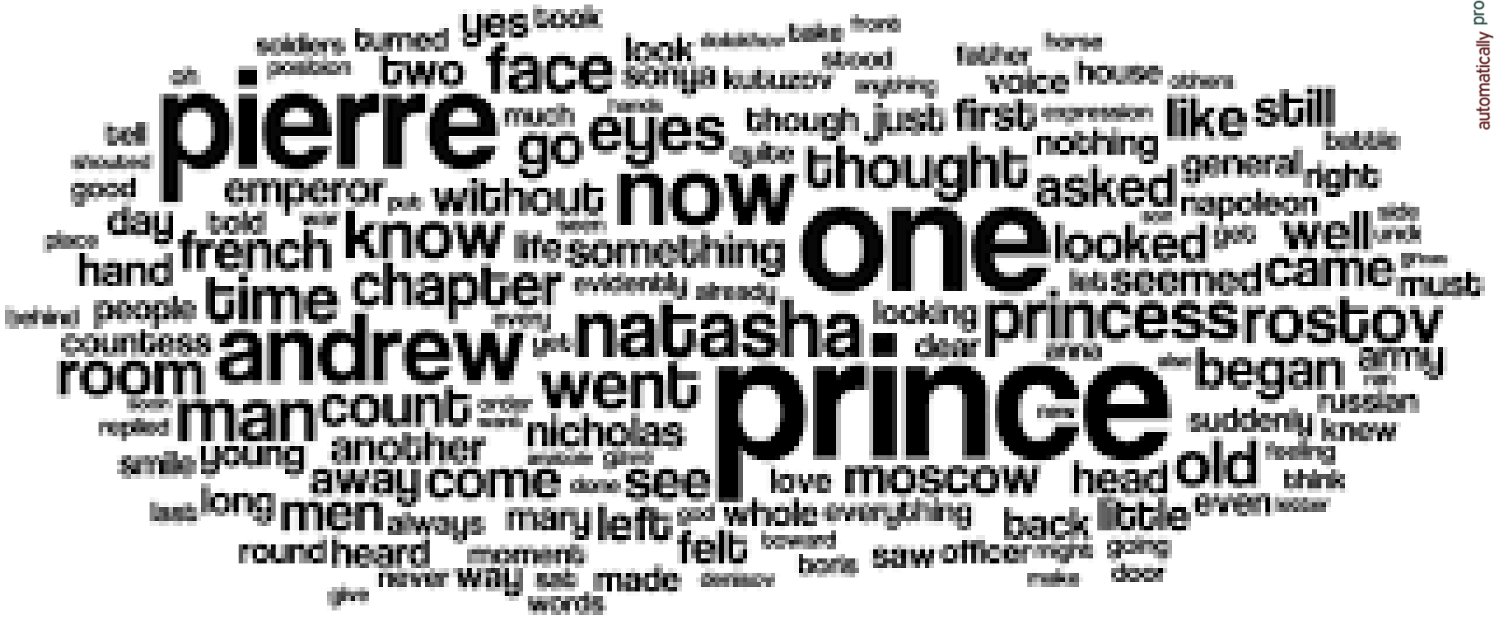
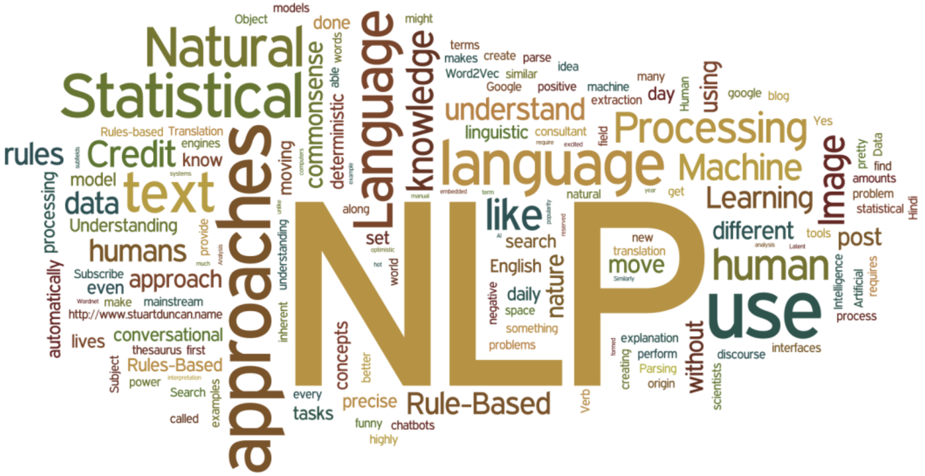
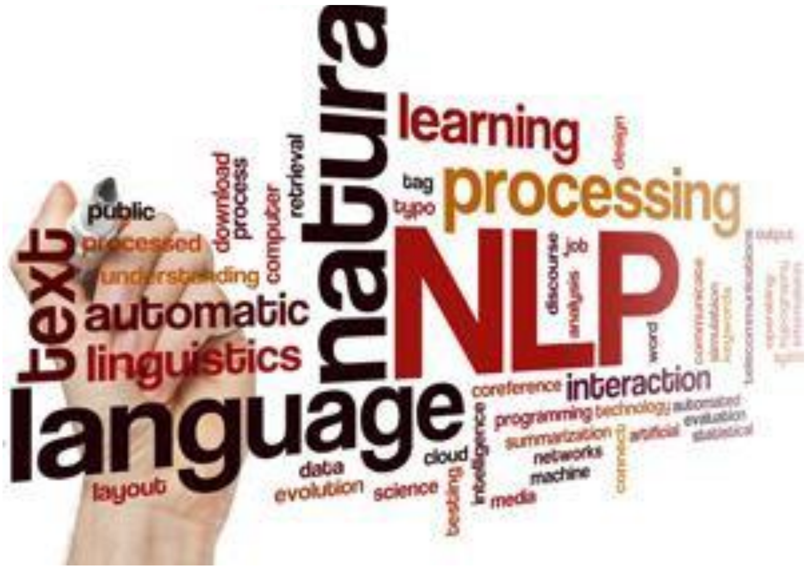
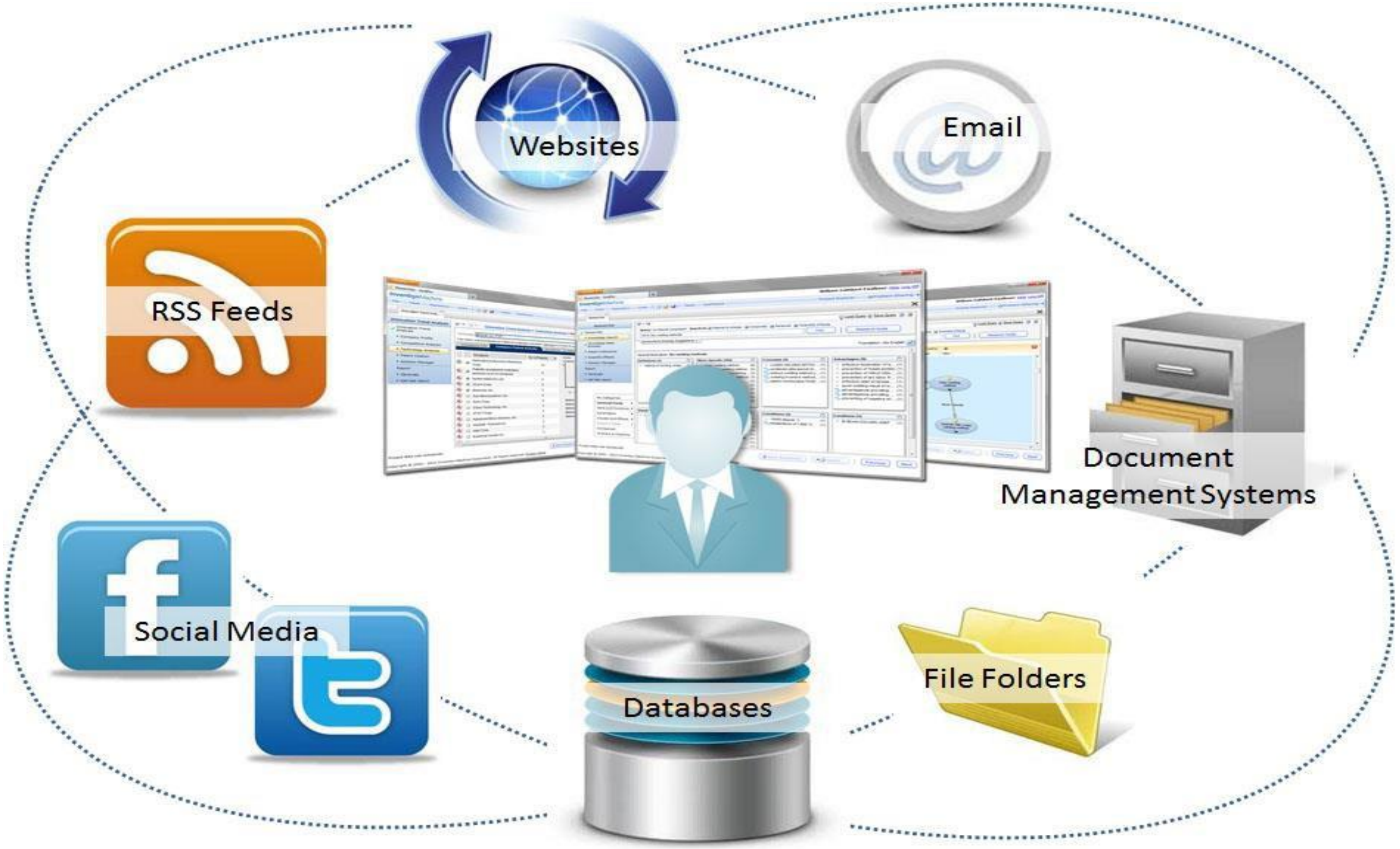
Natural Language Processing & Applications

