Introduction

Institute of Computational Linguistics
Peking University
王厚峰
Wanghf@pku.edu.cn

Basic Course Information

- 1: Intro
- 2: ML in NLP
- 3: KNN,Decision-Tree(*)
- 3: ANN(Artificial Neural Net)
- 4: SVM
- 5: MEM
- 6: Naïve Bayes
- 7: Intro-to-Graphical_Model
- 8: Inference-GM

Basic Course Information

- 9: **Student's Presentation I**
- 10: Learning-GM
- 11: GM-HMM-MEMM-CRF
- 12: Clustering
- 13: Semi_Supervised Learning
- 14: **Student's Presentation II**
- 15: Ensemble

Basic Course Information

- Instructor: Wang Houfeng
  - Email: wanghf@pku.edu.cn
  - Office phone: 10-62753081 (ext.106)
- Score count:
  - Projects
  - Final exam (or paper)

Outline

- **What's Computational Linguistics?**
- Why is Natural Language Processing (NLP) important (Applications)?
- Difficulties
- Brief History

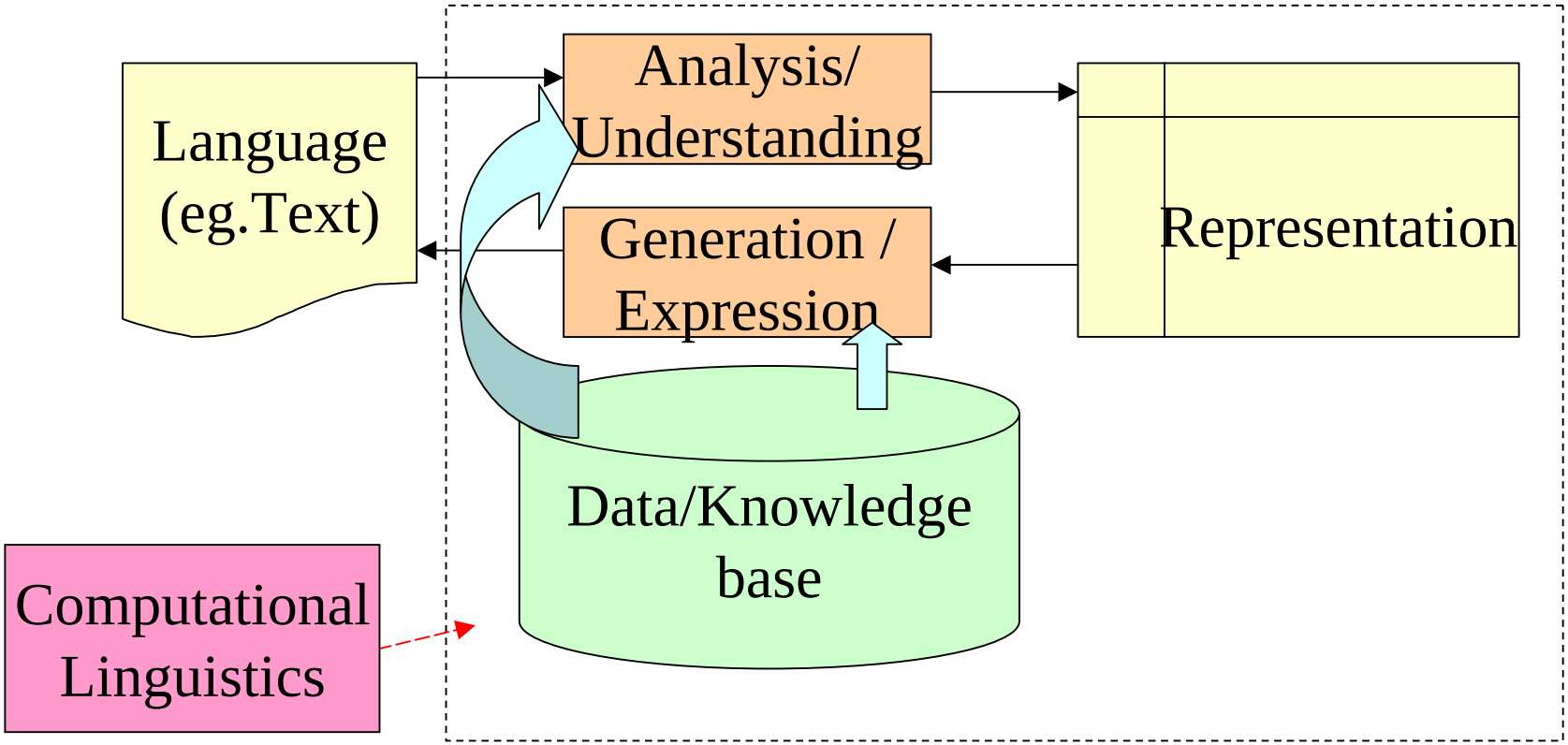
Computational Linguistics

- Computational Linguistics or NLP?
- Task (Main Topics)

Frame of Question–Solving

- Data Representation:
  - Input form: Human → Computer
  - Inter-form: Computer → Computer
  - Output form: Computer → Human
- Knowledge Representation (AI **)
  - Relation: used as reasoning
- Algorithm + Infering Mechanism(**)
  - Manipulate representation and produce desired action or result

Computational Linguistics



Computation on NL

- Popular Applications of computation?
  - Numerical Processing(Input: Numerical value);
  - Simple symbol (character) processing;
  - Table processing;
  - Graph / image processing;
  - Speech/voice processing;
  - ……ultimate aim (AI)——Human Language Processing?
- The Study of Computational Processing of Natural Language (shared above processing frame):
  - Data + Algorithms , or
  - Knowledge + Reasoning
  - The object: Natural Language

