Between Computational Linguistics and Computation for Linguistics

Khalil Sima'an

Institute for Logic, Language and Computation (ILLC)
University of Amsterdam, The Netherlands

Israeli Seminar on Computational Linguistics (ISCOL)
The Israel Association for Theoretical Linguistics (IATL)
Technion, Haifa – 22 June 2005

Computational Linguistics Research

Old *"Computation for linguistics"*
Formalization + computation for linguistic theorizing.
E.g. Formal grammars, parsing algorithms, ...
Examples: HPSG/LFG, parsing algorithms and complexity.

New Probabilistic models, statistical inference, ML methods, corpora, applications etc.

"... anyone who cannot at least use the terminology persuasively risks being mistaken for kitchen help at the ACL banquet." (Abney 1996).

What is this new Computational *Linguistics*?

Computational Linguistics Research

Old *"Computation for linguistics"*
Formalization + computation for linguistic theorizing.
E.g. Formal grammars, parsing algorithms, ...
Examples: HPSG/LFG, parsing algorithms and complexity.

New Probabilistic models, statistical inference, ML methods, corpora, applications etc.

"... anyone who cannot at least use the terminology persuasively risks being mistaken for kitchen help at the ACL banquet." (Abney 1996).

What is this new Computational *Linguistics*?

Computational Linguistics Research

Old *"Computation for linguistics"*
Formalization + computation for linguistic theorizing.
E.g. Formal grammars, parsing algorithms, ...
Examples: HPSG/LFG, parsing algorithms and complexity.

New Probabilistic models, statistical inference, ML methods, corpora, applications etc.

"... anyone who cannot at least use the terminology persuasively risks being mistaken for kitchen help at the ACL banquet." (Abney 1996).

What is this new Computational *Linguistics*?

A Methodological Frenzy

Initially: Language models
Inspired by success in speech-recognition.

Later: A swelling number of approaches

- Joint vs. conditional models/estimation ...
- HMM, Prob. CF/TS/TI/TA Grammars, Prob. LFG ...
- ML, MaxEnt, Cond. Random Fields ...
- SVMs, MBL, Decision-Trees ...

This talk: – Review a stripped skeleton of CL research
– Contemplate the meaning of this all

A Methodological Frenzy

Initially: Language models
Inspired by success in speech-recognition.

Later: A swelling number of approaches

- Joint vs. conditional models/estimation ...
- HMM, Prob. CF/TS/TI/TA Grammars, Prob. LFG ...
- ML, MaxEnt, Cond. Random Fields ...
- SVMs, MBL, Decision-Trees ...

This talk: – Review a stripped skeleton of CL research
– Contemplate the meaning of this all

Bits and Pieces of Current CL

Approach:	Input-output (blackbox) models of language use
Methods:	Probabilistic models+statistical inference
Data+:	Corpora for estimation + smoothing
Domain:	Specific language use (e.g. WSJ text)
Technology:	Tools for technological applications
Evaluation:	Evaluation using quantitative measures

Do these pieces fit together as in a puzzle?

Where is the (linguistic) structure?

Bits and Pieces of Current CL

Approach:	Input-output (blackbox) models of language use
Methods:	Probabilistic models+statistical inference
Data+:	Corpora for estimation + smoothing
Domain:	Specific language use (e.g. WSJ text)
Technology:	Tools for technological applications
Evaluation:	Evaluation using quantitative measures

Do these pieces fit together as in a puzzle?

Where is the (linguistic) structure?

Bits and Pieces of Current CL

Approach:	Input-output (blackbox) models of language use
Methods:	Probabilistic models+statistical inference
Data+:	Corpora for estimation + smoothing
Domain:	Specific language use (e.g. WSJ text)
Technology:	Tools for technological applications
Evaluation:	Evaluation using quantitative measures

Do these pieces fit together as in a puzzle?

Where is the (linguistic) structure?

Language Use in Technological Environments

Goal: Computational models for language technology, e.g.

