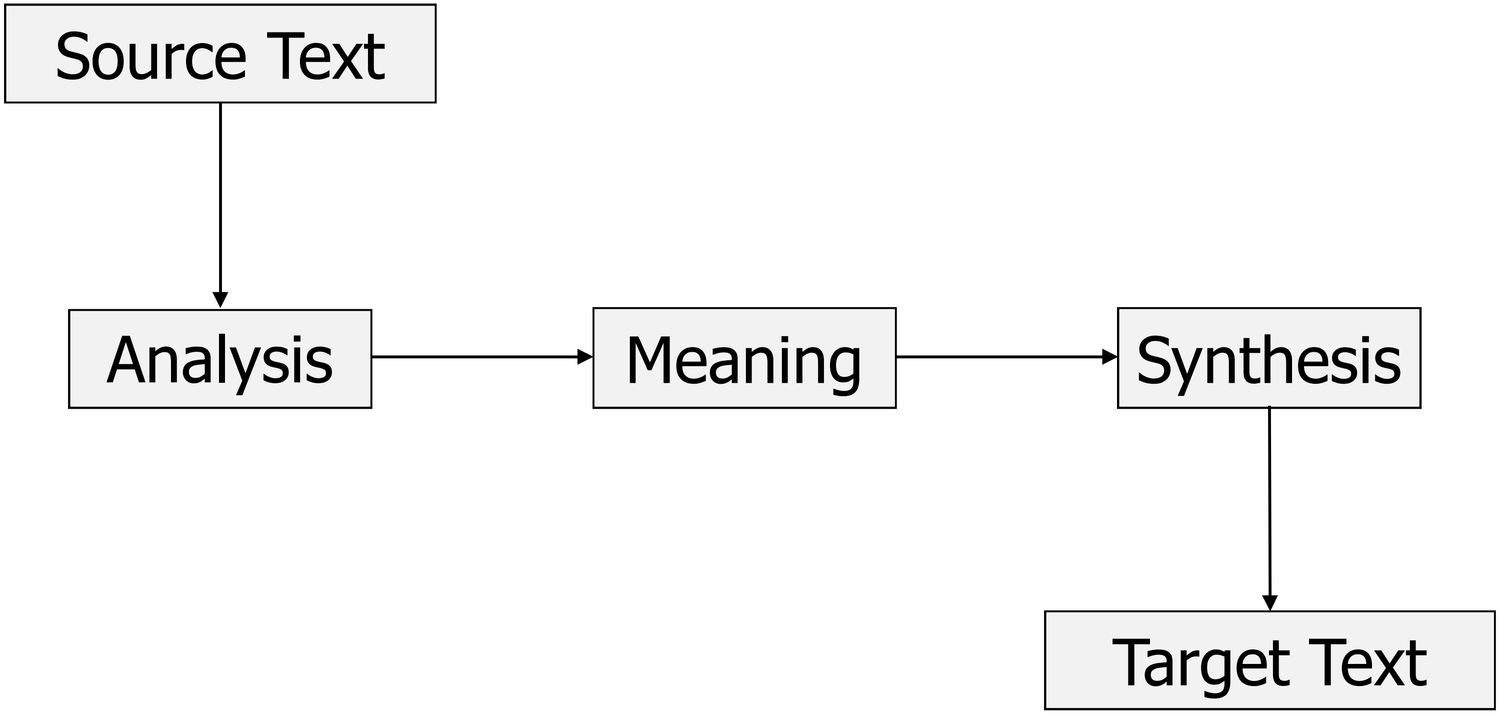
- Article 1 is' held a competition based on qualifications and possible interview for the allocation of n. 1 scholarship of 6 months' total amount of 12,000 Euros for the development at the Department of Computer Science of a study entitled "Techniques of statistical machine translation based on syntax analysis".

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

Translation Process



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

Let us look at the History of MT first.

Historical Overview

- Four periods:
  - Optimistic beginnings
  - Disillusion
  - 70ies: partial successes
  - Commercial application
- Generations of translation techniques

History- Optimistic Beginnings

- 1942 first computer -> condition for development of machine translation created
- Very optimistic attitude towards the problems, high expectations
- Attempts for developing resources (e.g. Bi-lingual dictionaries) to support MT
- 1952 first MT conference
- 1954 Georgetown Experiment (Russian to English) -> enormous success – suggesting that the problem was almost solved;   research intensified

George Town Conference - 1954

Quotable quotes:

Mechanical translation was not only feasible, but far
Closer to realizations than possibly the audience recognized.

In about 2 years (from August 1957) we shall have a device
Which will at one glance read a whole page and feed what it
has read into a tape recorder. And thus remove all human
Cooperation on the input side of translation.