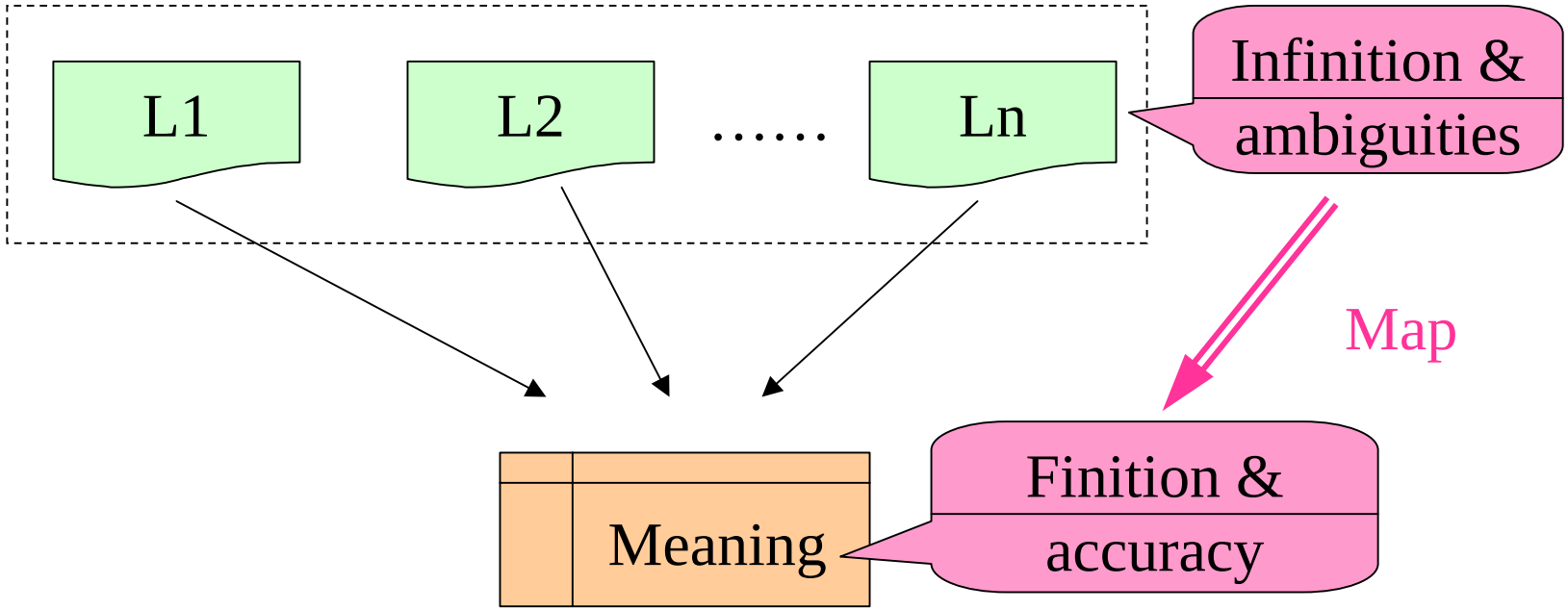
Characters of Computational Linguistics(1)

- Linguistic Essentials: Big integration
- (phonetics, syntax, semantics) Vs.(sound, form, meaning)

  all kinds of data representation (linguistic): sound, sound pieces, word, word pieces, sentence, sentence pieces...
  - Linguistics
  - psychology (Cognitive Science)
  - mathematics

Characters of Computational Linguistics(2)

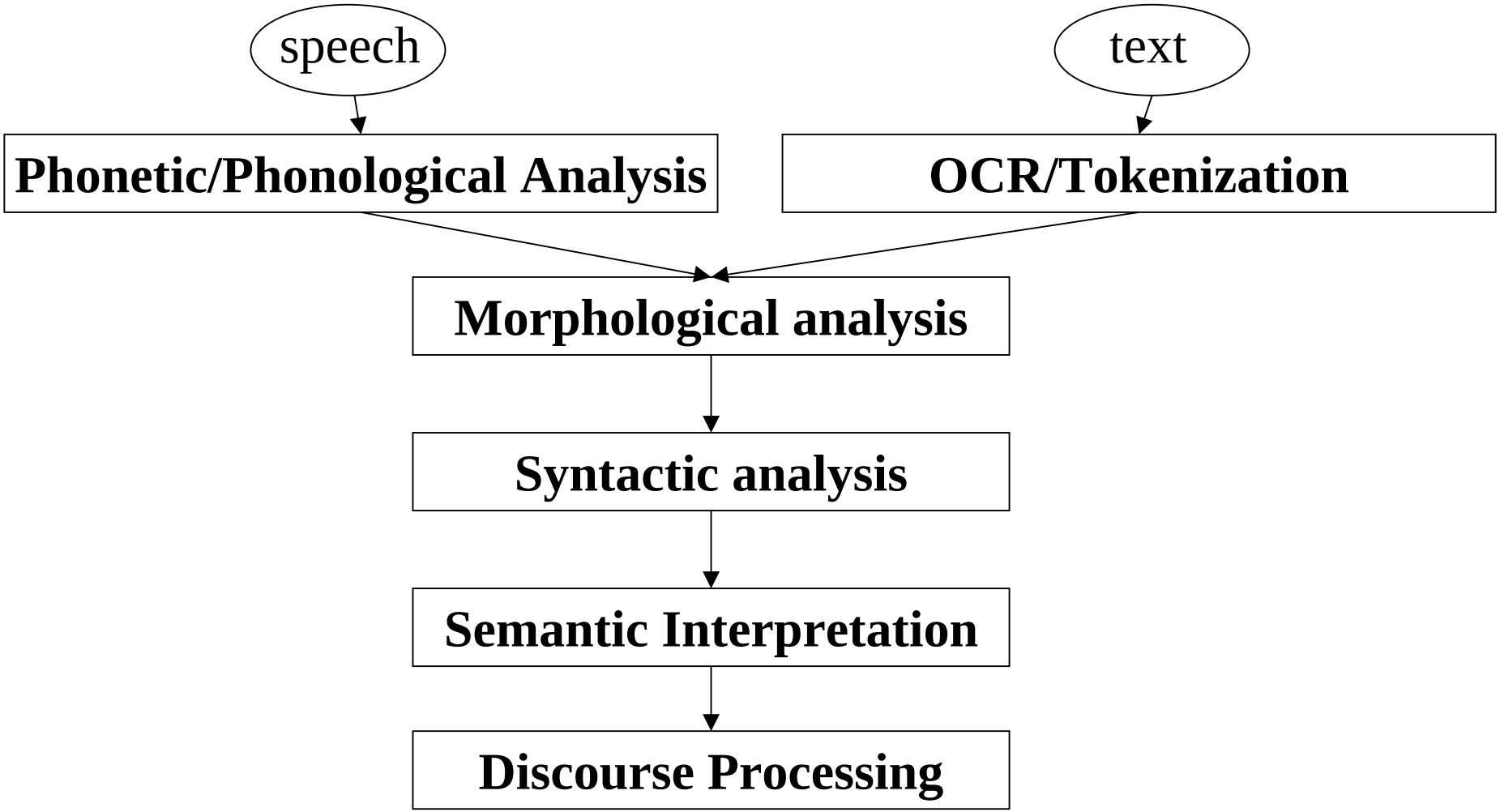
- Formalization of Natural language
  - Mathematics, Statistics, logic etc.
  - Computation Science , Artificial Intelligence etc.



