[Day-1] Introduction to Statistical Natural Language Processing

(Part I: Introduction I)

Keh-Yih Su and Jing-Shin Chang
kysu@bdc.com.tw, shin@nlp.csie.ncnu.edu.tw

(2002/8/17)

Behavior Design Corporation (BDC)
No. 5, 2F, Industrial East Road IV, Science-Based Industrial Park
Hsinchu, Taiwan 30077, R.O.C.

Department of Computer Science and Information Engineering
National Chi-Nan University
Puli, Nantou, Taiwan 545, R.O.C.

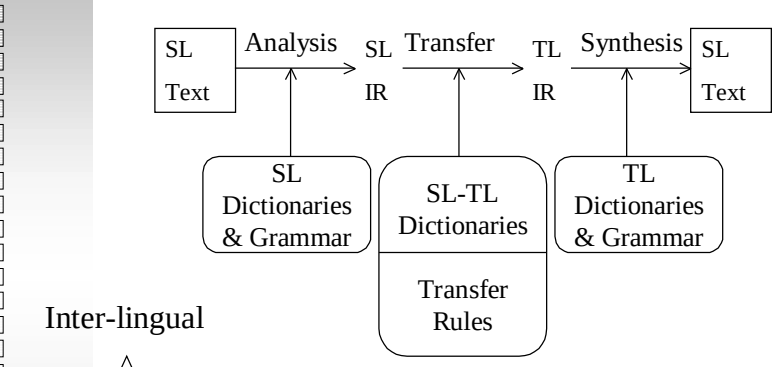
Day-1: Introduction to Statistical Natural Language Processing (mainly on Supervised Learning)

- **Part I: Introduction (1)**
  - **Problems and Characteristics of Natural Language Processing**
- Part II: Introduction (2)
  - What, When and Why Statistical Approach
- Part III: Basic Concepts and Background
  - Feature Space, Probability, Estimator, Stochastic Process, Data Set Classification, and Performance Measure
- Part IV: Typical Applications
  - Word Segmentation, Tagging, Selecting Parse Tree, Aligning Bilingual Corpus
- Part V: Techniques for Improving Performance
  - Smoothing, Class-Based Model, Adaptive Learning, Tips for Checking
- Part VI: Advanced Topics: SVM, ME
  - Support Vector Machine, Maximum Entropy Models
- Appendix: Related Techniques
  - Parameter Estimation, Fractional Factorial Experiment Design, Decision Tree

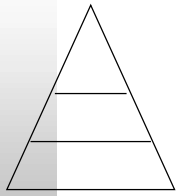
Part I: Introduction

- A Natural Language Processing Example: Machine Translation
  - Process Flow
- Main Problems in NLP
  - Ambiguity & Ill-formedness
- Characteristics of NLP
  - Non-deterministic, Requires huge and messy knowledge
- Main Tasks in NLP
  - Tasks required in NLP, Bottlenecks
- Knowledge Acquisition in NLP
  - How it is related to the Knowledge Representation Form

An Example: Machine Translation (Transfer Approach)



Inter-lingual



IR: Intermediate Representation Form

- Analysis is target independent and generation is source independent

Example: Miss Smith put two books on this dining table.

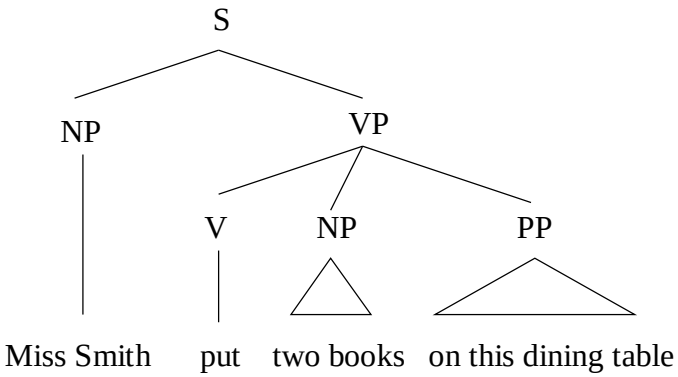
- Analysis
  - (1) Morphological Analysis

Miss
Smith
put (+ed)
two
book+s
on
this
dining table.

Example: Miss Smith put two books on this dining table.

Analysis (Cont.)

(2) Syntactic Analysis



Example: Miss Smith put two books on this dining table.