History-Disillusion

- Expectations not fulfilled – high expectations without the theoretical background.
- 1966 very negative report from ALPAC (*Automatic Language Processing Advisory Committee*) citing poor-quality technology and availability of cheap manual labour.
- Post-editing is at par with human translation – both cost and efficiency wise.
- Danger of over-promising abilities was visible.
- Interest in MT decreased - Much less research funds

History-70s

- Development of Artifical intillegence and knowledge based techniques.
- Revival of MT research
  - EUROTRA by the EC to provide MT of all the members nations‘ languages.
  - Initiative by Japan government and industries.

No looking back!!!!

Different 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
Context based	(CBMT)	--	Carbonell

Are the major ones.

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].

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

Different Paradigms

- 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.
- 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, along with its translation in TL.
- This example is adapted suitably to generate a translation of the given input.

Different Paradigms

Context Based :

- recently proposed
- has not been explored in detail
- use statistical techniques, but in a different way

Characteristics:

- a lightweight translation model
- utilizing a full-form bilingual dictionary
- a sophisticated decoder using long-range context
  via long n-grams and cascaded overlapping.
- in-language substitution of tokens and phrases
- substitution utilizes a synonym and near-synonym generator
- corpus-based unsupervised learning process.

First Commercial Systems

Meteo (Montreal – 1966) weather forecast.

Systran (1968) – Russian English, (Defence)
(US Airforce) French- English

Interlingua-Based Systems

Interlingua - formal representation of semantics
(1980-90) independent of specific language

Pangloss - (Southeren California)

Catalyst (CMU)

Considered better than approaches which use low-level mapping of lexical/ syntactical units - as proper theory of meaning is aimed to formalize.

Existing Systems

Google language Tools translates among 50 language pairs

If we search in Google we can find at least 40 commercial systems.

In this department we have a working system.
http://semawiki.di.unipi.it/translate/

But still lot of improvements need to be done.

- development for resource poor countries.
- improvements for existing ones.

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 Lancia*
  - *He has a Flavia*
- **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.*

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.

- I took rice *with* fish.
- I took rice *with* a spoon
- I took rice *with* a friend.

Complex Semantic Ambiguity

- **Homonymy:** to understand the sentence specific sense has to be used.
  - ○ *The box is in the pen.*

- **Metonymy:** substituting the name of an attribute or feature for the name of the thing itself
  - ○ *They counted heads.*
  - ○ *While driving John hit the tree.*

Langauge Specific Features

- Metaphors
- Idioms
- Proverbs
- Symbols

Are often difficult to translate.

Some Illustrations with Italian

He came by car.

Egli è venuto in auto.

He came by three o'clock.

Egli è venuto da tre

He came by London.

Egli è passando da Londra.

He came by himself.

Venuto da solo.

He came by night.

È venuto di notte.

He came by village.

Egli è venuto per villaggio.

Some Illustrations with Italian

I canned fish.	I pesci in scatola.
I canned apple	I mela in scatola.

But

I can fish	Posso pesce
I can run	I possibile eseguire

Sometimes perhaps it does not matter!!

The computer is printing data. It is numeric.

Il computer è la stampa di dati. E 'numerica.

Le Computer sont des données d'impression. Il est numérique.

The computer is printing data. It is fast.

Il copmputer è la stampa di dati. È veloce .

Le Copmputer sont des données d'impression. Il est rapide .

Sometimes perhaps it does not matter!!

I eat rice with spoon

Je mange du riz avec une cuillère

Mangio riso con un cucchiaio

I eat rice with friends

Je mange du riz avec des amis

Mangio riso con gli amici

I eat rice with fish

Je mange du riz avec du poisson

Mangio riso con pesce

But Sometimes it does

I eat rice with spoon

main chammach ***se*** *chawal khaataa hoon (H)*

aami chamaoch ***diye*** *bhaat khaai (B)*

I eat rice with friends

main dost ***ke saath*** *chawal khaataahoon*

aami bondhu-r ***saathe*** *bhaat khai*

I eat rice with fish

main machhli ***ke saath*** *chawal khaataahoon*

aami maachh ***diye*** *bhaat khaai*

Selection of Right Word

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

E.G

Uncle -> mama, kaka, chacha, jethu,
pise, meso (Bengali)

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*

Other Difficulties

Domain dependency: bat (in a game; in animal)

Type of text: News Article vs. stories.

Recursive nature:

- This is the house that Jack built.
- This is the malt that lay in the house that Jack built
- This is the rat that ate the malt
  That lay in the house that Jack built.
- This is the cat that killed the rat
  That ate the malt that lay in the house that Jack built.
- This is the dog that worried the cat
  That killed the rat that ate the malt
  That lay in the house that Jack built.