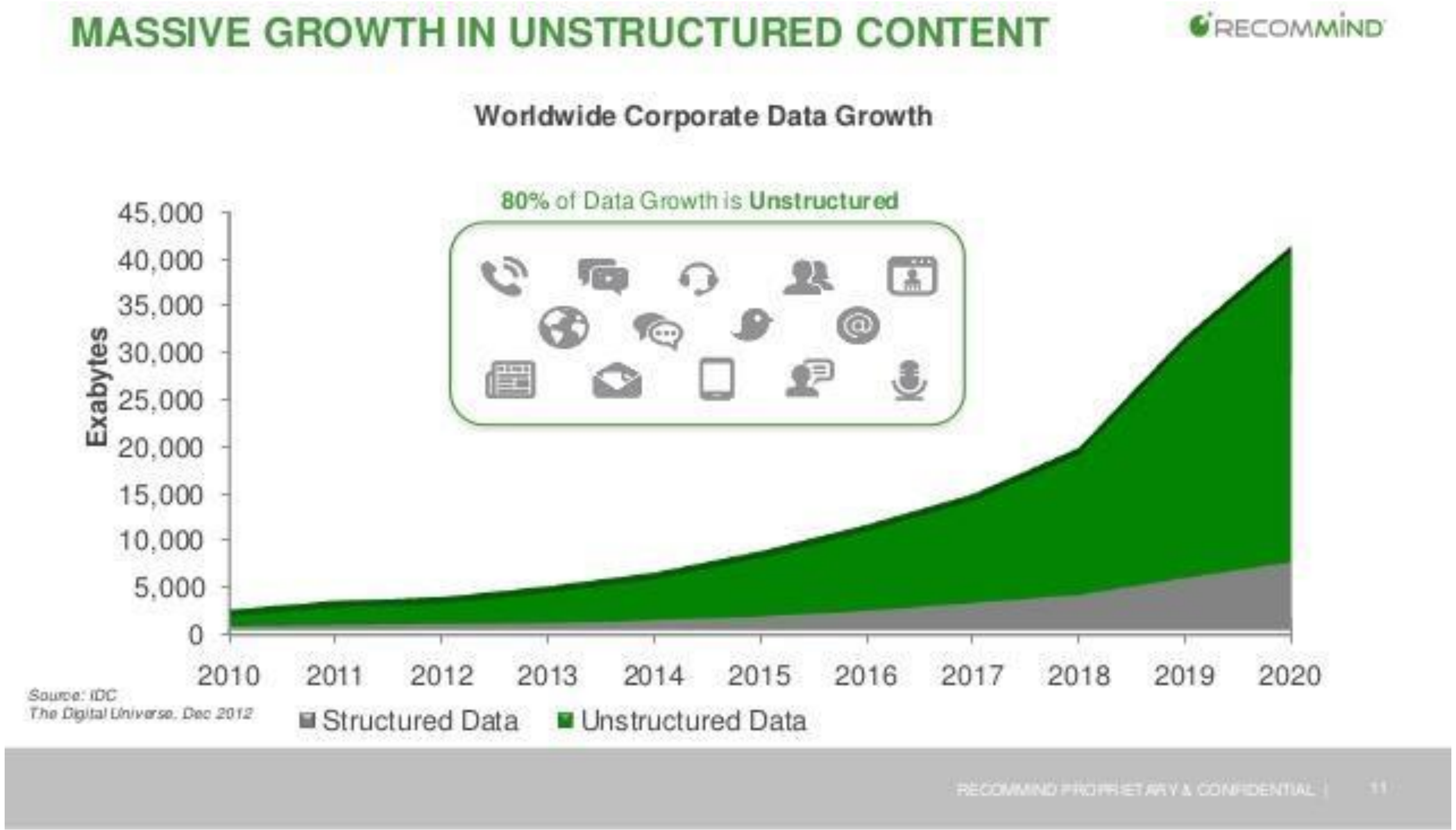






Why Text ?


Websites
Email
RSS Feeds
Document
Management Systems
Social Media
Databases
File Folders



MASSIVE GROWTH IN UNSTRUCTURED CONTENT

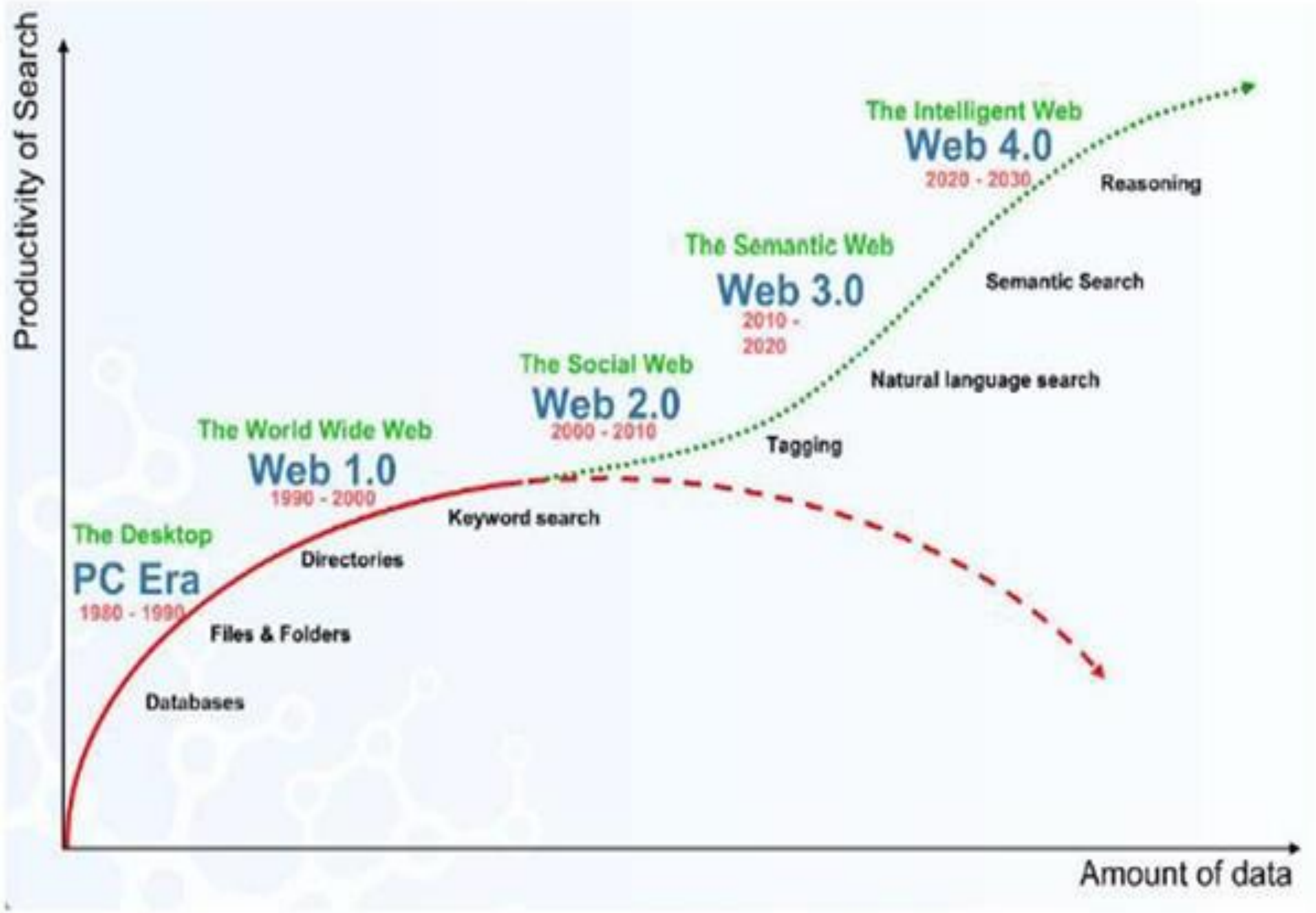
RECOMMIND

Worldwide Corporate Data Growth

Source: IDC
The Digital Universe, Dec 2012

RECOMMIND PROPRIETARY & CONFIDENTIAL | 11

Source : RECOMND


Productivity of Search
The Intelligent Web
Web 4.0
2020 - 2030
Reasoning
The Semantic Web
Web 3.0
2010 -
2020
Semantic Search
Natural language search
The Social Web
Web 2.0
2000 - 2010
Tagging
The World Wide Web
Web 1.0
1990 - 2000
Keyword search
The Desktop
PC Era
1980 - 1990
Directories
Files & Folders
Databases
Amount of data

