Probabilistic Models of Natural Language Processing

Empirical Validity and Technological Viability

Khalil Sima'an

Institute For Logic, Language and Computation

Universiteit van Amsterdam

FIRST COLOGNET-ELSNET SYMPOSIUM

Trento, Italy, 3-4 August 2002

Speech and Language Technology

What is common for these applications ?

- Document Retrieval, Document Categorization,$\cdots$
- Question Answering, Information Extraction,$\cdots$
- Text Summarizing, Dictation systems, Machine Translation,$\cdots$
- Speech Understanding, Speech-based Dialogue systems,$\cdots$

  $\vdots$

A Model of Natural Language Processing (N L P) ?

Common Wisdom – Current Experience

Practice: *advanced NLP models* do not work !

Common Speech-Tech wisdom

Hiring linguistics hurts the company's shares

Common IR-Tech wisdom

Linguistic models do not help retrieval

Can there be a role for NLP in applications ?

This talk: *Empirical Validity and Technological Viability*

Empirical Validity vs. Technological Viability

Empirically valid model: cognitive ? black-box view ?$\cdots$

Technologically viable model: what applications/resources ?$\cdots$

We leave psycholinguistics aside and concentrate now on the joint requirements (black-box model):

Technological: Correctness, robustness and efficiency

Cognitive: Correctness, robustness and efficiency

- **Where does the common wisdom come from ?**
- **How can we meet these requirements ?**

The Paradigmatic Role of Syntactic Processing

Syntactic processing (parsing) is interesting because:

Fundamental: it is a major step to utterance understanding

Well studied: vast linguistic knowledge and theories

Example role: formal devices of syntactic processing can be examples for subsequent processing (semantics, discourse,...)

Infrastructure: data and test-suits are available

Exploitable: applications can benefit from good parsing

“Shallow parsing” is already entering applications

Structure of Talk

- Set-theoretic (categorical) approach to parsing and where it fails
- Probabilistic approach: new life to the set-theoretic approach ?
- Advantages of the probabilistic approach: empirical validity
- Technological viability of the probabilistic approach
- Examples of existing parsing models
- A view on future research

Set-theoretic Approach to Parsing

Assigning linguistic structure to input utterances with the goal of facilitating semantic interpretation.

A Language is a **set** of sentence-analysis pairs

Formal devices: A language is described by a formal generative device
e.g. Context-Free / Unification Grammar,...

Belief: A formal grammar is suitable for processing utterances in order to extract syntactic structure

Does the set-theoretic approach satisfy the requirements set on applied/cognitive models ?

Problems of Set-Theoretic Approach

Ambiguity: Multiple analyses associated with the same sentence !

`BUT: Humans do select a single preferred analysis`

$\neg$ **Robustness:** Input is not in the set describing the language !

`BUT: Humans do understand ‘‘weird" utterances`

Inefficiency: *Worst case complexity* under grammar types !

`BUT: Humans process utterances efficiently`

Can the set-theoretic approach deal with these problems ?

CLAIM: THREE FACES OF UNCERTAINTY

UNCERTAINTY

Ambiguity **Lack-of-robustness** **Inefficiency**

Problem	**Uncertainty w.r.t.**
Ambiguity	Output: which output is best ?
¬Robustness	Input: what inputs to expect ?
Inefficiency	Processing algorithm: how to navigate ?

Ambiguity and Uncertainty

Ambiguity due to contextual / linguistic / extra-linguistic factors, e.g.

- **Word-sense:** `bank` (of the river) vs. `bank` (e.g. ABN-AMRO)
- **Part-of-speech:** `list` as verb/noun; `following` as verb/adj/noun
- **Sentence structure:** `I saw the man/dog with the telescope`
  `The telegraphy and telephone services are important`

Uncertainty due to hazard (in technological applications) e.g.

- **Spelling:** `I was` <u>`teading`</u>, (teading$\in\{$leading,reading,feeding,$\cdots\}$)
- **Speech:** `Travel to Almelo/Ermelo/marmalade/Elsloo`

Coverage and Robustness

What utterances are "grammatical" ? example problems:

- **"Ungrammatical" use:** `He say no to mom !` (third person agreement)

- **Infrequent use:** `Cats eat, tigers devour` (subcat frames)

What utterances might occur in the input ? example problems:

- **Speech utterances:** repetitions, corrections, hesitations,...