- Transfer
  - (1) Lexical Transfer

Miss	⇒	小姐
Smith	⇒	史密斯
put (+ed)	⇒	放
two	⇒	兩
book+s	⇒	書
on	⇒	在…上面
this	⇒	這
dining table	⇒	餐桌

  - (2) Phrasal Transfer

    小姐史密斯放兩書在上面這餐桌
    史密斯小姐放兩書在這餐桌上面

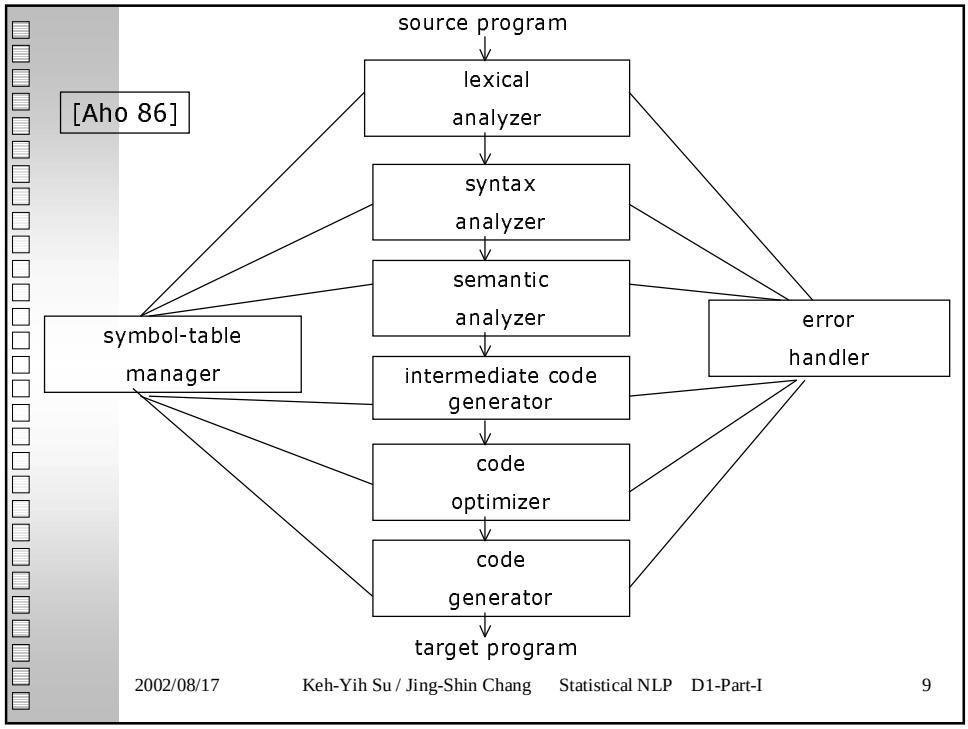
Example: Miss Smith put two books on this dining table.