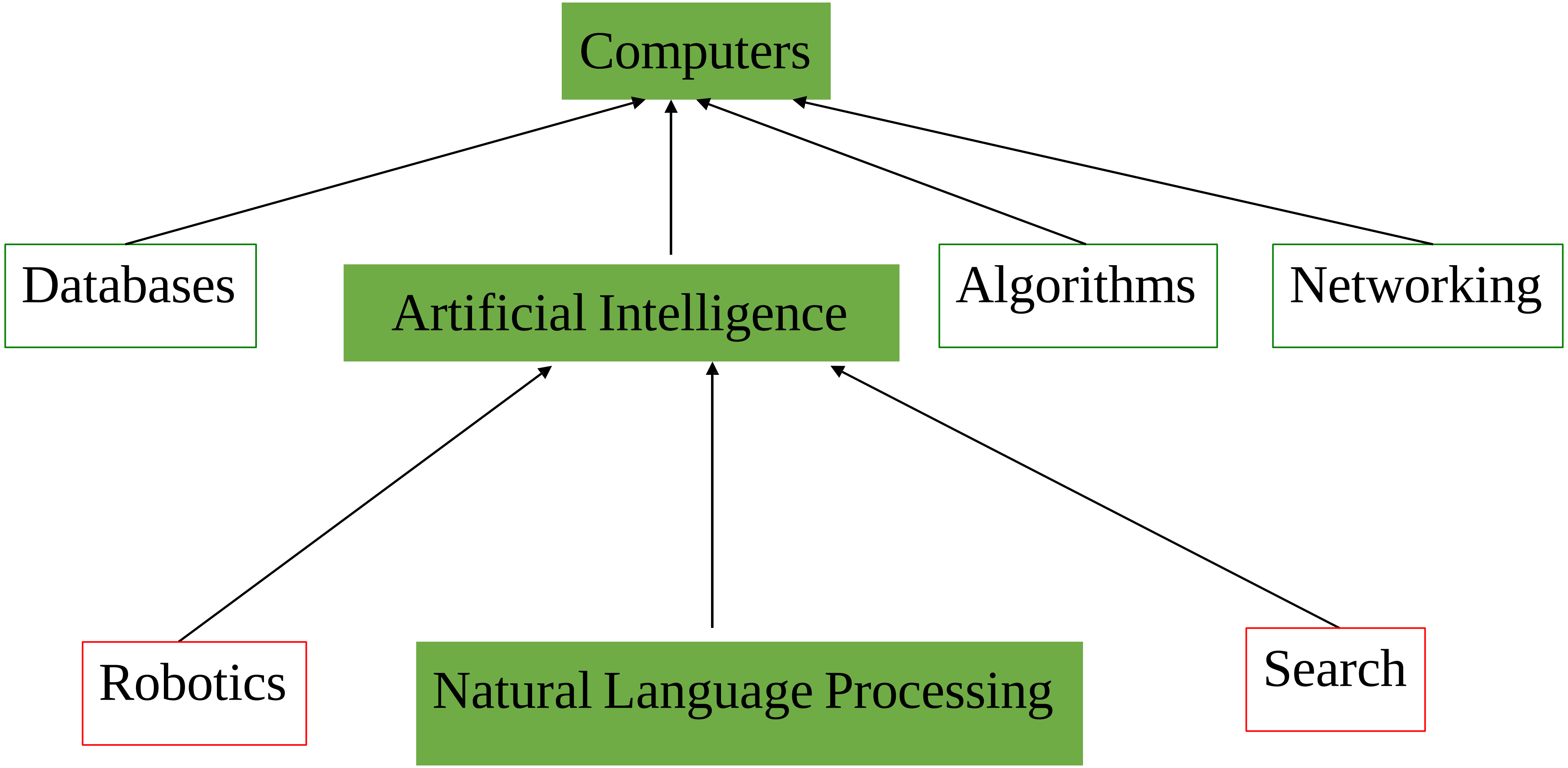
Natural Language Processing

- A hallmark of human intelligence.
- Natural Language Processing
  - Natural Language Understanding
  - Natural Language Generation
  - Process information contained in natural language text.
  - Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)
- Can machines understand human language?

Ultimate goal

Analyze, understand and generate human languages just like humans do.

Fitting in CS taxonomy



NLP- Tasks

Natural Language Understanding

Taking some spoken/typed sentence and working out what it means

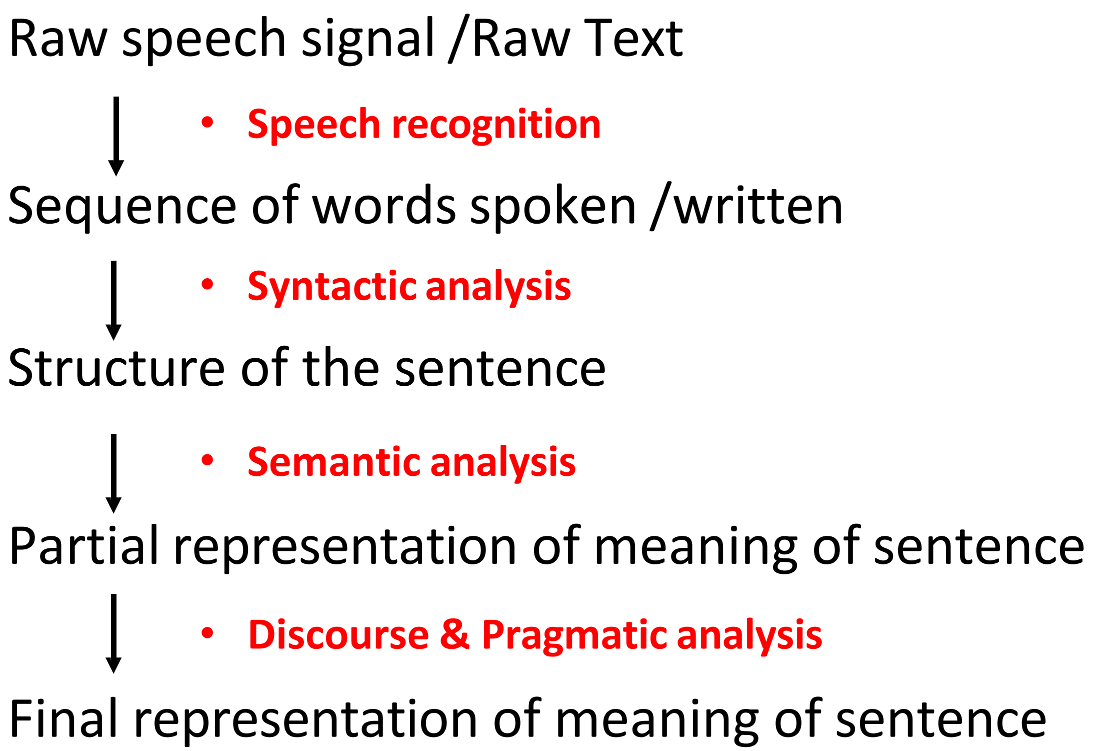
Natural Language Generation

Taking some formal representation of what you want to say and working out a way to express it in a natural (human) language (e.g., English)

Working towards

- Applying computational techniques to language domain.
- Use the theories to build systems that can be of social use.
- Make computers learn our language rather than we learn theirs.