Characters of Computational Linguistics(3)

- Multi-level(multi-mapping):
  - *Phonetics(speech, sound pattern in language)*
  - *Morphology (word and its function)*
  - *Phrase structure or Syntax(form sentence)*
  - *Thematic structure: (who did what to whom)*
  - *Semantics (word meaning and combination into sentence meaning)*
  - *Discourse(Pragmatics & context)*

Pipeline


speech
text
Phonetic/Phonological Analysis
OCR/Tokenization
Morphological analysis
Syntactic analysis
Semantic Interpretation
Discourse Processing

Formal models

- State Machines:
  - FSAs, Markov Models, ATNs etc.
- Formal Rules
  - CFG, Dependency Grammar, PCFG etc.
- Logic-based
  - First Order Logic, High Order Logic,etc.
- Uncertainty Model
  - Naïve Bayse, Probability Theory, Etc.

Algorithms

- State Space search(Parsing,MT)
  - Classical AI Problem
  - NP hard?
- Dynamic Programming
  - Avoiding recomputing

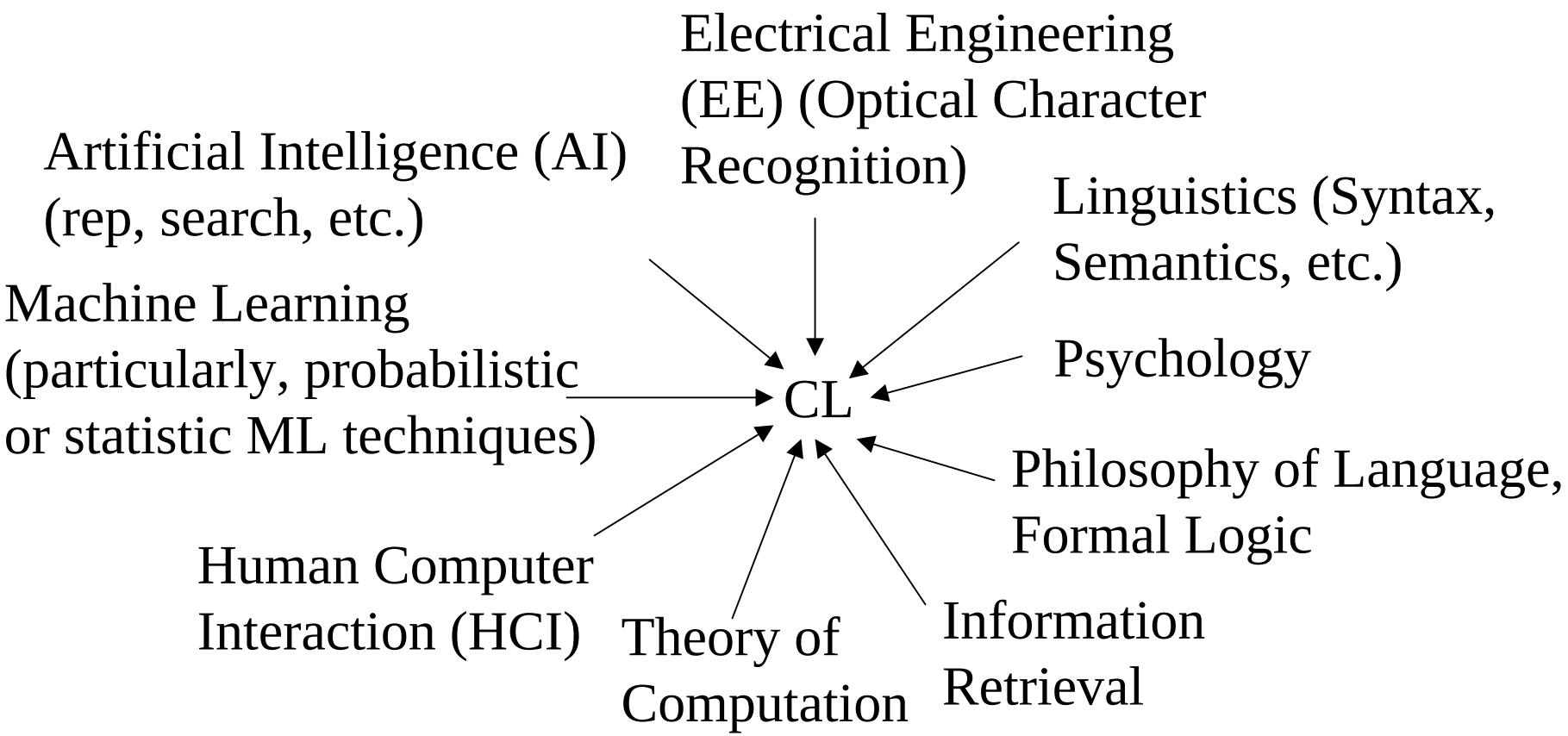
Computational Linguistics Vs. NLP

- Computational Linguistics: Principle
  - How to compute (Computers dealing with language)?
  - Modeling how people do
  - Fundamental Research: Universal methods
- NLP: Applications (implementation)