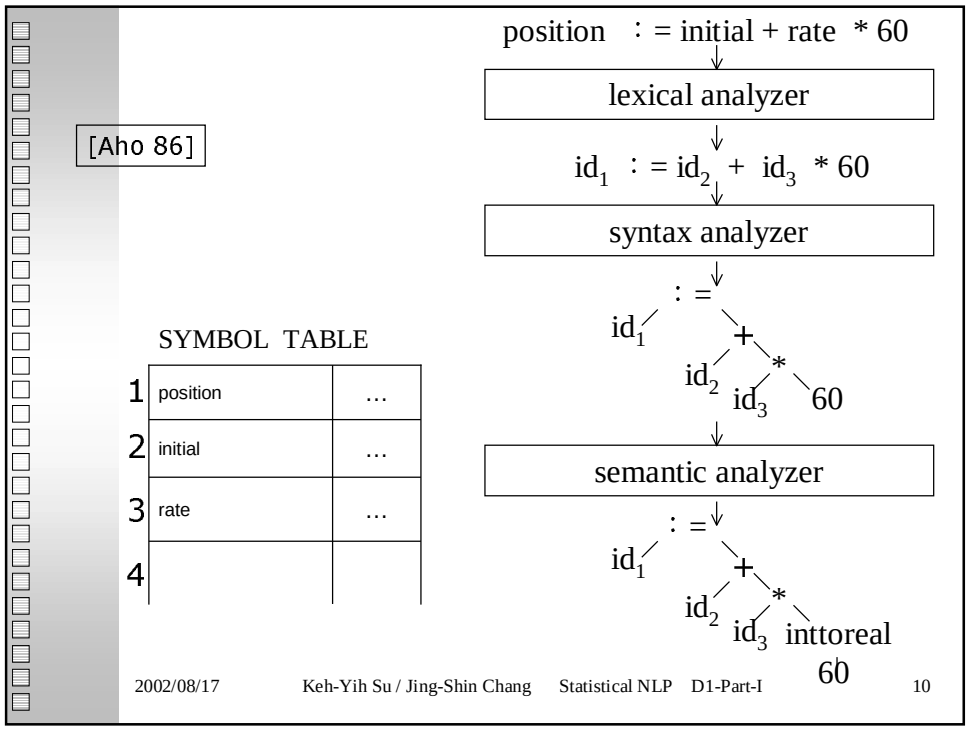
- Generation

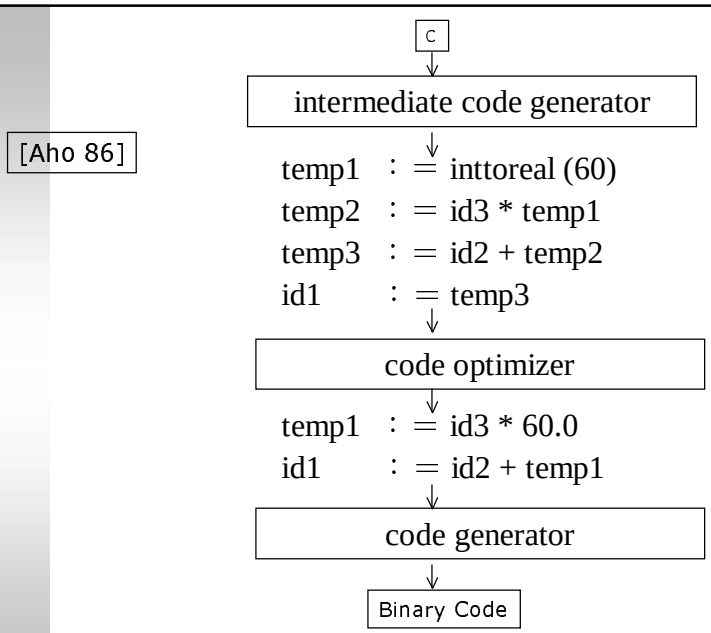
  史密斯小姐放兩書在這餐桌上面

  史密斯小姐(把)兩(本)書放在這(張)餐桌上面

中文翻譯：
⇒ 史密斯小姐把兩本書放在這張餐桌上面


[Aho 86]
source program
lexical analyzer
syntax analyzer
semantic analyzer
intermediate code generator
code optimizer
code generator
target program
symbol-table manager
error handler
2002/08/17 Keh-Yih Su / Jing-Shin Chang Statistical NLP D1-Part-I 9


[Aho 86]
position : = initial + rate * 60
lexical analyzer
id1 : = id2 + id3 * 60
syntax analyzer
: =
id1
+
id2
*
id3
60
semantic analyzer
: =
id1
+
id2
*
id3
inttoreal
60
SYMBOL TABLE
1 position …
2 initial …
3 rate …
4
2002/08/17 Keh-Yih Su / Jing-Shin Chang Statistical NLP D1-Part-I 10



Detailed Steps (1): Analysis

- Text Pre-processing (separating texts from tags)
  - Clean up garbage patterns (usually introduced during file conversion)
  - Recover sentences and words (e.g., <B>C</B> omputer)
  - Separate Processing-Regions from Non-Processing-Regions (e.g., File-Header-Sections, Equations, etc.)
  - Extract and mark strings that need special treatment (e.g., Topics, Keywords, etc.)
  - Identify and convert markup tags into internal tags (de-markup; however, markup tags also provide information)
- Discourse and Sentence Segmentation
  - Divide text into various primary processing units (e.g., sentences)
  - Discourse: Cue Phrases
  - Sentence: mainly classify the type of “Period” and “Carriage Return” in English (“sentence stops” vs. “abbreviations/titles”)

Detailed Steps (2): Analysis (Cont.)