Natural language understanding



Raw speech signal /Raw Text

- **Speech recognition**

Sequence of words spoken /written

- **Syntactic analysis**

Structure of the sentence

- **Semantic analysis**

Partial representation of meaning of sentence

- **Discourse & Pragmatic analysis**

Final representation of meaning of sentence

Aspects of Language Processing

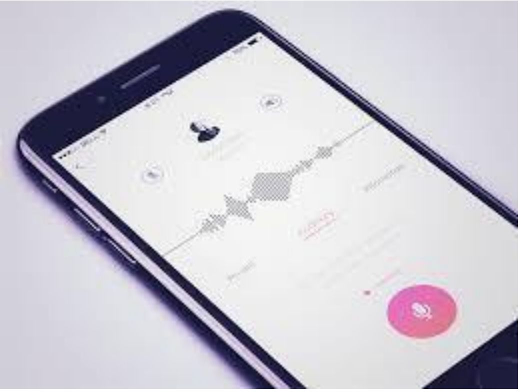
- Phonology – Speech processing
- Word, lexicon: lexical analysis
  - Morphology, word segmentation
- Syntax
  - Sentence structure, phrase, grammar, …
- Semantics
  - Meaning
- Discourse analysis
  - Meaning of a text
  - Relationship between sentences
- Pragmatics

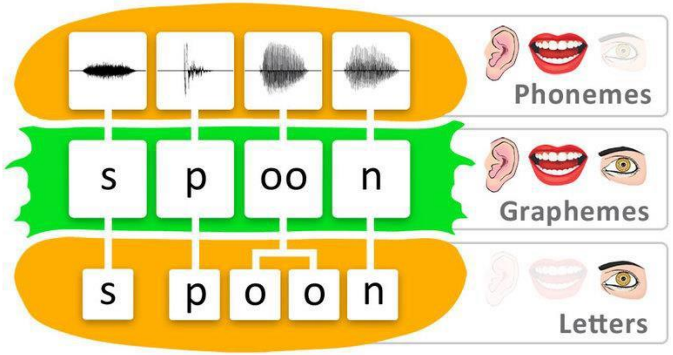
  The study of meaning in different contexts of use

Phonology

Speech processing

- Humans process speech remarkably well.
- Speech interface can replace keyboards and monitors.
- Convert Acoustic signals to Text.
- Phonemes are the smallest recognizable speech unit in a language.

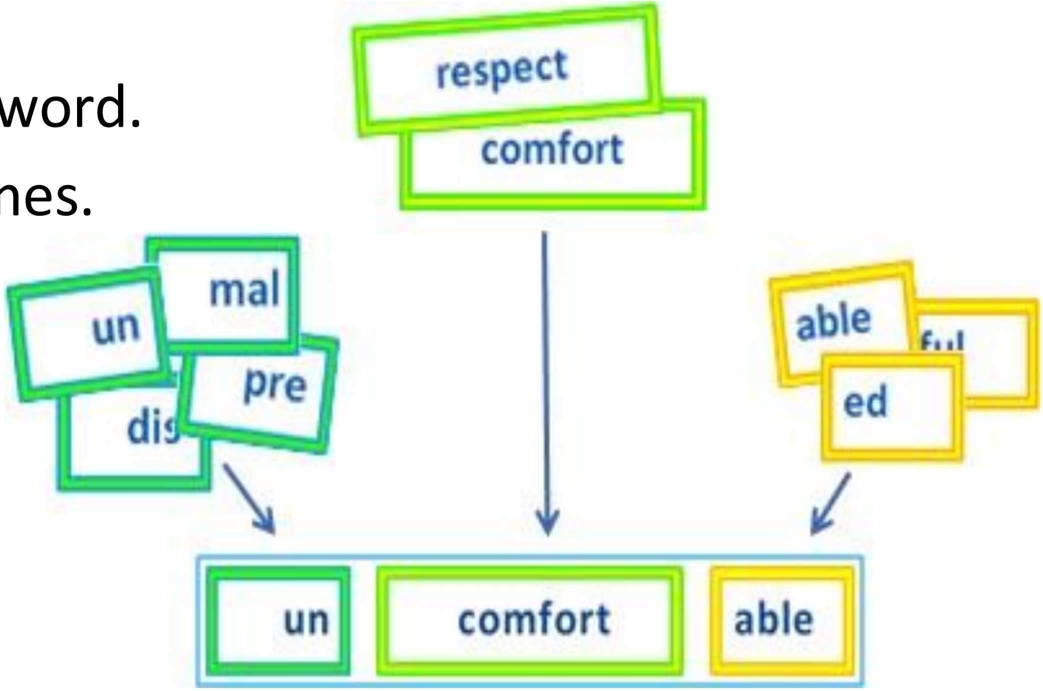




Morphology

- Structures and patterns in words
- Words are a sequence of Morphemes.
  - Morpheme – smallest meaningful unit in a word.
- Analyses how words are formed from morphemes.

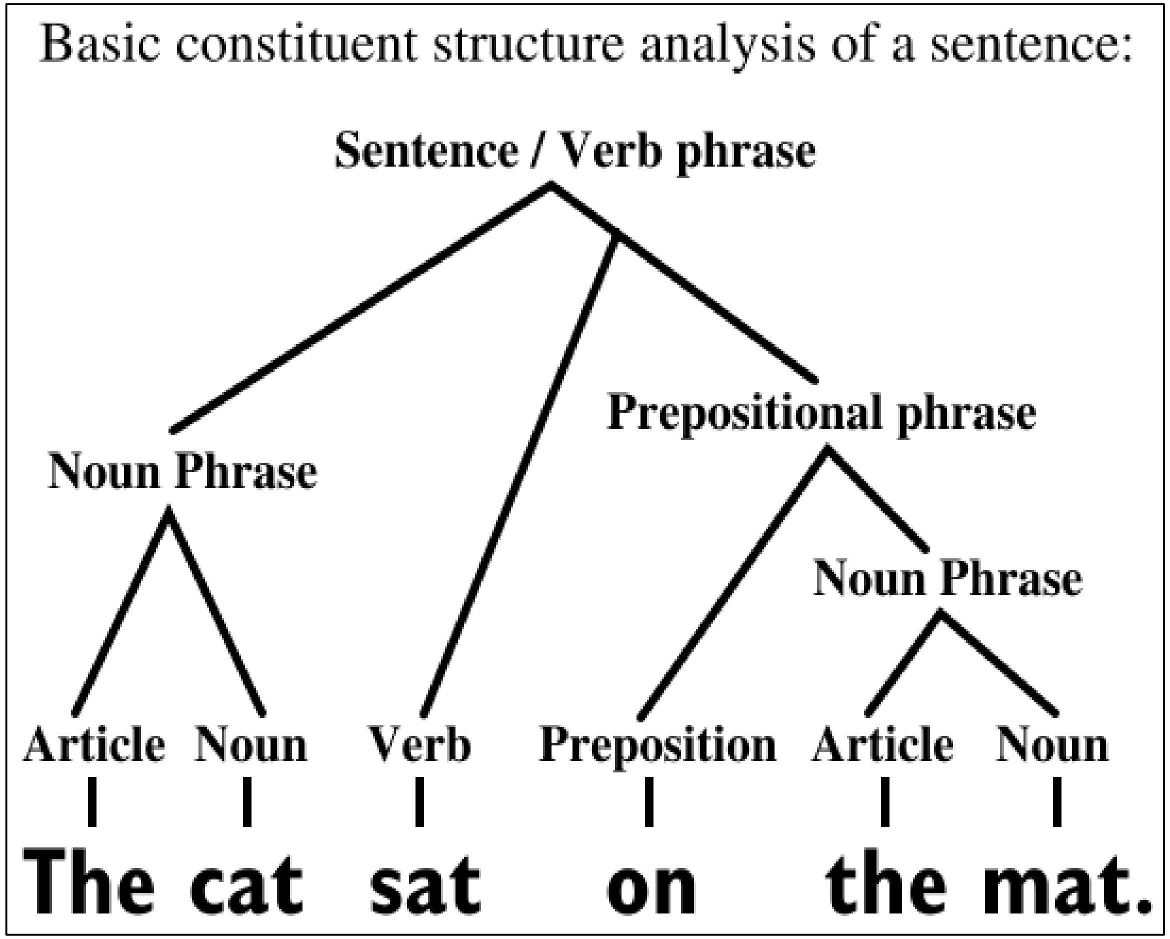
  e.g., dogs= dog+s.
- Inflectional Morphology – Same Part of Speech
  - Buses = Bus + es
  - Carried = Carry + ed
- Derivational Morphology – Change PoS.
  - Destruct + ion = Destruction (Noun)
  - Beauty + ful = Beautiful (Adjective)
- Affixes – Prefixes, Suffixes Rules govern the fusion.



Syntax

- Words when put together they convey more.
- Syntax is the grammatical structure of the sentence.
- Syntactic Analysis (Parsing)

    Process of assigning a parse tree to a sentence.