- Document/information retrieval systems, question-answering systems
- Speech recognition/generation; dialogue systems
- Translation and summarization systems
- Spelling-correction and grammar checking
- Other, e.g. Optical character recognition (OCR)
- ...
- Tools: parsers, POS taggers, ...

The challenges of Technology

Diversity The applications are diverse: speech vs. language, parsing vs. OCR? translation vs. spelling-correction?

Performance The models will have to perform in actual situations:
Performance evaluation measures
Robustness, accuracy (and efficiency)

What is the single concept that will support all this?

Noise as Information Concept

Linguists are fascinated with select phenomena, e.g.
"Colorless green ideas sleep furiously"

CLs seem to be fascinated with...

Noise

Claim:

Linguistics		**Current CL**
Grammaticality		Noise

From noise all the way to grammar?

Noise as Information Concept

Linguists are fascinated with select phenomena, e.g.
"Colorless green ideas sleep furiously"

CLs seem to be fascinated with...

Noise

Claim:

Linguistics		**Current CL**
Grammaticality		Noise

From noise all the way to grammar?

Noise as Information Concept

Linguists are fascinated with select phenomena, e.g. "Colorless green ideas sleep furiously"

CLs seem to be fascinated with...

Noise

Claim:

Linguistics		**Current CL**
Grammaticality		Noise

From noise all the way to grammar?

Noisy Communication – Shannon 1949

- A message is communicated through a noisy channel
- An observation is obtained at the end of the channel
- Need "best guess" of the original message from observation.

Was "noise" Shannon's actual problem?
Is CL studying models of noisy communication?

The Noisy-Channel Task

Guess the original message m given the observation o with a minimum expected ratio of errors.

Alfabet A finite set over which define m and o

Needed A set of hypotheses M
A set of observations O

Black Box: The best bet m^* fulfills

$$P(m^*) = \max_{m \in M} P(m|o)$$

Why? Gives us the smallest risk (on average)

Optimization: Relative judgment!

The Bayesian Inversion

$$argmax_{m \in M} P(m|o) = argmax_{m \in M} P(o|m)P(m)$$

Conceptual (and technical) separation of concerns

Language model: $P(m)$ (or source)

Language = a distribution over M (vs. set M)

"Task" model: $P(o|m)$ (or noisy channel)

Probability of noisy observation given hypothesis message

Language applications $\geq$ language models!

Examples: Language $P(m)$ and Task $P(o|m)$

Speech-Rec. o = acoustic signal; m = word sequence

Spelling-Cor. o = corrupted word (sequence)
m = "dictionary" word (sequence)

Machine Trans. Source language sentence s;
Target language sentence t

$$argmax_{t \in T} P(t|s) = argmax_{t \in T} P(s|t)P(t)$$

- T is set of possible sentences in target language
- $P(t)$ is target language model
- $P(s|t)$ is "translation" model ($t \rightarrow s$)

What Kinds of Noise?

In language processing tasks, noise stands for a blend of what we call

- **Channel:** *Acoustic waves*, telephone hazard, printing ink stains,... stuttering, hesitation...
- **Lack of knowledge** of language use, of task, world-knowledge, situation, cultural influence, ...
- **Other factors:** Frequency effects in human processing,...,

Which bit of noise originates from which source?

Shannon's noise: irregularity relative to expected regularity

Noise is interpreted as *uncertainty*

What Kinds of Noise?

In language processing tasks, noise stands for a blend of what we call

Channel: *Acoustic waves*, telephone hazard, printing ink stains,... stuttering, hesitation...

Lack of knowledge of language use, of task, world-knowledge, situation, cultural influence, ...

Other factors: Frequency effects in human processing,...,

Which bit of noise originates from which source?

Shannon's noise: irregularity relative to expected regularity

Noise is interpreted as *uncertainty*

What Kinds of Noise?