- **Communication noise:** sending messages over a channel

Efficiency and Expectations

Beyond *worst-case complexity*

Expectations "as in human processing", e.g.

Frequency: Does frequency of occurrence affect processing speed ?

Domain: What domain of language use ?

Context: Where a phrase is likely / unlikely to appear ?

Prediction: What to expect after seeing only part of an utterance ?

Limited beam: Why explore the whole space ?

Give Up the Set-Theoretic Approach ?

In this *methodological* issue we think this could be **unwise**:

Structure and Probability:

> *Employ the set-theoretic approach as a* **first informed approximation** *of the preferred model structure, and recast the model in Probabilistic formulae.*

Structure and Data (Bayesian Learning):

$$\arg max_{m \in Models} P(m \mid data) = \arg max_{m \in Models} P(m) \times P(Data \mid m)$$

Structured Probabilistic Language Models

Language Models: Extending Sets

A *language model* is a probability mass function over utterances-analyses:

$$\boxed{P : U \times T \longrightarrow [0,1]} \qquad \boxed{\sum_{\langle u,t\rangle \in (U\times T)} P(\langle u,t\rangle) = 1}$$

The probabilistic view provides:

- a generalization over sets + an established solution to uncertainty
- direct empirical interpretation: *Statistics*
- direct links to theories of *learning*
- methodological advantages, e.g. *model integration, optimization, hypothesis testing, evaluation*

Aspects of Language Models

- How do language models:

  (Q_1) Achieve disambiguation/robustness/efficiency ?

  (Q_2) Link to Learning, Statistics, (in)dependence and modularity ?

  (Q_3) Incorporate formal languages (probabilistic grammars) ?

- Briefly on state of the art:

  - Ambiguity resolution: Memory vs. Dependencies.

  - Robustness: smoothing by hidden structure.

  - Efficiency: pruning and model specialization.

Language Models and Ambiguity (Q_1)

Given a language model P:

Parsing utterances: for an input utterance u, output the pair

$$\langle u,t\rangle^* = arg\max_{\langle u,t\rangle} P(\langle u,t\rangle)$$

Ambiguous input: for an ambiguous input $U_x \subseteq U$, output

$$u^* = arg\max_{u\in U_x} \sum_{\langle u,t\rangle} P(\langle u,t\rangle)$$

How can we achieve correct disambiguation ?

Language Models and Robustness (Q_1 cont.))

A well-informed (e.g. linguistically) language model P might assign probability zero to some highly infrequent pair $\langle u,t \rangle \in U \times T$.

Smooth P to assign $P(u,t) \neq 0$ (e.g. Good-Turing, Katz)

Interpolate a weaker language model P_w with P

$$P_i = \lambda P + (1 - \lambda) P_w$$

Reveal latent structure μ for informed smoothing:

$$P(X) = \sum_{\mu} P(X, \mu) = \sum_{\mu} P(\mu) P(X|\mu)$$

$\boxed{\text{How can we achieve suitable robustness ?}}$

Language Models and Efficiency (Q_1 cont.)

Given an input utterance $u = w_1, \cdots w_n$:

Expectations: a probability mass function P_e over subutterance-subanalysis pairs $\langle u_1, t_1 \rangle$, given some preceding part $\langle u_0, t_0 \rangle$: $P(u_1, t_1 | u_0, t_0)$

Beam: prune distribution of subanalyses for $w_i, \cdots w_j$

Frequency: compile P such that more frequent utterances can be retrieved faster

How can we achieve satisfying efficiency ?

Statistics, Learning and Model Integration (Q_2)

Methodological advantages of language models:

- **Learning/Estimation:** parameter μ estimation from data D

Maximum-A-Posteriori	$arg\max_{\mu} P(\mu\|D)$	(Bayesian Updating)
Maximum-Likelihood	$arg\max_{\mu} P(D\|\mu)$	(Maximum-Entropy)
$\vdots$	$\vdots$	$\vdots$

Error-bounds: Bayesian classifiers.