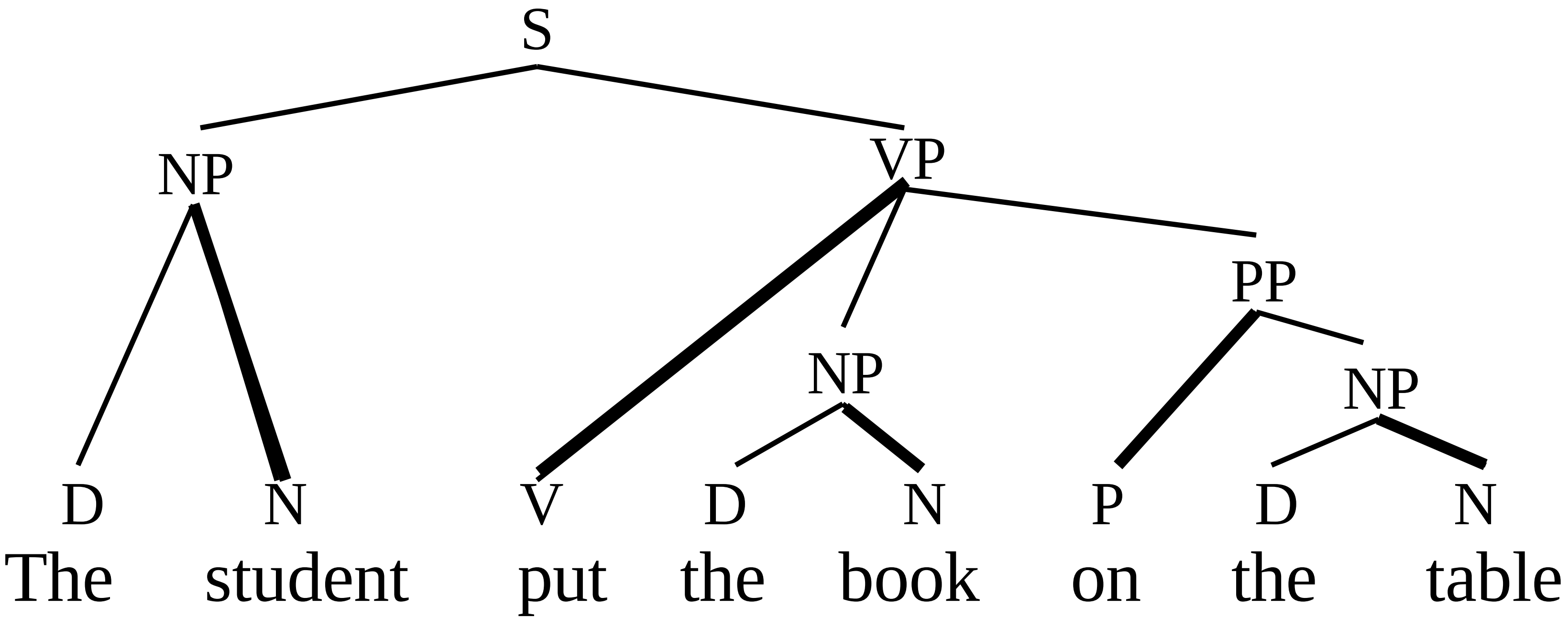


Syntactic Analysis - Grammar

```
sentence -> noun_phrase, verb_phrase
noun_phrase  -> proper_noun
noun_phrase  -> determiner, noun
verb_phrase  -> verb,
                noun_phrase
proper_noun  -> [mary]
noun -> [apple]
verb -> [ate]
determiner -> [the]
```

Parsing

- Analyze the structure of a sentence



Semantic Analysis

- What do you mean..?
- Words – Lexical Semantics
- Sentences – Compositional Semantics
- Converting the syntactic structures to semantic format – meaning representation.
- Semantics: the meaning of a word or phrase within a sentence

e.g., "I saw the prudential building flying into Boston"

Semantic Representations

- Meaning representation of the sentence from its syntactic structure(s)
- Ways of meaning representing the sentence:
  - Logical forms

    Representation: Ξx man(x) & tall(x) & plays(x, basketball)
  - Semantic role labelling

    Representation: Kill(Agent : X, Victim: 3 people, Instrument : Gun)

Discourse Analysis

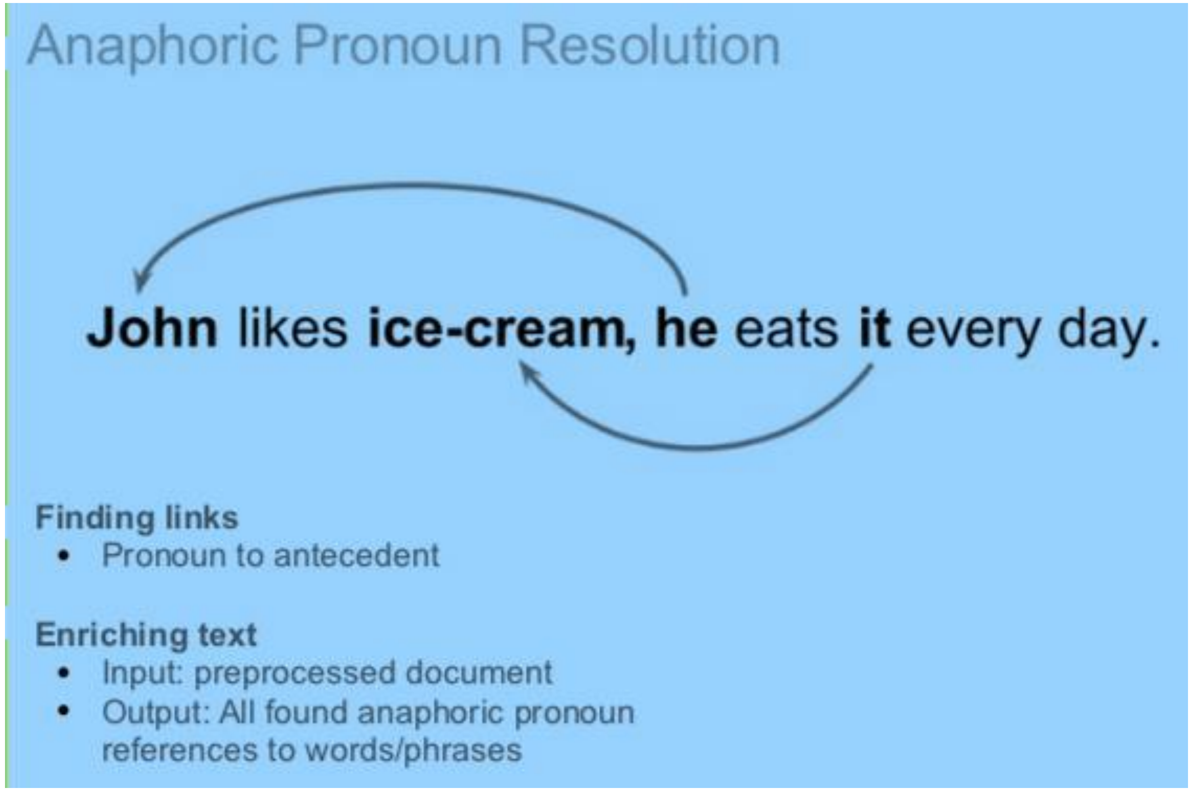
- The meaning of an individual sentence may *depend on the sentences that precede* it and may *influence the meaning of the sentence that follow it*.
- Issues related to discourse Integration
  - Anaphora

    Resolving the pronoun's reference. Co-reference resolution

    Coreference resolution is the task of finding all expressions that refer to the same entity in a text.
  - Ellipsis

    Incomplete sentences

Anaphora Resolution



Anaphoric Pronoun Resolution

John likes **ice-cream, he** eats **it** every day.