In language processing tasks, noise stands for a blend of what we call

- **Channel:** *Acoustic waves*, telephone hazard, printing ink stains,... stuttering, hesitation...
- **Lack of knowledge** of language use, of task, world-knowledge, situation, cultural influence, ...
- **Other factors:** Frequency effects in human processing,...,

Which bit of noise originates from which source?

Shannon's noise: irregularity relative to expected regularity

- **Noise** is interpreted as *uncertainty*

Modeling Input and Output

$$\arg\max_{m \in M} P(o|m)P(m)$$

The sets M and O must be known a priori

- A set of alternative, "well-formed" messages/observations
- A Formal, generative grammar of some sort defining $G \subseteq M \times O$: Finite-State, CF or more complex

- **Dilemma:** What to assume about M, O and G?
  - Coverage Assume too much $\Longrightarrow$ intolerant to noise
    - `Ex: Ling. grammar for spoken utterances`
  - Accuracy Assume too little $\Longrightarrow$ too permissive
    - `Ex: Word-based Finite-State for written text`

!! Use data as prior: Corpus

Modeling Input and Output

$$\arg\max_{m \in M} P(o|m)P(m)$$

The sets M and O must be known a priori

- A set of alternative, "well-formed" messages/observations
- A Formal, generative grammar of some sort defining $G \subseteq M \times O$: Finite-State, CF or more complex

- **Dilemma:** What to assume about M, O and G?

  Coverage Assume too much $\Longrightarrow$ intolerant to noise

  `Ex: Ling. grammar for spoken utterances`

  Accuracy Assume too little $\Longrightarrow$ too permissive

  `Ex: Word-based Finite-State for written text`

!! Use data as prior: Corpus

Modeling Input and Output

$$\arg\max_{m \in M} P(o|m)P(m)$$

The sets M and O must be known a priori

- A set of alternative, "well-formed" messages/observations
- A Formal, generative grammar of some sort defining $G \subseteq M \times O$: Finite-State, CF or more complex

- **Dilemma:** What to assume about M, O and G?

  Coverage Assume too much $\Longrightarrow$ intolerant to noise

  `Ex: Ling. grammar for spoken utterances`

  Accuracy Assume too little $\Longrightarrow$ too permissive

  `Ex: Word-based Finite-State for written text`

!! Use data as prior: Corpus

Corpora as "Priors"

Current research employs corpora

Why? exemplifies the structural and frequency variation in future data ("prior")

Which? Specific domains of language use
A priori known application (environment).

Best practice "Corpus grammars", extract from corpus

Grammar: A rough definition of $G \subseteq M \times O$

Statistics: Statistical estimates of P

Examples: Annotated corpora for speech-recognition, POS tagging and parsing

Linguistic annotations enter from the backdoor!

Corpora as "Priors"

Current research employs corpora

Why? exemplifies the structural and frequency variation in future data ("prior")

Which? Specific domains of language use
A priori known application (environment).

Best practice "Corpus grammars", extract from corpus

Grammar: A rough definition of $G \subseteq M \times O$

Statistics: Statistical estimates of P

Examples: Annotated corpora for speech-recognition, POS tagging and parsing

Linguistic annotations enter from the backdoor!

Why the Backdoor?

Linguistic grammars provide

Grammaticality Hopefully sign of input regularity

Structure Set (possibly $\emptyset$) of analyzes per input

However, grammars bring with them

Intolerance No account for domain-specific language, e.g. "sounding natural" etc.
No access to context

Structure enters via corpus into probabilistic grammars

Structured Language Models $P(m)$?

$$P : M \rightarrow [0,1] \qquad \sum_{m \in M} P(m) = 1$$

M is set of utterances (word sequences)

`Are language models over structure possible?`

Correct question: *How much and what kind of structure?*

Flat: Markov models over word sequences

Structured: $P(m) = \sum_{t \in T} P(m,t)$

where T is a set of disjoint/alternative structures

e.g. T might be sequence of POS tags, parse...