- **Model Integration:** Noisy-Channel (explicit assumptions):

$$P(m,t|u) = P(m|t,u)P(t|u) \qquad (\mathbf{m}eaning, \mathbf{t}ree, \mathbf{u}tterance)$$

Probabilistic Grammars as Language Models (Q_3)

Extend formal grammars to become probabilistic grammars:

Parameters: how to estimate probabilities of the set μ of rewrite-events given their contexts ? what kind of context ?

Stochastic processes: how to estimate probabilities of derivations, i.e. sequences of rewrite-events in context ?

Probabilities of pairs: how to estimate probability $P(\langle u,t \rangle)$?

Represent a language model P by a set of parameter values μ

- What constitutes a good language model ?

Tree-Bank Grammars

Probabilistic grammar is a generative device (vs. reduction system):

Generative view: Every parse-tree t is generated from the start-symbol of the grammar S

Stochastic processes: a parse is generated through an ordered sequence of rewrite-events $r_1, \cdots, r_n$, each with probability conditioned properly

$$P(r_1 \cdots r_n | S) = \prod_{i=1}^{n} P(r_i | r_1, \cdots, r_{i-1})$$

Tree-bank: a representative multiset of utterance-tree pairs

Tree-bank models: the rewrite-events and their probabilities are extracted from a tree-bank

Example: Prob. Context-Free Grammar

A Probabilistic CFG (PCFG) extends a CFG with a probability mass function P over the finite set of rewrite-rules $\mathcal{R}$ such that

Generative model: probability of $A \longrightarrow \alpha$ conditioned on A only

Similar statement: for all nonterminals A: $\sum_{\alpha: A \rightarrow \alpha \in \mathcal{R}} P(A \longrightarrow \alpha | A) = 1$

Independence: no context effects, i.e. probability of derivation $d = r_1, \cdots, r_n$ is estimated by $P(d) = \prod_{i=1}^{n} P(r_i)$

Simple extension (remains PCFG): add some context, e.g. condition on label of parent of A, as extracted from examples found in the tree-bank

Current Research on Parsing (1)

Tree-bank based: training material D is a multiset of utterance-analysis pairs (manually annotated/corrected, use of specific domains)

Example tree-bank: Penn Wall Street Journal(WSJ), 10^6 words, 5×10^4 sentences of average length 23 words (up to 115 words !) from the WSJ newspaper

Main questions: what rewrite units, context to extract ? with what probabilities ? how to smooth using linguistic knowledge ?

Evaluation: Labeled Recall/Precision (percentage of nodes that exactly match) over a test-set of 2400 sentences not involved in training

Current Research on Parsing (2)

Magnitude of problem: $\approx$ 75%/75% recall/precision for **broad-coverage linguistic grammars** (IBM; Probabilistic LFG (PARC)), *each developed over* $>$ 10 *linguistic-labour years*

Bilexical-dependency models: $\approx$ 91%/91%, a well-smoothed model with probabilities ranging over dependencies between head-words

Data Oriented Parsing: $\approx$ 91%/91% with a model that puts probabilities over large chunks of linguistic structure

Two successful kinds of rewrite-events (3)

Bilexical-dependencies: head-driven Markov Grammars, e.g. (Collins 97, Charniak 99)

Events: pairs $\langle Parent.word1, Child.word2 \rangle$.

