

**INTRODUCTION TO  
CORPUS-BASED  
STATISTICS-ORIENTED  
(CBSO)  
TECHNIQUES  
(PART I: INTRODUCTION)**

Keh-Yih Su  
Tung-Hui Chiang  
Jing-Shin Chang

Department of Electrical Engineering  
National Tsing-Hua University  
Hsinchu, TAIWAN 30043, R.O.C.  
email: [kysu@bdc.com.tw](mailto:kysu@bdc.com.tw)

# TASKS FOR LARGE SCALE NLP

---

- Knowledge Representation
  - How to organize and describe linguistic knowledge
- Knowledge Control Strategies
  - How to use knowledge for
    - efficient analysis
    - ambiguity resolution
    - ill-formedness recovery
- Knowledge Acquisition
  - How to systematically and cost-effectively set up knowledge bases, and
  - maintain knowledge base consistency
- Knowledge Integration
  - How to jointly consider various knowledge sources effectively

## Some Observations in NLP Research

- Knowledge acquisition is the major engineering bottleneck for a large scale NLP/MT system.
  - Required knowledge is usually huge, messy and fine-grained.
  - Knowledge acquisition is a very expensive and time-consuming task.
  
- Inducing consistent knowledge with corpus-based approaches is the promising approach for NLP applications:
  - Most knowledge required in NLP is inductive, not deductive.
    - Language → Linguistics
    - Linguistics  $\not\rightarrow$  Language
  - Inductive knowledge can be acquired easily with statistical methods.
  
- Need a cooperative way between human and machine:
  - Human is competent in abstract language modeling, but awkward in dealing with large and fine-grained knowledge.
  - It is not easy for people to maintain consistency of large rule bases and guarantee global improvement by enlarging rule bases.
  - Computers are good at processing large corpora.

# Knowledge Acquisition

- Induction by human:** Rule-based approaches
- Automatic Induction:**
  - Statistical Modeling
    - Pure statistical approach
    - CBSO approach
  - Symbolic Learning
  - Connectionist

## RULE-BASED APPROACHES

---

- Have a strict sense of well-formedness in mind
- Impose linguistic constraints on syntactic or semantic constructs to satisfy well-formedness
- Mostly based on linguistic knowledge that is linguistically interesting or *ad hoc* heuristics
  
- Example: (lexical disambiguation)
  - Heuristics: a “determiner” cannot be followed by a “verb”
  - Rule: if  $C_{i-1} = \text{“det”}$  then  $C_i \neq \text{“verb”}$

## **Rule-Based Systems:**

### **Advantages and Disadvantages**

Advantages:

- Easy to incorporate existing linguistic knowledge.
- Have better generalization (deduction) capacity.

Disadvantages:

- Hard to handle uncertainty (lack of objective preference measure).
- Hard to deal with complex and irregular knowledge (exception).
- Hard to maintain consistency among different persons at different time.
- Knowledge acquisition is costly and time consuming.
- Lack of systematic and automatic ways for learning rules in large-scale applications.
- Not easy to obtain high coverage (completeness) even for a given domain.
- Not easy to avoid redundancy.

# PURELY STATISTICAL APPROACHES

- No strict sense of well-formedness in mind
- Assume language generation to be a Markov chain
- Treat translation as a decoding process, no traditional linguistic knowledge is used in modeling
  
- Example I: (lexical disambiguation)
  - To discover the fact that  $p(v|det) = 0$ , one need to find  $p(y|x) = 0$  for all  $x \in \text{“det”}$  and all  $y \in \text{“verb”}$

x	y	$p(y x)$
the	ate	0
a	took	0
an	walks	0
	buy	0

- Have a large parameter space: for example
  - 100,000 words + Tri-gram word model  
=> number of parameters is  $10^5 \cdot 10^5 \cdot 10^5 = 10^{15}$

□ Example II: [Brown et. al, 89, 90]

- find possible translation lexicon for all words in source language
- English-to-French translation
  - cut a sentence into words
  - find corresponding French words
  - place the words in a bag
  - try to recover the sentence by rearranging all possible sequences and using tri-gram source model, fertility probability and distortion probability to find the preferred one
- assume free word order:
  - “John likes Mary.” is as good as
  - “Mary likes John.”