- Stemming
  - English: perform morphological analysis (e.g., -ed, -ing, -s, -ly, re-, pre-, etc.) and Identify root form (e.g., got <get>, lay <lie/lay>, etc.)
  - Chinese: mainly detect suffix lexemes (e.g., 孩子們, 學生們, etc.)
  - Text normalization: Capitalization, Hyphenation, …
- Tokenization
  - English: mainly identify split-idiom (e.g., turn NP on) and compound
  - Chinese: Word Segmentation (e.g., [土地] [公有] [政策])
  - Regular Expression: numerical strings/expressions (e.g., twenty millions), date, … (each being associated with a specific type)
- Tagging
  - Assign Part-of-Speech (e.g., n, v, adj, adv, etc.)
  - Associated forms are basically independent of languages starting from this step

Detailed Steps (3): Analysis (Cont.)

- Parsing
  - Decide suitable syntactic relationship (e.g., PP-Attachment)
- Decide Word-Sense
  - Decide appropriate lexicon-sense (e.g., River-Bank, Money-Bank, etc.)
- Assign Case-Label
  - Decide suitable semantic relationship (e.g., Patient, Agent, etc.)
- Anaphora and Antecedent Resolution
  - Pronoun reference (e.g., “he” refers to “the president”)

Detailed Steps (4): Analysis (Cont.)

- Decide Discourse Structure
  - Decide suitable discourse segments relationship (e.g., Evidence, Concession, Justification, etc. [Marcu 2000].)
- Convert into Logical Form (Optional)
  - Co-reference resolution (e.g., “president” refers to “Bill Clinton”), scope resolution (e.g., negation), Temporal Resolution (e.g., today, last Friday), Spatial Resolution (e.g., here, next), etc.
  - Identify roles of Named-Entities (Person, Location, Organization), and determine IS-A (also Part-of) relationship, etc.
  - Mainly used in inference related applications (e.g., Q&A, etc.)

Detailed Steps (5): Transfer

- Decide suitable Target Discourse Structure
  - For example: Evidence, Concession, Justification, etc. [Marcu 2000].
- Decide suitable Target Lexicon Senses
  - Sense Mapping may not be one-to-one (sense resolution might be different in different languages, e.g. “snow” has more senses in Eskimo)
  - Sense-Token Mapping may not be one-to-one (lexicon representation power might be different in different languages, e.g., “DINK”, “貶”, etc). It could be 2-1, 1-2, etc.
- Decide suitable Target Sentence Structure
  - For example: verb nominalization, constitute promotion and demotion (usually occurs when Sense-Token-Mapping is not 1-1)
- Decide appropriate Target Case
  - Case Label might change after the structure has been modified
  - (Example) verb nominalization: “… that you (AGENT) invite me” ⇔ “… your (POSS) invitation”

Detailed Steps (6): Generation

- Adopt suitable Sentence Syntactic Pattern
  - Depend on Style (which is the distributions of lexicon selection and syntactic patterns adopted)
- Adopt suitable Target Lexicon
  - Select from Synonym Set (depend on style)
- Add “de” (Chinese), comma, tense, measure (Chinese), etc.
  - Morphological generation is required for target-specific tokens
- Text Post-processing
  - Final string substitution (replace those markers of special strings)
  - Extract and export associated information (e.g., Glossary, Index, etc.)
  - Restore customer’s markup tags (re-markup) for saving typesetting work

Common Phenomena in NLP

- An early step might need information from later steps
  - For example, identifying split-idiom in the Tokenization step needs to verify a specified constituent (e.g., turn NP on)
  - One way to handle that is to adopt a Black-Board approach; however, it is not efficient (ref. Verbmobile report [Wahlster 00]).
- Output may not be unique
  - Zero, when a rule-based approach encounters ill-formed input
  - Usually several candidates are possible (even under Unification Grammar Formalism)

Translation is hard (even by human)

- Sometimes the "same" phrase means different things in different geographical areas:
  - Example: Knock somebody up (Margaret King)
    - Wake them in the morning
    - Get them pregnant
- Sometimes contradictory phrases might mean the same thing in different geographical areas:
  - Example: Valid Ticket and Invalid Ticket (Martin Kay)

Machine Translation is harder

- The computer system has to make choices even when the human isn't (normally) aware that a choice exists.
  - Example from Margaret King:
    - The farmer's wife sold the cow because she needed money.
    - The farmer's wife sold the cow because she wasn't giving enough milk.
  - Another example:
    - The mother with *babies* under four….
    - The *mother* with babies under forty…..