General Term

- Computational Linguistics
- Natural Language Processing
- Human Language Technology

Relation of CL to Other Disciplines



Outline

- What's Computational Linguistics?
- **Why is Natural Language Processing (NLP) important (Applications)?**
- Difficulties
- Brief History

We need NLP Applications(1)

- Chinese Character Input:
  - Input by keyboard
    - Automatically translate from string of Pin-Yins to string of Chinese characters
    - Precision (People vs. Computer)
  - Input by Voice
- Input Check(correct):
  - Spell check(eg. English),
  - grammar check, (OCR, Pin-Yin, etc.)…

Applications(2)

- Text Retrieval
  - Internet & digital Library:
    - Query (Keywords, Phrase or Sentences: units of NL)
    - Search and find relevance documents
    - Example:

华人 Search

1.

新浪教育出国之海外生活 提供海外华人生活琐事、奋斗经历、见闻趣事等。
生活服务 > 出国服务 > 移民

2.

新华通讯社 中华人民共和国的国家通讯社，是中国最大的新闻信息采集和发布中心。
新闻媒体 > 新闻机构组织 > 通讯社 > 新华社

3.

中华残疾人服务网 残疾人发起的、自己制作的、为华人圈残疾人（残障人）服务的网站。
社会文化 > 残障

4.

亚洲交友中心 有数百万华人会员的大型交友社区，让你安全而能有隐私的交友，寻找属于自己的缘分。
生活服务 > 婚恋与交友服务 > 婚介交友

5.

中华人民共和国卫生部 地址：北京西城区后海北沿44号；邮编：100725。
政法军事 > 政府与国家机构 > 中央政府/组织机构 > 国务院各部委 > 卫生部

Applications(3)

- Question-Answering

  Where is summer palace? How get there?

  – Weather
  – Airline
  – Buying
  – Find person
  – Etc.

Applications(4)

- Text Categorization:
  - News? Education? Sport? Amusement…

$C=\{c_1,c_2,\ldots,c_m\}$, set of predefined categories;

$D=\{d_1,d_2,\ldots,d_n\}$, set of docements to be categorized

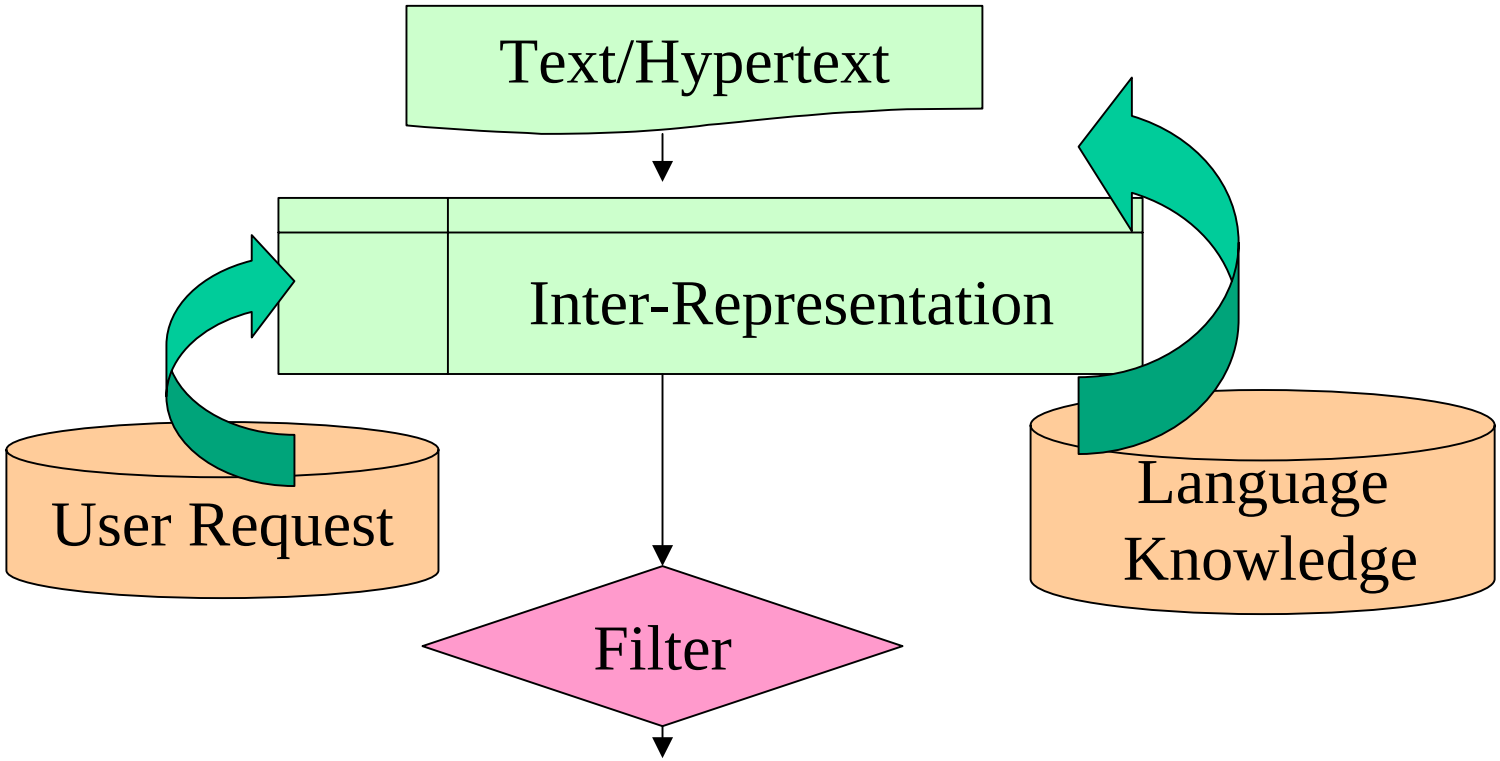
Classification:

$$f(d_i) = c \in C$$

How to represent each d_i & assign a class tag to it ?