Finding links

- Pronoun to antecedent

Enriching text

- Input: preprocessed document
- Output: All found anaphoric pronoun references to words/phrases

Discourse Structures- Ellipsis

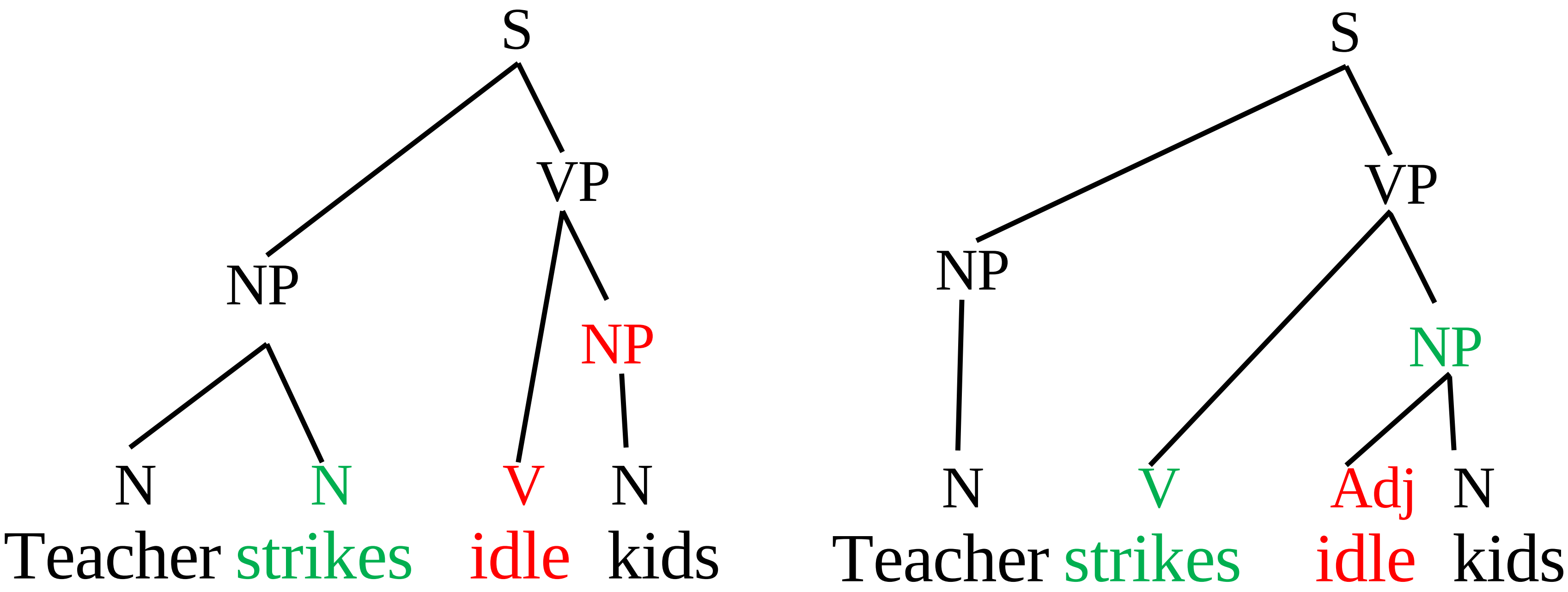
- Ellipsis – Incomplete sentences

The second sentence is not complete, but what it means can be inferred from the first one.

Pragmatics

- Uses context of utterance
    - Where, by who, to whom, why, when it was said
    - ***Intentions:*** *inform, request, promise, criticize,* …

Syntax Ambiguity



Semantic Ambiguity

Semantic ambiguity: **“I saw the prudential building flying into Boston”**

Semantic Restriction, Domain Knowledge - Ontology

Pragmatics Ambiguity

Pragmatic ambiguity: **“you’re late”**

What’s the speaker’s intention: informing or criticizing?

Computing Techniques

- Stemming
  - Reduce *words* to base form.
- POS Tagging
  - Determine for each word whether it is a noun, adjective, verb, …..
- Parsing
  - sentence ⇒ parse tree
- Word Sense Disambiguation
  - orange juice vs. orange coat
- Semantic analysis
  - Semantic representations and Evaluation

Application Areas

- Machine Translation
- Information Retrieval

  Selecting from a set of documents the ones that are relevant to a query

- Text Categorization

  Sorting text into fixed topic categories

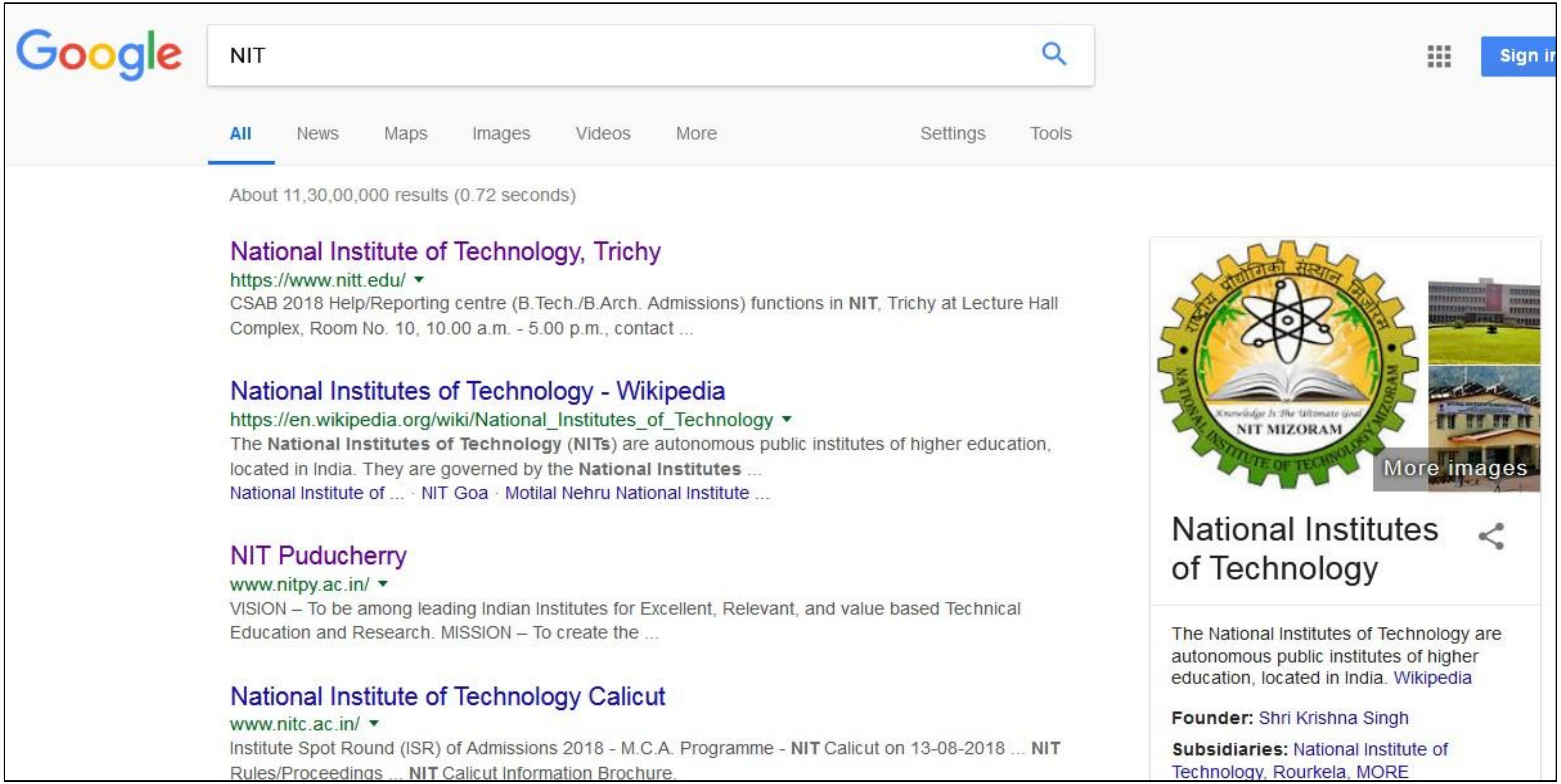
- Question Answering
- Information Extraction

  Converting unstructured text into structured data

Application Areas (cont..)

- Spoken language control systems
- Spelling and grammar checkers
- Sentiment Analysis
- Text-to-Speech & Speech recognition
- Natural Language Dialogue Interfaces to Databases
- Plagiarism detection

Information Retrieval



About 11,30,00,000 results (0.72 seconds)

National Institute of Technology, Trichy
https://www.nitt.edu/ ▾
CSAB 2018 Help/Reporting centre (B.Tech./B.Arch. Admissions) functions in **NIT**, Trichy at Lecture Hall Complex, Room No. 10, 10.00 a.m. - 5.00 p.m., contact ...

National Institutes of Technology - Wikipedia
https://en.wikipedia.org/wiki/National_Institutes_of_Technology ▾
The **National Institutes of Technology** (**NITs**) are autonomous public institutes of higher education, located in India. They are governed by the **National Institutes** ...
National Institute of ... · NIT Goa · Motilal Nehru National Institute ...

NIT Puducherry
www.nitpy.ac.in/ ▾
VISION – To be among leading Indian Institutes for Excellent, Relevant, and value based Technical Education and Research. MISSION – To create the ...