Main Problems in NLP (1)

- Ambiguity
  - Sentence Segmentation:
    - An English “Period” might not be the Sentence-Delimiter (e.g., Eq. 3)
    - Should Chinese “Comma” be regarded as the Sentence-Delimiter?
    - order: several candidates per sentence
  - Tokenization:
    - English Split-Idiom and Compound-Noun matching
    - Chinese Word Segmentation
    - order: several to tens of candidates per sentence
  - Lexical:
    - "current": noun vs. adjective
    - order: hundreds of candidates per sentence

Main Problems in NLP (2)

- Ambiguity (Cont.)
  - Syntactic:
    - "[saw [the boy] [in the park] [with a telescope]]"
    - "[saw [the boy in the park [with a telescope]]]"
    - order: several hundreds to thousands
  - ⇔ Analogy in artificial language: dangling-else problem [Aho 86]
    - " [If (...) then [if (...) then (...) else (...)]] "
    - " [If (...) then [if (...) then (...)] else (...)] “
    - Choose the nearest “THEN”, if not particularly specified
  - Semantic:
    - Lexicon-Sense: Bank (Money vs. River)
    - Case: Agent vs. Patient
      --- "[the police] were ordered [to stop drinking] by midnight”
  - Pragmatic:
    - Example: “你好厲害哦”

Main Problems in NLP (3)

- Ill-Formedness
  - **Possible Forms:**
    - Unknown Words (not found in dictionaries)
      - Missing in lexicon-database: Vocabulary size, Proper Noun, Typing error, Breeding words (e.g., Singlish: Singapore English), new technical terms (e.g., bioinformatics)
    - Known Words, but Missing desired information (e.g., part-of-speech)
      - New usage (e.g., “Please xerox a copy to me.”)
      - Known usage, but Not listed in the dictionary: e.g., “that” should be “whn” in “*You may want the extra protection **that** a power-conditioner can give you*”; however, it is missing in the dictionary (which only have dem, pron, and comp).

Main Problems in NLP (4)

- Ill-Formedness (Cont.)
  - **Possible Forms (Cont.):**
    - Un-grammatical sentences (cannot be parsed by the given grammar)
      - Example: "Which one?" …
    - Violate semantic constrain
      - Example: My car drinks gasoline like water. (subject-verb agreement)
    - Violate ontology
      - Example: There is a plastic bird on the desk. Can this bird fly? (Sowa 2000).

Main Problems in NLP (5)

- Ill-Formedness (cont.)
  - **Possible Sources:**
    - Source Contamination (careless preparation)
      - Typing Error: mis-spelling, missing words, extra words, etc.
      - Bad writing: missing verbs, etc.
      - Garbage introduced by file transmission/conversion
      - Gargled by extra tags: typesetting formats, XML (RTF, SGML) tags, etc
    - Languages continuously evolve
      - New lexicons
      - New usage patterns

Main Problems in NLP (6)

- Ill-Formedness (cont.)
  - **Possible Sources (Cont.):**
    - Linguistically uninterested/unresolved problems:
      - Real language is dirty: e.g., different ways to express a Date (e.g., “Aug. 17, 2002”, “17, Aug 2002”, “17-Aug-02”, “17/08/02”, etc.)
      - Colloquium usage is loose: e.g., "Which one?" …
    - Designing Tradeoff
      - Number of Ambiguities v.s. grammar coverage rate
    - Implementation Limitation
      - Legal candidates are pruned out by limited searching Beam-Width in early stages (known as “searching errors”)

Characteristics of NLP (1)

- Knowledge required to handle the above-mentioned problems is huge, messy & fine-grained
  - NLP is a very complicated process
  - Real language is dirty (not regular)
  - Construct Knowledge by hands is very expensive and time-consuming
- Interpretation is dynamic, not static
  - Interpretation highly depends on its context (Knowledge Soup [Sowa 00]: "A bird can fly" might not be true.)
  - Ontology is difficult to build, and many situations cannot be covered
- Most knowledge required in NLP is inductive, not deductive
  - Language ----> Linguistics
  - Linguistics - x-> Language (e.g., Esperanto)
- Even human do not give the same answer
  - Human is competent in abstract language modeling, but awkward in consistently dealing with large and fine-grained knowledge
  - Performance Upper-bound Exists (a Golden book is not truly golden)

Characteristics of NLP (2)