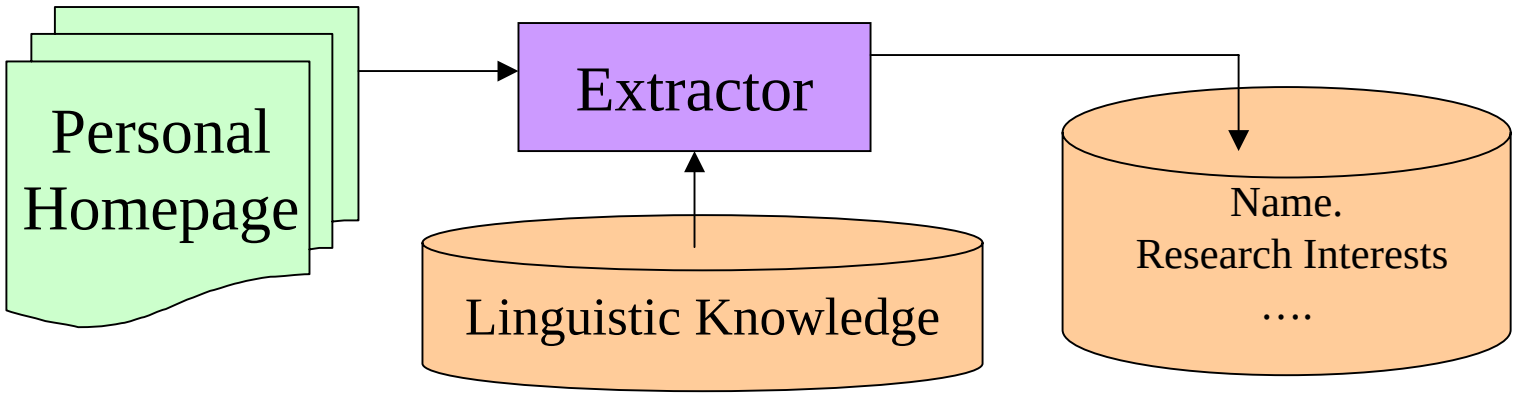
Applications(5)

- Text Information Filter on Internet



Applications(6)

- IE(Information Extraction)
  - Elicit interesting information from free or semi-structure text and represent it in structure data (data base)
  - Example: extract faculty information from personal homepage(CV) and generate data base:
    - Name . Research interests . Teaching . Position . Email



Applications(7)

- Topic detection & Tracking
  - Main Task:
    - In Time Order
    - Detecting new topic in news flow
    - Tracking the same topic in time order
  - Analysis of Topic / Content in Text

Applications(8)

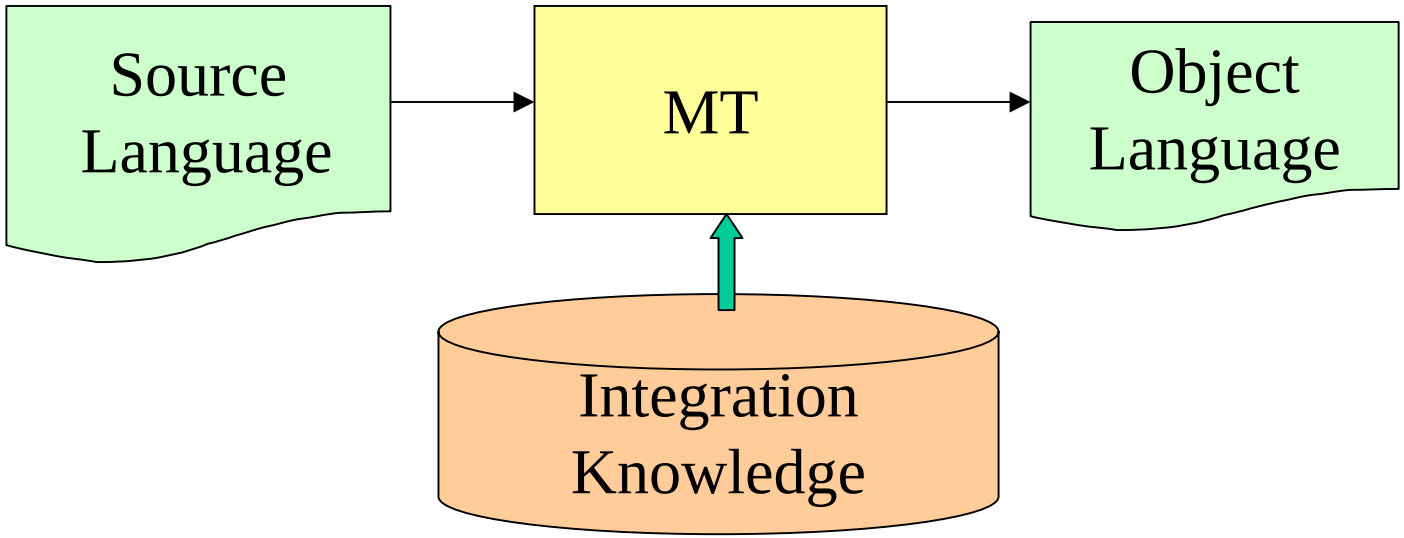
- Keywords Extraction and indexing
  - Word and phrase recognition
  - importance analysis
  - Filtering unimportant ones

