CS565: Intelligent Systems and Interfaces

Lecture 1

13th Jan, 2016

Semester: Jan – May 2016

Ashish Anand

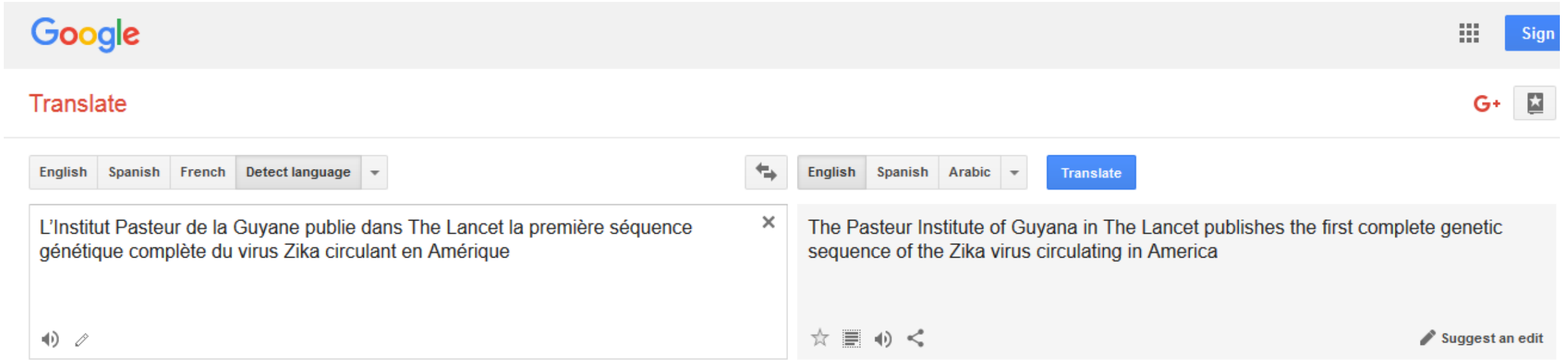
IIT Guwahati

Objective of the lecture

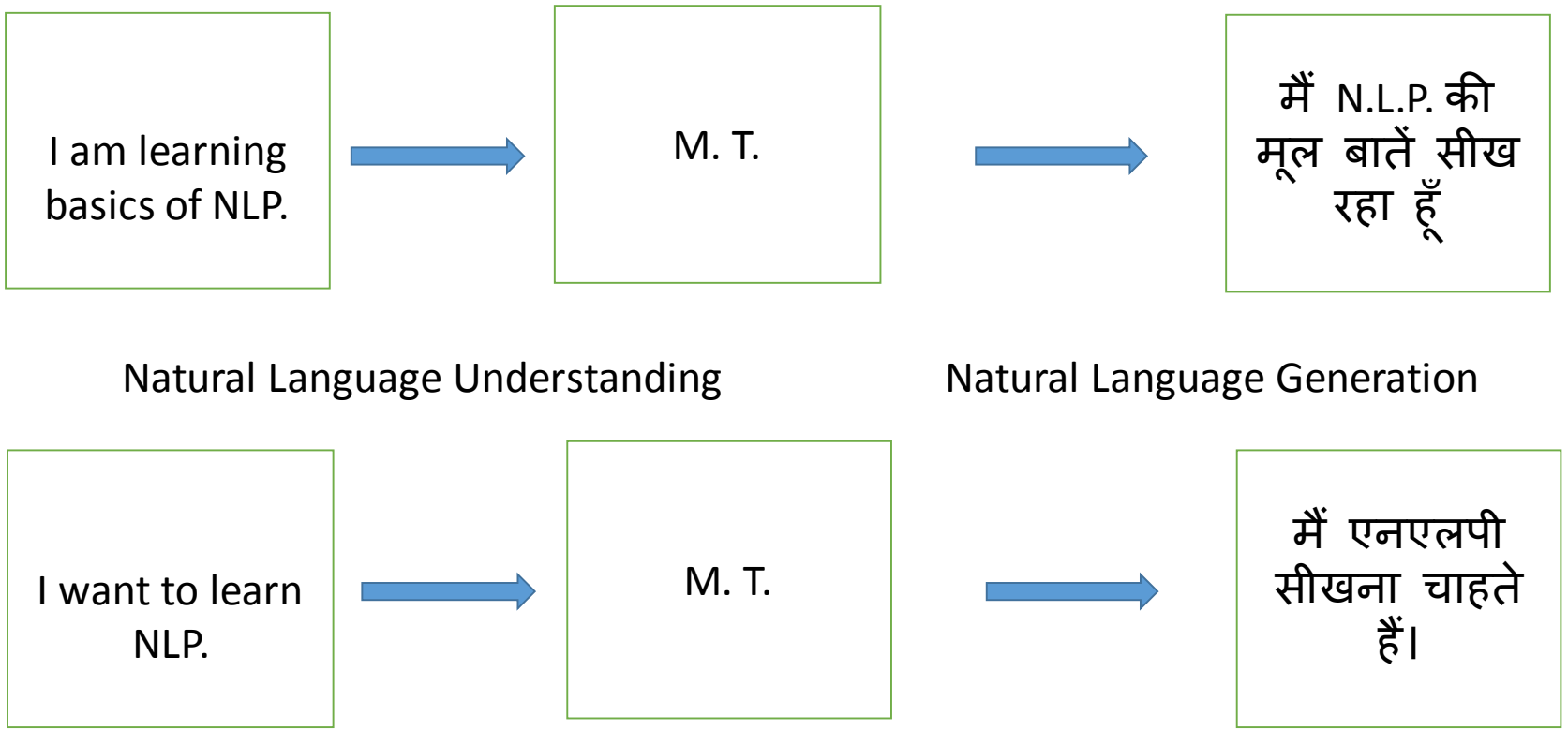
- Understanding what is NLP
- Get an idea why NLP is hard
- Get started with dealing with natural language