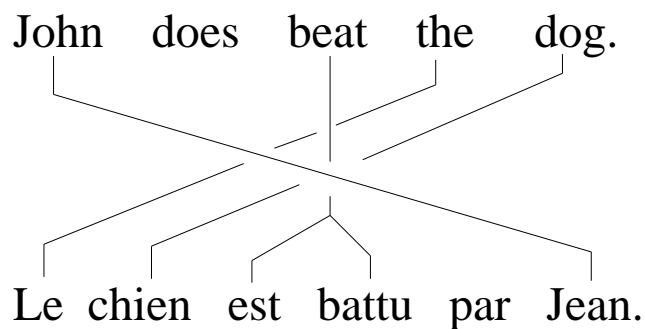
□ Example II: [Brown et. al, 89, 90] (cont'd)

- Regard "Translation Mechanism" as a Decoding Process

—Find  $\max P(t|s)$ , where (s,t) is a source language and target language sentence pair.

- Translation Model:

$$\begin{aligned} P(s | t) = & P(f_1 = 1 | John) \cdot P(Jean | John) \cdot \\ & P(f_2 = 0 | does) \cdot \\ & P(f_3 = 2 | beat) \cdot P(est | beat) \cdot P(battu | beat) \cdot \\ & P(f_4 = 1 | the) \cdot P(le | the) \cdot \\ & P(f_5 = 1 | dog) \cdot P(chien | dog) \cdot \\ & P(f_6 = 1 | null) \cdot P(par | null) \end{aligned}$$



## **Purely Statistical Approaches: Advantages and Disadvantages**

### Advantages:

- Do not need to establish linguistic models.
- Computation is easy.

### Disadvantages:

- The parameter space is too large.
- A very large training corpus is required due to the large number of parameters.
- Only local dependency can be handled.

—Noam Chomsky's Debate (1956):

No finite-state Markov Process can serve as an English grammar.

Example:

—If S1, then S2.

—Either S3, or S4.

—The man who said that S5, is arriving today.

- Cannot make use of prior existing knowledge induced by linguists

# WHAT IS A CBSO APPROACH?

---

- Use Corpus as the Main Information Source
  - Only a Few Well-Justified Linguistic Rules are Used
    - mechanical induction processes are performed by computers
  - Most Knowledge is Obtained from Corpora
    - Except modeling
  
- Language model is proposed by human which is, in general, more complex than the simple Markov chain.
  
- Use Statistical Method to Acquire Knowledge
  - Knowledge is Interpreted in Statistical Sense
  - Parameters are Automatically Learned from Corpus

# Characteristics of the CBSO Approaches

---

- CBSO = Statistical Language Model + Language Parameters
- No strict sense of well-formedness in mind
- Assume that a language generation process has stochastic behavior when a proper language model is used
- Make use of well-justified linguistic knowledge to construct high-level stochastic language models
- A promising paradigm for integrating linguistic and statistical knowledge

□ Example: (lexical disambiguation)

- use the known lexical knowledge to drastically reduce the parameter space and the required training data

$$\left. \begin{array}{cccc} w_1 & w_2 & w_3 & w_4 & \dots \\ c_1 & c_2 & c_3 & c_4 & \dots \\ \\ w_1 & w_2 & w_3 & w_4 & \dots \\ c_1 & c_2 & c_3 & c_4 & \dots \end{array} \right\} p(\text{verb}|\text{det}) = 0$$

□ Have a moderate and manageable parameter space:

- e.g., 100 parts of speech + Tri-gram model in part of speech  
=> number of parameters is  $10^2 \cdot 10^2 \cdot 10^2 = 10^6$

□ Why linguistic models are needed?

- Linguistic models group words into equivalent classes (e.g., parts of speech, phrase elements, words with the same semantic features), and hence reduce the number of required parameters drastically.

# CBSO: Advantages and Disadvantages

---

## Disadvantages:

- It is sometimes hard to couple with some existing linguistic theories.
- Generalization (deduction) capability is poor with small database.
- Large corpora are required to get reliable statistics.