Structured Language Models $P(m)$?

$$P : M \rightarrow [0,1] \qquad \sum_{m \in M} P(m) = 1$$

M is set of utterances (word sequences)

`Are language models over structure possible?`

Correct question: *How much and what kind of structure?*

Flat: Markov models over word sequences

Structured: $P(m) = \sum_{t \in T} P(m,t)$

where T is a set of disjoint/alternative structures

e.g. T might be sequence of POS tags, parse...

Choosing the Right Building Blocks

Generative grammars define

- $P(m,t) = \sum_{der} P(m,t,der)$ where *der* is a derivation
- derivation *der* involves rewrite rules and operations
- This is just like tiling $\langle m,t \rangle$ with *basic building blocks*

- How to:
  - define the right kind of building blocks and their probs?
  - estimate these probs from corpora?

Grammar	probability	building blocks
Markov (SFSA)	$P(w_n \mid w_{n-1})$	bigrams
PCFG	$P(S \rightarrow NP\ VP \mid S)$	phrase-str. rules

Which building blocks are suitable for language structure?

Choosing the Right Building Blocks

Generative grammars define

- $P(m,t) = \sum_{der} P(m,t,der)$ where *der* is a derivation
- derivation *der* involves rewrite rules and operations
- This is just like tiling $\langle m,t \rangle$ with *basic building blocks*

- How to:
  - define the right kind of building blocks and their probs?
  - estimate these probs from corpora?

Grammar	probability	building blocks
Markov (SFSA)	$P(w_n \mid w_{n-1})$	bigrams
PCFG	$P(S \rightarrow NP\ VP \mid S)$	phrase-str. rules

Which building blocks are suitable for language structure?

Example: Which Rules? (1)

Parsing the Penn WSJ corpus:

Task Build a parser for WSJ text.

Parser Sentence s and set of parses T

$$\arg\max_{t \in T} P(t|s) = \arg\max_{t \in T} P(t)P(s|t) = \arg\max_{t \in T} P(t)$$

Corpus Penn WSJ treebank (phrase-structure)

Baseline Extract Probabilistic Context-Free Grammar (PCFG) directly from treebank.

Results %75 labeled recall and precision (Charniak 1996)

Example: Which Rules? (2)

More internal structure:

Heads Phrases + heads over phrase-structure

Dependency predicate-argument (e.g. selectional preferences).

$$P(a\_dog \overset{subject}{\longrightarrow} bark) \geq P(snake \overset{subject}{\longrightarrow} bark)$$

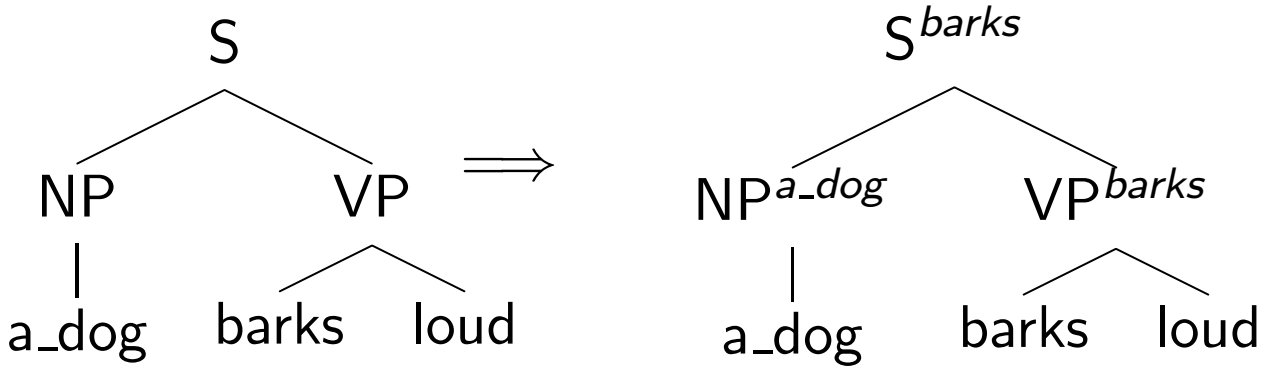
$$P(pizza\ with\ egg \overset{object}{\longrightarrow} eat) \geq P(pizza\ with\ fork \overset{object}{\longrightarrow} eat)$$

Approach:

- Transform Penn treebank into phrase-str. annotated with heads.
- Extract PCFG from this new treebank.