Applications(9)

- Automatic Text Summarization
  - Single Text Summarization
  - Multi-Text Summarization
- Main Task:
  - Parsing / Analysis
  - Condensing(eliminate unimportant units)
  - Generation

Applications(10)

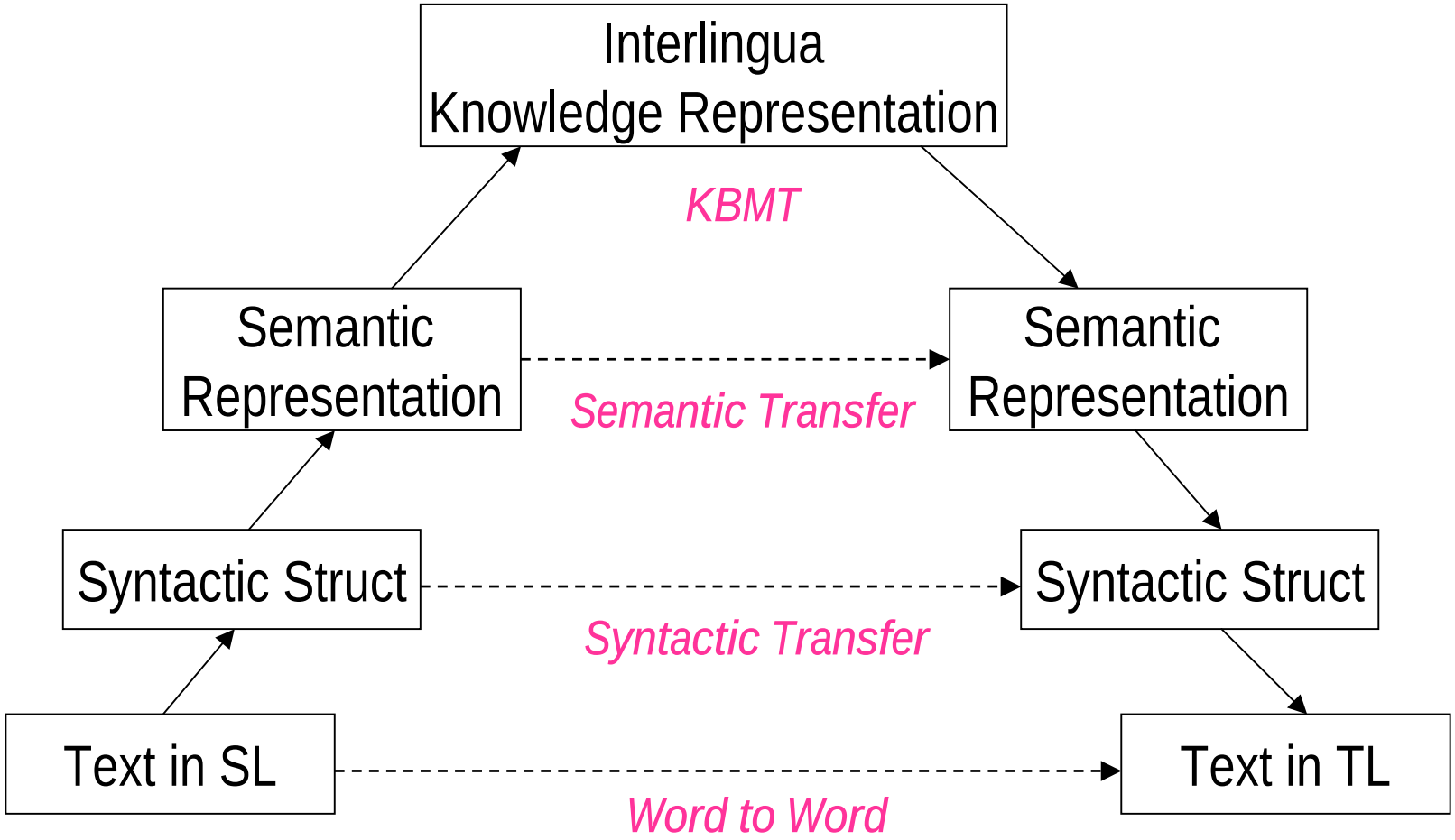
- Multi-Lingual System & Machine Translation
  - Cross-Lingual IR
  - MT needed



Outline

- What's Computational Linguistics?
- Why is Natural Language Processing (NLP) important (Applications)?
- **Difficulties**
- Brief History

General Model of MT(Analysis/Generation)



Why NLP difficult

- Could NL be fully formalized? How?
  (knowledge representation)
  - 主席团台上坐
  - 主席团坐台上
  - 台上坐主席团
  - 台上主席团坐
- Learning
- Common sense reasoning
- ambiguity resolution : do there exist any deterministic algorithms (Combinatorial explosion)

Ambiguities (1)

- Pinyins → Chinese Characters

  Input : Pinyin character without tone

  Output : Chinese Characters

  Eg. : yishishiweiyiju

  Step1: segmentation;

  yi shi shi wei yi ju.

  Step2: Search

意识	实时	示威	偎依	移居
仪式	事实	侍卫	位移	依据
遗失	实施	市委	唯一	一举
(28)	(28)	(9)	(9)	(6)

Ambiguities (2)

- Segmentation
  - Sentence segmentation:
    - Famous example: 下雨天留客天留我不留
  - English: Cannot $\rightarrow$ Can not
  - Chinese Word Segmentation
    - Problem 1: word & unknown word?