## Advantages:

- Uncertainty or preference is interpreted objectively and consistently.
- Consistency can be easily maintained even in large scale systems.
- Well-established statistical theories and techniques are available.
- (Semi-)automatic training is possible with least human intervention.
- Remove the burden of rule induction from linguists to machine.
- Long distance dependency is manageable:
  - more contextual information can be used;
  - syntactic and semantic information could be consulted.

- It is suitable for feedback training because of the parameterized modeling.
  - Sparse data problem is less severe:
    - intermediate forms are introduced to reduce the parameter space;
    - parameter space is small with respect to purely statistical approaches.
  - It is usually robust to the ill-formed cases.
- The CBSO approaches manipulate stochastic behavior on top of nonterminal symbols. Hence, long distance dependency can be easily handled, and the size of the parameter space can be limited to a manageable scale.

# **CBSO DESIGN PHILOSOPHY**

---

## **Cooperative approach:**

- Admit the facts that
  - general linguistic knowledge has better generalization for unseen domain
  - uncertain knowledge can be more objectively quantified by statistic models
  - large and fine-grained language parameters can be acquired more cost-effectively by machine
  - human is competent in language modeling while machines are suitable for massive data processing
- Principles of cooperative approach
  - take advantages of well-recognized linguistic phenomena
  - setup probabilistic language model by humans
  - acquire language parameters from corpora (semi-) automatically

## **Systematic approach:**

- knowledge acquisition & integration should be done with systematic approach
- least human intervention should be involved via parameter learning



# WHY CBSO APPROACHES BECOME POPULAR?

---

- An appropriate integration of symbolic and statistical approaches can enhance both of the individual processing [Meteer 94].
  - “Traditionally, symbolic and statistical approaches have been kept separate and applied to very different problems.”
  - Symbolic techniques apply best:
    - “when we have well-developed knowledge of the language or the domain.”
    - “when the application of a theory or study of selected examples can help leverage and extend our knowledge.”
  - Statistical approaches apply best:
    - when the randomness of existing linguistic theory is observed.
    - “when the results of decisions can be represented in a statistical model.”
    - “when we have sufficient training data to accurately estimate the parameters of the model.”

- The demand for large scale practical NLP applications is increasing. Thus, corpus-based research is required to reflect the real usage of the languages and provide the real world view for daily encountered linguistic problems.
  - Wide coverage toward unrestricted domain is required in real NLP applications; practically encountered phenomena must be resolved even if they are not theoretically interesting.
  - Robustness: NLP systems must show good performance both in laboratory and real environment (avoid over-tuning).
  
- Justification of various theoretical frameworks against a large and possibly common text corpus is required to evaluate all frameworks.
  
- Human intervention is costly and time consuming; corpus processing techniques are required to reduce the cost of knowledge acquisition

# A COMPARATIVE REVIEW

---

Rule-Based Approaches

- Heuristic Rule: a “determiner” cannot be followed by a “verb”

Purely Statistical Approach

$x$	$y$	$p(y x)$
the	ate	0
a	took	0
an	walks	0
	buy	0

- Conclusion: a word in  $x$  cannot be followed by a word in  $y$

CBSO Approach

$$\left. \begin{array}{l}
 w_1 \ w_2 \ w_3 \ w_4 \ \dots \\
 c_1 \ c_2 \ c_3 \ c_4 \ \dots \\
 \\
 w_1 \ w_2 \ w_3 \ w_4 \ \dots \\
 c_1 \ c_2 \ c_3 \ c_4 \ \dots
 \end{array} \right\} p(\text{verb}|\text{det}) = 0$$

# WHEN TO ADOPT A CBSO APPROACH?

---

- When we should adopt rule-based approach
  - Not easy to establish a large-scale database
  - The size of rule-base needed is not large (Phenomena can be governed by a small number of rules, which have been well justified.)
  - Rules with large coverage exist
  - Extensional knowledge is important to the system
  
- When should we adopt corpus based statistics oriented approach
  - Establishing a large-scale database is affordable
  - Knowledge needed to solve the problem is huge and intricate, not easy to acquire by human.
  - Intensional knowledge is enough for the system
  - A good model or formulation can be found