National Institute of Technology Calicut
www.nitc.ac.in/ ▾
Institute Spot Round (ISR) of Admissions 2018 - M.C.A. Programme - **NIT** Calicut on 13-08-2018 ... **NIT** Rules/Proceedings ... **NIT** Calicut Information Brochure.

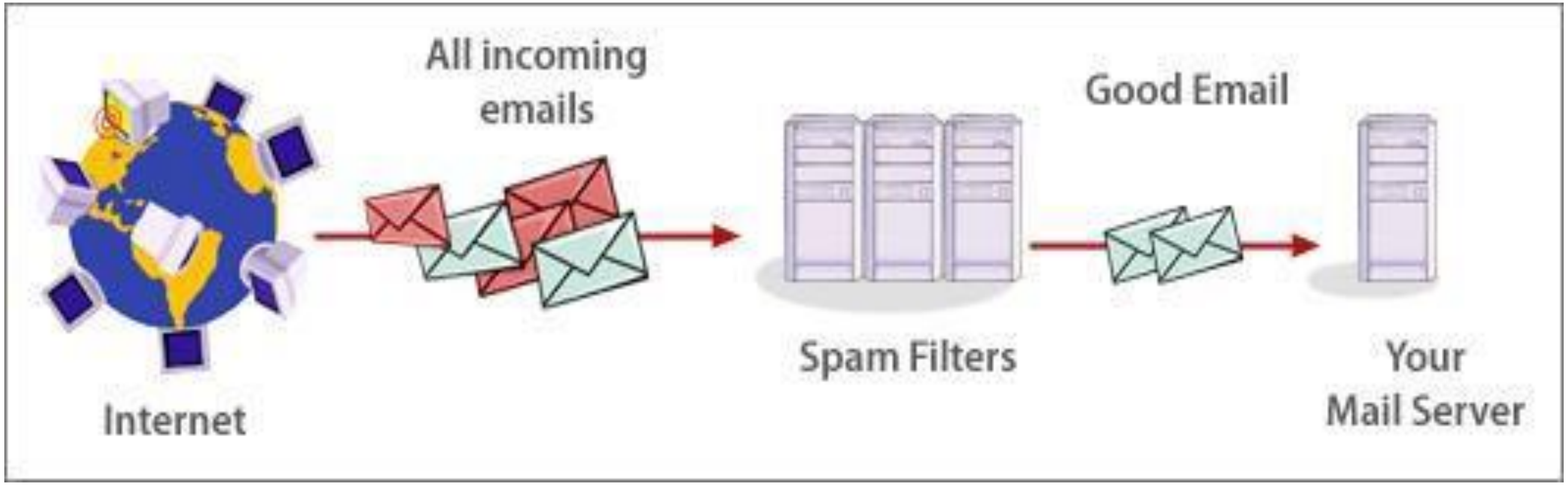
National Institutes of Technology

The National Institutes of Technology are autonomous public institutes of higher education, located in India. Wikipedia

Founder: Shri Krishna Singh

Subsidiaries: National Institute of Technology, Rourkela, MORE

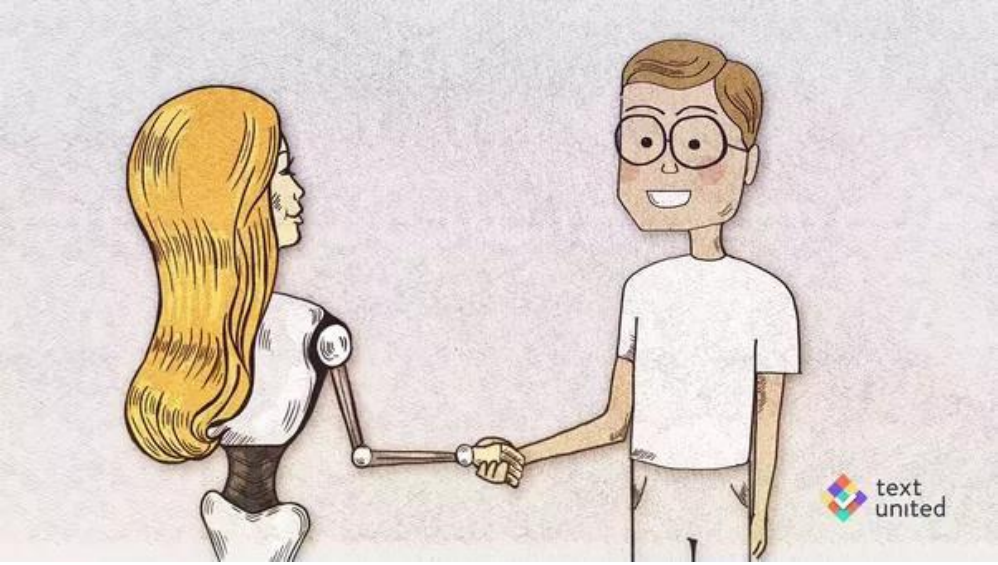
Email Spam Filtering/Categorizing

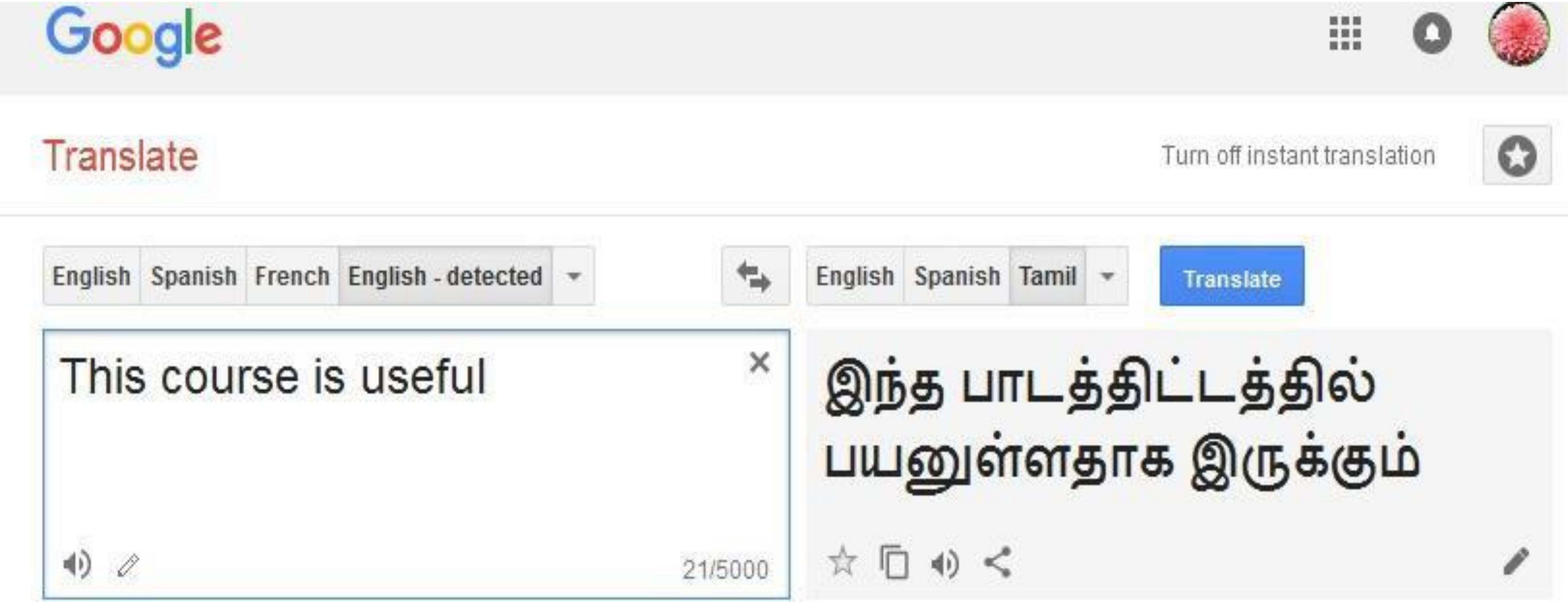


Machine Translation

- Multilingual Usage
- Machine-assisted human Translation
- Scope

  Creating Language resources.