    洗澡(take bath) vs. 洗脸(wash face)/洗手(wash hands)/洗脚(wash foot)/洗头(wash head)…

    者 $\Rightarrow$ 作者，[以散文写作为主业] 者
    - Problem2: Too many Ambiguities

New Words: (Especially, new words on internet, Chat, BBS…)

美国的武器大家都知道，奇贵无比，比如巡航导弹要几百万美刀一个，一百个就是好几亿啊 。（中华网，2002年11月19日）

羽绒裤一般不穿在表面，因此不存在款式问题……很便宜，100人刀左右。如果你的人刀不是主要问题，推荐你购买×××面料羽绒服。

这是我穿过的最好的羽绒服……售价大约是2600港刀，有时打折，价格在2000港刀以下。（新浪网·新浪生活，2001年2月27日）……

- 庸医治病害死人：

  庸医 / 治 / 病害 / 死人

  庸医 / 治 / 病害 / 死 / 人

  庸医 / 治 / 病 / 害 / 死人

  庸医 / 治 / 病 / 害 / 死 / 人 (√)

  庸 / 医治 / 病害 / 死 / 人

  庸 / 医治 / 病 / 害 / 死 / 人

  庸 / 医 / 治 / 病害 / 死 / 人

  庸 / 医 / 治 / 病 / 害 / 死 / 人

  …… if “治病” is also a word ?

Ambiguities (3)

- Part of Speech

  more than one POS:

  book(N, V)

  She had read this book

  He wanted to book a ticket

把/q-p-v-n 这/r 篇/q 报道/v-n 编辑/v-n 一/m-c 下/f-q-v

Ambiguities (4)

- Formal Grammar
- General G=(N, T, S, P)
  - Type 0: $\alpha \rightarrow \beta$ Turing Machine
  - Type 1: $\alpha\ A\ \beta \rightarrow \alpha\ \gamma\ \beta$ Type1 $\subseteq$ Type0 and $|\gamma| \neq 0$
    Liner bounded Automata
  - Type 2: $A \rightarrow \beta$ Type2 $\subseteq$ Type1 and $A \in N$
    pushdown automata (****)
  - Type 3: $A \rightarrow a$ or $A \rightarrow aB$, $a \in Vt$, $A,B \in N$
    finite-state automata

Type 2= Context Free Grammar(CFG)

CFG = Phrase Structure Grammar(PSG)

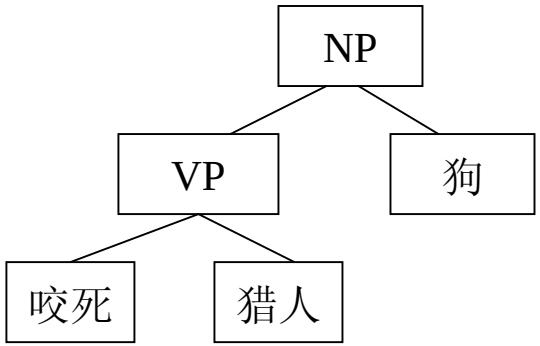
$S \rightarrow NP + VP$

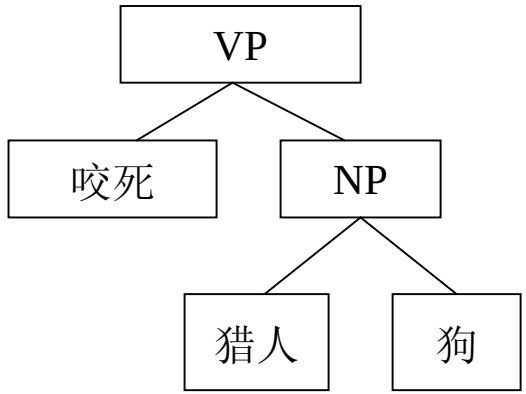
$NP \rightarrow adj + NP \mid NP + NP \mid n\ldots$

Structure ambiguity (Example)

I saw the boy with telescope
咬死猎人的狗 (two meanings)

$VP \rightarrow VP + NP \mid v$
$NP \rightarrow NP + NP \mid NP + 的 + NP \mid VP + 的 + NP \mid n$





Structure Ambiguity(Example)

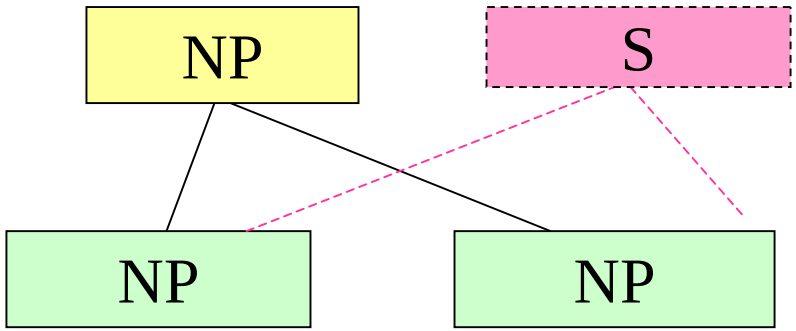
[[鲁迅$_{np}$ 先生$_{np}$] $_{np}$ 浙江人$_{np}$] $_{s}$

[[关羽][身高 八尺]]

CFG:

$S \rightarrow NP + NP$

$NP \rightarrow NP + NP \mid n$



Special Cases in Chinese

- 妈妈/n 叫/v 我/p 去/v 幼儿园/n 接/v 明明/n

  $S \rightarrow NP + VP$; $NP \rightarrow n|p$; $VP \rightarrow v$

  – 去/v 幼儿园/n

  – 接/v 小聪/n

  $VP \rightarrow VP + NP$

  – [去幼儿园]+[接明明]

  $VP \rightarrow VP + VP$

  [[[去幼儿园] [接]] [明明]]

  $VP \rightarrow VP + VP$

  $VP \rightarrow VP + NP$

Much Serious Ambiguity

- More than 10000 trees on some long sentences !
- extremum : (all combinations: enumeration)

  $Tree(n) = \Sigma\ Tree(k)Tree(n-k)$; $Tree(2)=Tree(1)=1$

  $Tree(9)=1430$; $Tree(10)=4862$; $Tree(11)=16796$; $Tree(12)=58786$

  $Tree(13)=208012$; $Tree(14)=742900$; $Tree(15)=2674440$
- Why?
  - Too many rules are needed for covering all kinds of sentences
  - there exist many conflicts among rules.