Introduction: NLP at High Level

Machine Translation



At the very high level



Information Extraction

10TH DEGREE is a full service advertising agency specializing in direct and interactive marketing. Located in Irvine CA, 10TH DEGREE is looking for an Assistant Account Manager to help manage and coordinate interactive marketing initiatives for a marquee automative account. Experience in online marketing, automative and/or the advertising field is a plus. Assistant Account Manager Responsibilities Ensures smooth implementation of programs and initiatives Helps manage the delivery of projects and key client deliverables . . . Compensation: $50,000-$80,000 Hiring Organization: 10TH DEGREE

⇓

INDUSTRY	Advertising
POSITION	Assistant Account Manager
LOCATION	Irvine, CA
COMPANY	10TH DEGREE
SALARY	$50,000-$80,000

Source: Taken from Collins slide
(https://www.coursera.org/course/nlangp

What do we mean by NLP?

- Natural Language – spoken or written language for expressing ourselves. Example: Hindi, English, Sanskrit, German ..

- NLP
  - Computational processing techniques to process spoken as well as written language.
  - Inclusive Definition

Why it is hard?

- Ambiguity
  - Example: *I made her duck*
  - It can any of the following
    - *I cooked duck for her*
    - *I cooked duck belonging to her*
    - *I caused her to quickly lower her body*

Different Levels of NLP

- Word
  - Phonology: study of linguistic sounds
  - Morphology: study of meaningful components of words
- Syntax: structural relationship between words
- Semantic: study of meaning
- Pragmatic: study at discourse level.

Lets begin: what it takes to make an NLP system

Source

- Corpora (plural for *corpus*: large, (un)structured set of texts)
  - Brown corpus: 500 samples of English texts published in the US in 1961, approx. 1 million words
  - Access to multiple corpus from tools like *NLTK*
  - Building from databases such as PubMed.

Source

- Caution: One shoe does not fit all.

Looking at Text: Basic pre-processing

Tokenization

- Definition: Process to divide the input text into units, also called, *tokens*, where each is either a *word* or a *number* or a *punctuation mark.*
- Should we remove all punctuation marks ?

What counts as a word?

- Kucera and Francis (1967) defined "*graphic word*" as follows :
  - " a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctuation marks"

Problem with graphic word definition

- Should we consider “$12.20” or “Micro$oft” or “:)” as a word?
- We can expect several variants especially in forums like Twitter etc which may not obey exact definition but should be considered as a word.
- Simple Heuristic: *Whitespace*
  - *“*a *space* or *tab* or the *new line*” between words.
  - Still to deal with several issues.

Defining words: Problems

- Periods
  - Abbreviations at the end vs. in the middle
- Single apostrophes
  - Contractions such as I'll, I'm etc.: should be taken as two words or one word?
  - *Penn Treebank* split such contractions.
  - Phrases such as *dog's vs. yesterday's* in "The house I rented yesterday's garden is really big".
  - Orthographic-word-final single quotation such as "boys' toys".

Defining words: Problems

- Hyphenation
  - Again the same question – “do sequences of letters with a hyphen in between count as one word or two?
  - Occurrences like *e-mail*, *co-operate* vs. *non-lawyer, so-called, text-based*
  - Inconsistency in using words like “cooperate” as well as “co-operate”
  - Line-breaking hyphen vs. actual hyphen happens at the end of line [*haplology*]
- Word with a whitespace between its parts
  - New Delhi, San Francisco
  - … the New Delhi-New Jalpaiguri special train …

Defining words: other issues

- Morphology
  - Different forms of words
    - Go, went, gone
    - Fox, foxes
  - Stemming and Lemmatization

Dealing with cases: Main issue

- Can we make all letters in same case
  - Should we treat “*the*”, “*The”, and “THE”* differently vs. “*Mr. Brown*” and “*brown paints*”

Dealing with cases: A Heuristic

- Convert all capital letters to lowercase
  - At the beginning of a sentence, and
  - In headings, titles etc.

- Do we see any problem in this heuristic ?

Problems with the heuristic

- Dependency on correct detection of sentence boundary
- All names appearing in the beginning of the sentence or in places like titles, gets converted
- More importantly, loss of information
  - Example: words in the middle of a sentence but started with capital letter for emphasizing an important point.

- Objective of the study should determine our decision.

Defining Sentence Boundary

- Something ending with a ‘.’, ‘?’, or ‘!’
  - Language specific
- Problem with ‘.’
  - Still 90% of periods are sentence boundary indicators.
- Sub-sentence structure with the use of other punctuation
  - “The scene is written with a combination of unbridled passion and sure-handed control: In the exchanges …….. inexorability of separation”
- Other issues
  - “You remind me,” she remarked, “of your mother.”

Defining Sentence Boundary: A heuristic

- Put putative sentence boundaries after occurrences of ., ?, ! (and may be ;, :, -)
- Check presence of following quotation marks, if any move the boundary.
  - “You remind me,” she remarked, “of your mother.”
- Disqualify a period boundary if –
  - It is preceded by a known abbreviation that does not generally occur at the end of sentence such as Dr., Mr. or vs.
  - It is preceded by a know abbrev. that is not followed by an uppercase word.
- Disqualify a boundary with a ? or ! If
  - It is followed by a lowercase letter (or name)

Issues with Heuristic or set of pre-defined rules

- Is it possible to define such rules without the help of linguists?
- Will it work for all languages?