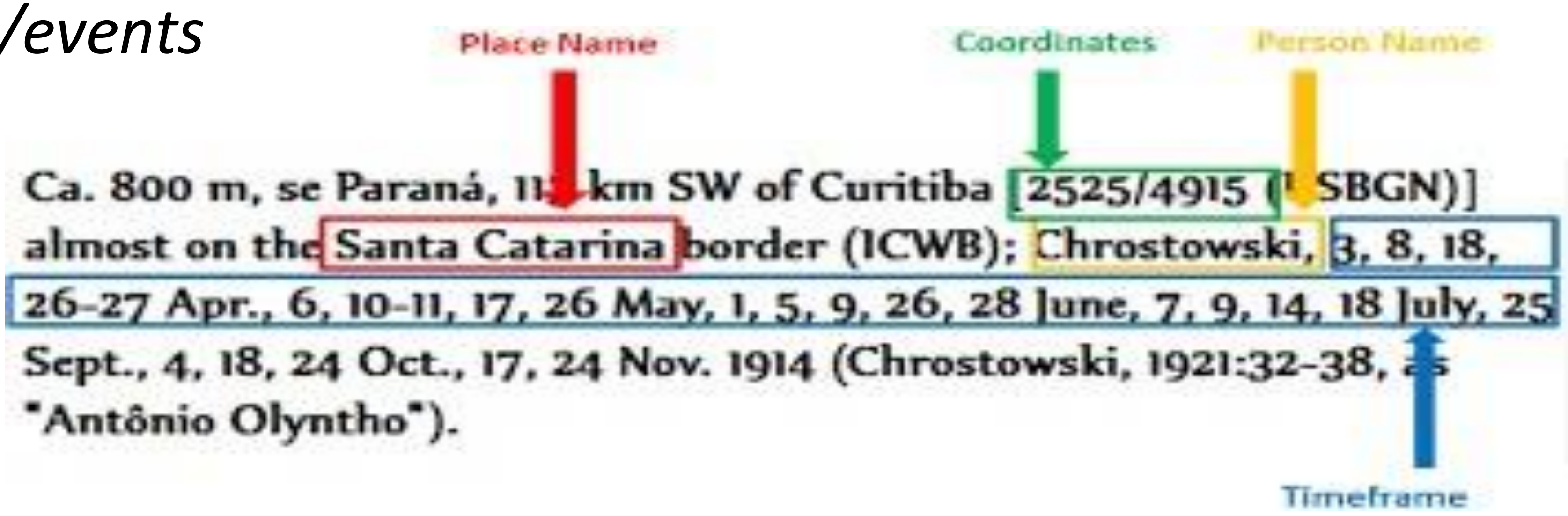
This course is useful

இந்த பாடத்திட்டத்தில் பயனுள்ளதாக இருக்கும்

Information Extraction

Information extraction systems

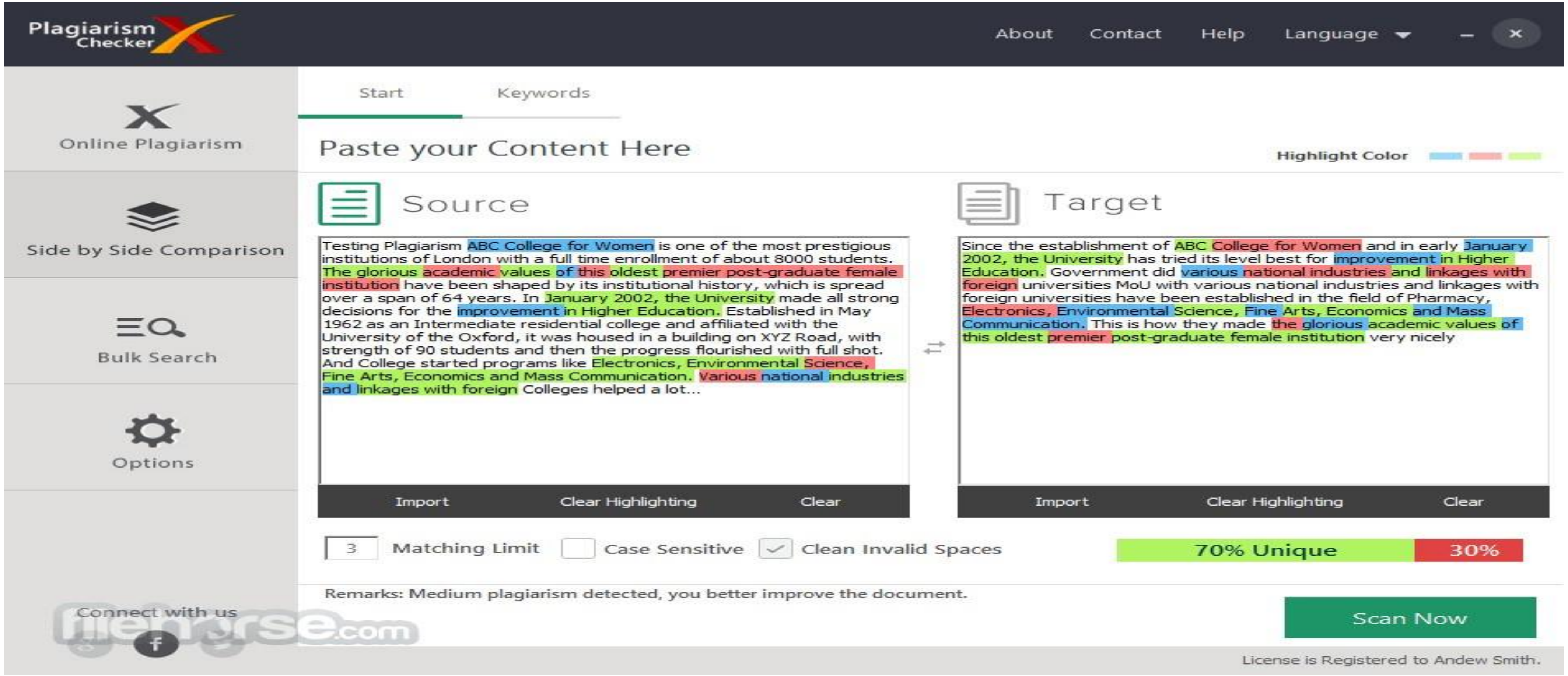
- Find and understand relevant parts of text.
- Produce a structured representation of the relevant information from text, in the form of :
  - entities,
  - relations between entities ,
  - events in which the entities are involved.
- Produce a structured representation of the relevant information- *relations/events*



Ca. 800 m, se Paraná, 11[illegible] km SW of Curitiba [2525/4915 (¹ SBGN)] almost on the Santa Catarina border (ICWB); Chrostowski, 3, 8, 18, 26-27 Apr., 6, 10-11, 17, 26 May, 1, 5, 9, 26, 28 June, 7, 9, 14, 18 July, 25 Sept., 4, 18, 24 Oct., 17, 24 Nov. 1914 (Chrostowski, 1921:32-38, as "Antônio Olyntho").

Place Name
Coordinates
Person Name
Timeframe

Plagiarism Detection



Source:

Testing Plagiarism ABC College for Women is one of the most prestigious institutions of London with a full time enrollment of about 8000 students. The glorious academic values of this oldest premier post-graduate female institution have been shaped by its institutional history, which is spread over a span of 64 years. In January 2002, the University made all strong decisions for the improvement in Higher Education. Established in May 1962 as an Intermediate residential college and affiliated with the University of the Oxford, it was housed in a building on XYZ Road, with strength of 90 students and then the progress flourished with full shot. And College started programs like Electronics, Environmental Science, Fine Arts, Economics and Mass Communication. Various national industries and linkages with foreign Colleges helped a lot...

Target:

Since the establishment of ABC College for Women and in early January 2002, the University has tried its level best for improvement in Higher Education. Government did various national industries and linkages with foreign universities MoU with various national industries and linkages with foreign universities have been established in the field of Pharmacy, Electronics, Environmental Science, Fine Arts, Economics and Mass Communication. This is how they made the glorious academic values of this oldest premier post-graduate female institution very nicely

3 Matching Limit ☐ Case Sensitive ☑ Clean Invalid Spaces

70% Unique 30%

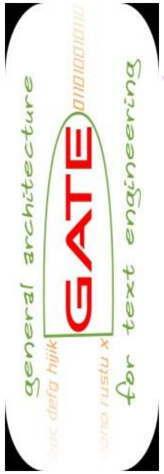
Remarks: Medium plagiarism detected, you better improve the document.

License is Registered to Andew Smith.

Current Technology Giants

- IBM Watson
- Google
- Microsoft Oxford
- Wolfram Natural Language Understanding System

TOOLS


GATE
general architecture for text engineering

OPEN NLP ™

Stanford University
S
NLP
Natural Language Processing

DBpedia Spotlight

spaCy

OPEN CALAIS
THOMSON REUTERS

Conclusion

- Complete human-level natural language understanding is still a distant goal
- Develop Algorithms for each level.
- Find appropriate match between application domain and the available methods

References

Books

- Allen.J, "Natural Language Understanding ", second edition, Benjamin Cummings Publishing Co .
- Manning.C.D and Schutze.H, "Foundations of Statistical Natural Language Processing ", MIT Press.

Web references

- research.microsoft.com/en-us/groups/nlp/
- nlp.stanford.edu/