- NLP is a non-deterministic process
  - Natural language is non-deterministic in nature. (not clearly expressed, or intentionally making jokes)
  - Unavoidable in modular pipeline control flow system design. (Early stages lack the required knowledge to be generated in later modules.)
- Ambiguity resolution strategies often conflict with the system coverage rate.
  - More constraints for less ambiguity => increase ill-formedness
  - Restrict possible word-senses by domain dictionary => uncovered senses
- Domain Dictionary is not enough
  - A Domain Dictionary implicitly reduces the degree of complexity by restricting the number of senses allowed; however, the sentence including rate is a product of the including rate of each lexicon (which is not 100%)
  - Sense is often implied by the contextual (dynamic) information
  - Domain Knowledge is required (even human translators/writers are classified by their expertise, not just giving them different domain dictionaries)

Main Tasks for Building NLP Systems

- Knowledge Representation
  - How to organize and describe intra-linguistic, inter-linguistic, and extra-linguistic knowledge.
- Knowledge Control Strategies
  - How to efficiently use knowledge for ambiguity resolution and ill-formedness recovery
- Knowledge Integration
  - How to jointly consider the information from different stages (e.g., syntactic score, semantic score, etc.): Natural language contains redundant information in different levels, they will enhance each other if they can be jointly considered
  - How to jointly consider knowledge from various sources effectively (e.g., WordNet, Hownet, various dictionaries, translation-memory, etc.)
- Knowledge Acquisition
  - How to systematically and cost-effectively set up knowledge bases
  - How to maintain the consistency of knowledge base

Main Bottleneck: Knowledge Acquisition

- Knowledge Acquisition is usually the bottleneck
  - Language usage is complex (not governed by any simple and elegant model), and dynamic (which changes with different groups, locations, and time)
  - Required knowledge is huge, messy and fine-grained
  - Inducing rules by human is usually very expensive, and time-consuming
  - Consistency is difficult to maintain when the system scales up
- Seesaw phenomenon is generally observed
  - Traditional rule-based approaches are very hard to ensure global improvement, even if it is possible. (Human can only jointly consider 5-9 objects at the same time.)
- Need cheap and systematic ways to acquire knowledge
  - Complex problems need a large amount of knowledge, which is very difficult and expensive to build and maintain
  - Machine Learning seems to be the only way to go

Knowledge Acquisition in NLP (1)

- Knowledge can be represented in different forms
  - Knowledge can be represented either explicitly (such as rules) or implicitly (such as parameters).
    - Example 1: IF [C_{i-1} is *Det*], then [C_i cannot be a *Verb*]
    - Example 2: $P(C_i = Verb \mid C_{i-1} = Det) = 0$
    - Example 3: weighting coefficients in neural-network
  - We usually classify various approaches by their associated Knowledge Representation Form (e.g., Rule-Based, Example-Based, etc.)
- The Task of Knowledge Acquisition is closely coupled with its Knowledge Representation Form
  - Changing the Knowledge Representation Form usually also changes the way to acquire knowledge (Rules <= human, Parameters <= computer)

Knowledge Acquisition in NLP (2)

- We should consider the Knowledge Representation Form from the Knowledge Acquisition point of view
  - Since Knowledge Acquisition is the bottleneck, we should consider it first
  - First select the suitable knowledge acquisition mode, then decide the corresponding appropriate knowledge representation form
- As the required knowledge is huge and messy, machine learning (not acquired by human) is preferred
  - What kind of knowledge is suitable for machine learning?
    - With complex interaction between different features (not easily handled by human)
    - Uniform, large quantity, can be easily derived from those observable data
  - Parametric form is most suitable for machine learning
    - A collection of a large number of simple, but *adjustable*, units can also demonstrates smart behavior. (for example, neurons and IBM Deep Blue)

Knowledge Acquisition in NLP (3)

- Integrated Approach is better for NLP applications (also classified as “hybrid” approaches by some researchers)
  - Motivation:
    - Learning abstract forms (e.g., model) has not yet demonstrated its success in machine learning
    - Final performance is judged by how closely the result matches the human preference (and human knows how the decision is made); usually, linguists have no problem to find out the possible features; they just have difficulty to handle the complex dependency between different features
  - Approaches:
    - Human select suitable features, then derive the appropriate parametric language model which possesses a lot of parameters
    - Parameter values are then acquired via machine learning from corpora
  - Hybrid approaches are the most promising in the next decade (at least)