Ambiguities (5)

- What is Meaning?
  - Lexical semantics:

A story about "意思"： What's the meaning of "意思"？

人们以为他对她有"意思**(爱)**"，于是，建议他对她"意思意思（表示）"。他说，他没那种"意思"。她则反问，你们是什么"意思（用意）"。大伙中有的觉得很有"意思（趣味）"，有的则认为真没"意思"。

  - Sentence semantics:

I saw the boy with telescope

该来的没有来，不该走的走了！

- How many Meanings in Natural Language and how to represent?

Ambiguities (6)

- Anaphora / Coreference Resolution
  - Whether does an anaphor refer one Antecedent or not?
    - 他们正在加紧研制那种使一架飞机可以携带多个不同类型弹头的新型装置。
  - How to find Antecedent
    - 张三看到了李四，当时他在公共汽车上。由于车子开得太快，没看清和他在一起的那位女孩子是谁。
    - 德国政府明确表示，中国是德国亚洲政策的重点。正是在这一方针指导下，近年来中德关系迅速改善。

The [Programmer i] successfully combined [Prolog j] with C, but [he i] had combined [it j] with Pascal last time.

The [Programmer i] successfully combined Prolog with [C j], but [he i] had combined Pascal with [it j] last time.

Outline

- What's Computational Linguistics?
- Why is Natural Language Processing (NLP) important (Applications)?
- Difficulties

➢ **Brief History**

Machine Translation(1930s)

- 1933: France Georges Artsrouni & Russia Peter Trojanskij: Mechanical mutillingual dictionary
- 1946-1947:Andrew Booth and Warren Weaver, MT.
- 1950s: Yehoshua Bar-Hillel(MIT): report on MT and held 1st MT conference(MIT),in 1952. Where Leon Dostert(Georgetown Univ.) suggestions: Demonstration (to attract funding).
- 1st System developed by IBM & Georgetown (250 words & 6 grammar rules),1954 Russia — English
- 1st Journal: Mechanical Translation(1953-1970) by MIT, William Locke & Victor Yngve. 1st Doctoral Thesis in 1953(MIT), Anthony G. Oettinger —Russia Mechanical Dictionary.

A. Turing Test(1940s)

- Language as Turing Test (Person-Machine Dialogue)
- Three Participants:
  - One Computer
  - One person as Participant
  - One Interrogator(person)
- Different goal:
  - Interrogator: tell who is computer and who is person
  - Computer: fool interrogator
  - Person:help interrogator to reach his goal.

Theory (1950s)

- Formal language (Chomsky, Kleene, Backus).
  - Formal characterization of classes of grammar (context-free, regular)
  - Association with relevant automata
- Probability and information theory:language understanding as decoding through noisy channel (Shannon)
  - Use of information theoretic concepts like entropy to measure success of language models.

Symbolic vs. Stochastic

- Symbolic
  - Use of formal grammars as basis for natural language processing and learning systems. (Chomsky, Harris)
  - Use of logic and logic based programming for characterizing syntactic or semantic inference (Kaplan, Kay, Pereira)
  - First toy natural language understanding and generation systems (Woods, Minsky, Schank, Winograd, Colmerauer)
  - Discourse Processing: Role of Intention, Focus (Grosz, Sidner, Hobbs)
- Stochastic Modeling
  - Probabilistic methods for early speech recognition, OCR (Bledsoe and Browning, Jelinek, Black, Mercer)

The ALPAC report(1966)

- Automatic Language Processing Advisory Committee(ALPAC) set up by National Science Foundation(1964, USA)
- Reasons of report:
  - The current state of MT: Semantic Barrier
  - Bar-Hillel's cristicism: real world knowledge is needed (Example: the box was in the pen)
- The ALPAC reports (1966) :
  - "there is no immediate or predictable prospects of useful machine translation"—— Ends funding MT.
  - Only support fundamental research in CL

Recovery(1970s)

- 1971 W. Woods: Lunar, IR(ATN)
- 1972: T. Winograd SHRDLU (Lisp). Existence proof of AI & NLP. (Q&A):

  Q: Which cube is sitting on the table?

  A: The large green one which supports the red pyramid

  Q: ……
- 1973: Schank: Concept Dependency (CD Theory)
  - MARGIE 1975: (Meaning, Analysis, Response Generation and Inference on English) on CD: NLU
  - SAM(Script Applier Mechanism) on CD

AI & Knowledge Base (1980s)

- Separation of Processing (Parsing) from descriptions of Linguistic Knowledge
- Representations of meanings (CD) & Lenat's CYC
- MT in limit domains (Meteo)
- Syntax-Semantics and Lexical semantics Theory: GPSG(Gazdar), GB(N. Chomsky) , FUG(Functional Unification Grammar, M. Kay)

Empiricism(1990s)

- IBM's P.Brown(1988-2nd TMI, 1990-CL,1993-CL): Statistical Approach on MT;
- Statistical and corpus-based gradually become dominate;
- Emphasis on very large corpora and real text;
- Emphasis on Machine Learning and automated knowledge acquisition;
- Speech recognition becomes usable;
- Large-Scale Evaluation.

In the future(maybe)

- Emphasis on integration of techniques.
- Combination of Rational and Empirical methods
- More integrations of NLP components into many application systems
- Emphasis on knowledge and meaning representations

Notice

- Class Time: 18:00-21:00 Tuesday (**from next time**)

Questions

- Someone thinks NLP as classification Problem, What do you think of that?