Translation 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*
this book will read

Vs.

John will attend this meeting
→ *John iss sabhaa mein jaayegaa*
this meeting to will go

Structural Divergence

Verbal Object: : Noun Phrase (NP) in SL

→ Prepositional Phrase (PP) in TL

John will read this book

→ john leggerà questo libro

Vs.

John will attend this meeting

→ john sarà partecipare a questa riunione

Conflational Divergence

- The verb of a source language sentence needs incorporation of additional words in the target language.

To love	-	*pyaar karnaa*
To slap	-	*thaappad maarnaa*
To borrow	-	*udhaar lenaa*
	Vs.	
To kick	-	*payr se maarnaa*
To stab	-	*chaaku se maarnaa*
To hurry	-	*jaldii se jaanaa*

Categorial Divergence

Predicative Adjunct → Verb

She is in trouble →

wah musiibat mein hai.

she trouble in is

BUT

She is in tears → *wah ro rahii hai*

she cry ...ing is

Categorial Divergence

Predicative Adjunct → Verb

She is in trouble → Lei è nei guai

BUT

She is in tears → Lei è in lacrime

Thematic Divergence

- Object → Subject upon translation
- Subject → Modifier upon translation

The shopkeeper ran out of vegetables →

dukaandaar	*ke paas*	*sabjiyaan*	*samaapt*	*ho gayii thii*
shopkeeper	*to*	*vegetables*	*finished*	*has been*

John misses Mary →

Mary manque à John (F)

Demotional Divergence

Main verb → the subjective complement upon translation

These two sofas face each other

~ *yeh do sofa ek dusre ke saamne hain*

these two sofa one other in the front is

The soup lacks salt

~ *soup mein namak kam hai*

soup in salt less is

Pronominal Divergence

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

It is running

~ *wah bhaag rahaa hai*

it run ...ing is

BUT

It is raining

~ *barsaat ho rahii hai*

rain happen ..ing is

Pronominal Divergence

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

It is running

~ E 'in esecuzione

BUT

It is raining

~ piove

Possessional Divergence

Focus is on sentences with “have/has" as the main verb.

He has a book ~ *uske paas ek kitaab hai.*
with him a book is

He has a headache ~ *use sirdard hai*
upon him headache is

These birds have sweet voice.
~ *ye chidiyon kii vaanii miithii hai*
these birds of voice sweet is

This city has a museum.
~ *iss shahar mein ek sangrahaalay hai*
this town in one museum is

Lexical Divergence

Here new lexical elements need to be added for conveying the sense.

The sky is cloudy

~ *aakaash par baadal chhaaye huye hai*

sky on cloud spread over is

They ran into the room

~ *weh daurte huye kamre mein ghus gaye*

they running room in entered

Some Examples from European languages

Thematic:

John misses Mary → Mary manque à John (FR)

Promotional

Il est probable que Jean viendra (FR) →

Jean will probably come

Demotional

Er liest gern (DE) → He likes reading

Some Examples from European languages

Structural

He aims the gun at him → Er zielt auf ihn mit dem gewehr (DE)

Categorial

John is fond of music → John aime la musique(FR)

Lexical

Give a cry → Pousser un cri (FR)

Identification of Divergence

Interlingua Approach [Dorr, 1993] SYSTRAN, UNL

Transfer Approach [Han et. al., 2000; Watanabe, 2000] --Transfer rules

Generation-Heavy Machine Translation Approach [Habash, 2003] -- Statistical approach

Rule-based Approach[Gupta & Chatterjee, 2003]
FT & SPAC based

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, 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 SMT.

References

- Dorr B. J. (1993). *Machine Translation: A View from the Lexicon.* MIT Press, Cambridge, MA.

- Margaret King (1987). Machine Translation Tutorial .

- Dorr, B. J., Jordan P. W., and Benoit, J. W. 1999. A survey of current paradigms in machine translation. In *Advances in Computers, M. Zelkowitz, Ed. Vol. 49. Academic Press, 1–68.*

- N. Chatterjee, Shailly Goyal and Anjali Naithani: Pattern Ambiguity and its Resolution in English to Hindi Translation, in the proceedings of International Conference "Recent Advances in Natural Language Processing-2005", ISBN: 954-91743-3-6, Borovets, Bulgaria, 2005, pp 152 – 156.

- D. Gupta and N. Chatterjee. Identification of Divergence for English to Hindi EBMT. Proc. MT Summit IX, New Orleans, LA, 2003, pp 141 – 148.

Thank You!!