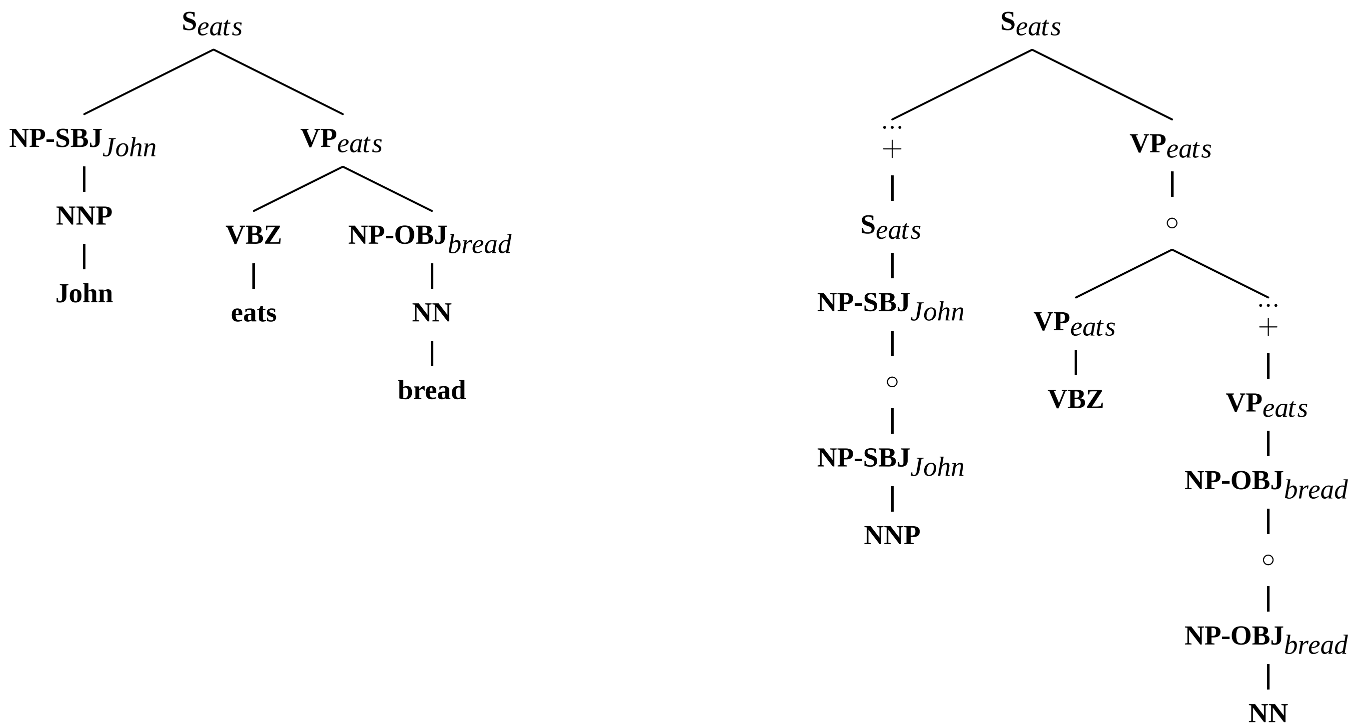
Probability: $\boxed{P(\langle Parent.word1, Child.word2 \rangle \mid Parent.word1)}$.

Structural relations: e.g. DOP (Scha 90; Bod 95; Sima'an 99).

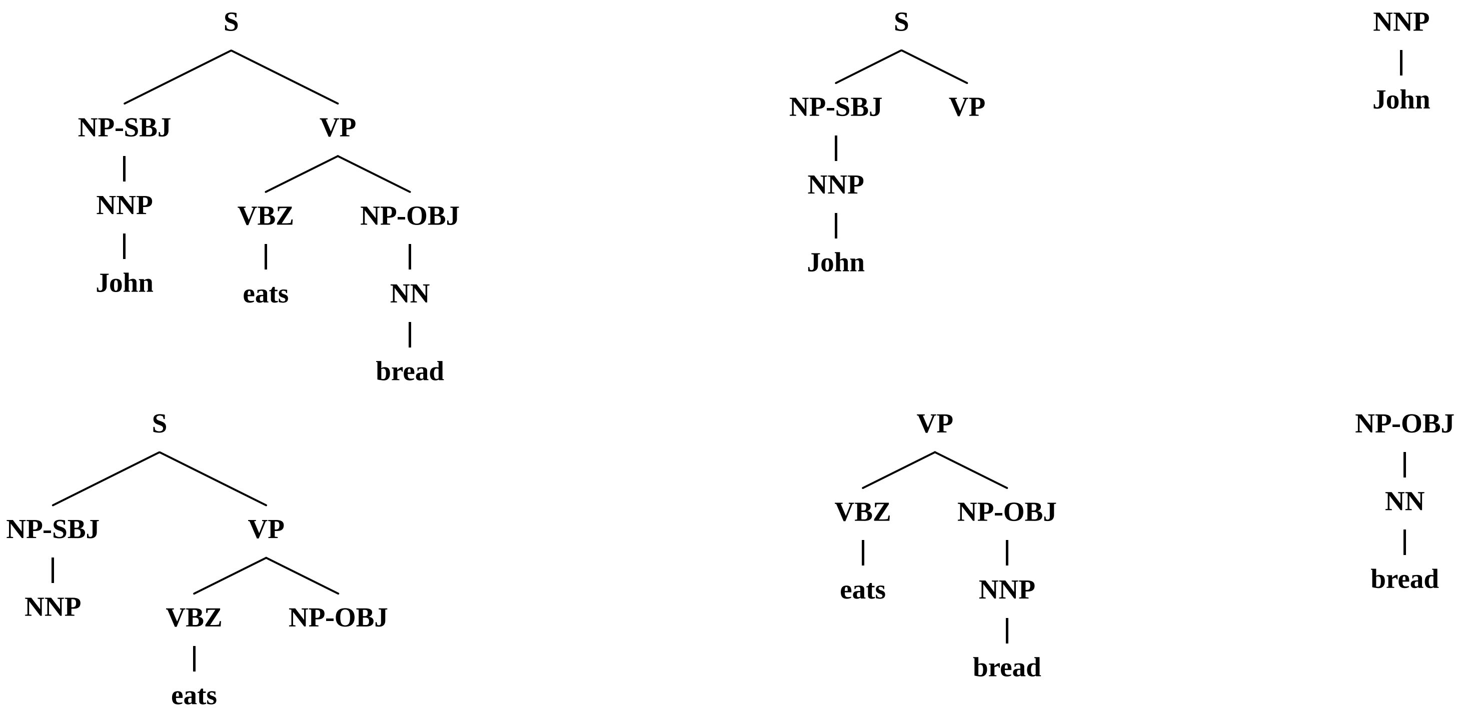
Events: arbitrary size syntactic structures, e.g. DOP subtrees are connected CFG-rules,

Probability: $\boxed{P(subtree \mid label.root(subtree))}$

Example: Bilexical-dependencies



Example: Data Oriented Parsing (some subtrees)



Smoothing for Robustness: Examples

Backoff: many possibilities (Katz method):

$$P(\langle A, X\rangle \rightarrow \alpha | \langle A, X\rangle) \approx \Theta(\ P(A \rightarrow \alpha | A)\ P(X \rightarrow \alpha | X)\ )$$

Markov Grammar: smoothing PCFGs for flat Phrase-Structure (Collins 1996, Charniak 1999):

$$P(A \rightarrow B_1 \cdots B_n | A) = P(B_1 | A) \times \prod_{i=2}^{n} P(B_i | A, B_1, \cdots, B_{i-1})$$

Hidden Structure: Assume edit-operations (delete, insert,$\cdots$) on frames as a hidden process (Eisner 2001):

- Given set of frames, each with a probability given a verb,
- Expand by Expectation Maximization (EM) on large bodies of text.

Efficiency: e.g. Suitable Pruning

How to allow pruning of subanalyses $XP \longrightarrow *(w_i \cdots w_j)$?

Inside probabilities: Language models provide estimates for

$$P(XP \longrightarrow^* w_i \cdots w_j)$$

Outside probabilities: BUT they do not provide estimates for:

$$P(S \longrightarrow^* w_1 \cdots w_{i-1}\, XP\, w_{j+1} \cdots w_n)$$

Pruning: use approximations of the Outside probabilities, estimated on many examples from a given domain of language use.

Future: More Expected Utterances Processed Faster

Next Issues in Empirical NLP

- Feature-structures + **Distributional-Similarity** (“Prob. Unification”)
- Robust and correct semantics of utterances
  - Lambda expressions, compositionality and “cooccurence semantics”
  - Predicate-Argument structures, dropping/insertion of arguments
  - Distributions over Lambda-expressions: expressing underspecification
- $P(sem, syn, utter) = P(sem)P(syn|sem)P(utter|sem, syn)$

Future: Cooccurence Statistics over Structure for Processing