Example: Head-Lexicalized PCFGs

Example



Prob left $P(S \longrightarrow NP \;\; VP \mid S)$

Prob right $P(S^{barks} \longrightarrow NP^{a\_dog} \;\; VP^{barks} \mid S^{barks})$

Interpretation $P(a\_dog \longrightarrow barks,\; S \rightarrow NP \;\; VP \mid S^{barks})$

A lot of context and smoothing (!!)

Results: 85–89% labeled recall and precision

Examples: Magerman 1995; Collins 1997; Charniak 1999

Why "Corpus Grammars"?

Extract the rules together with probabilities from the corpus/treebank.

Why extract rules from corpus directly?

Annotation: Allows fitting the right rules to corpus sentences as annotation proceeds.

Context Can be extracted for the rules

Model: Direct extraction allows experimenting with different kinds of rules+context

Example: Parsing: a rule such a $NP \rightarrow DET\ NP$ is extracted with context, e.g. appears under an S, VP, NP etc.

Corpus Grammars: Experience and Memory

Interpreting Corpus Grammars:

- Suppose the corpus stands for the experience of an adult language user.
- Extracting the rules-in-context directly from corpus stands for memorizing.
- For analyzing new input, combine together rules-in-context and disambiguate using probabilities
- *How much to memorize?* *When to start forgetting?*

- Has CL shifted towards

  Probabilistic memory-based language processing?

Estimating $P : M \longrightarrow [0, 1]$ from a corpus

(1) $P : M \longrightarrow [0, 1]$ (2) $\sum_{m \in M} P(m) = 1$

Events M is the set of sentences

Hypotheses H is **some** set of distributions over M
$\forall h \in H$: constraint (2) holds
(usually other constraints also hold)

Corpora C^* is set of all corpora over M

Estimator $est : C^* \longrightarrow H$

Example Maximum-Likelihood: $est(c) = \arg\max_{h \in H} h(c)$

Properties of Estimators: Consistency

Target distribution $h^* \in H$

Corpora C^n = all corpora of size n sampled from $h^* \in H$.

Consistency:

- Estimator *est* is *consistent w.r.t.* h^* iff sum of probabilities of all $c \in C^n$ for which $est(c)$ disagrees with h^* for some $m \in M$ will diminish as $n \rightarrow \infty$.
- *est* is consistent with respect to H iff it is consistent w.r.t. every $h^* \in H$.

Assumption Given corpus, we know H!

Practice Consistency is necessary but not sufficient!

Challenge for Language Structure

For linguistic tasks (e.g. parsing, semantics etc):

- Can we find an estimator that will provide good estimates from any space of hypotheses H? *Very hard* .
- Can deeper structure of sentences help delimit "better" spaces H? *Most likely*.

Ex. {head-driven phase-str. PCFGs} vs. {phrase-str. PCFGs}

Current practice (mostly)

- Maximum-Likelihood Estimator is consistent[1]
- Worry only about defining space H!

[1] W.r.t. H if the ML estimate is in H!

Challenge for Language Structure

For linguistic tasks (e.g. parsing, semantics etc):

- Can we find an estimator that will provide good estimates from any space of hypotheses H? *Very hard* .
- Can deeper structure of sentences help delimit "better" spaces H? *Most likely*.

Ex. {head-driven phase-str. PCFGs} vs. {phrase-str. PCFGs}

Current practice (mostly)

- Maximum-Likelihood Estimator is consistent[1]
- Worry only about defining space H!

[1] W.r.t. H if the ML estimate is in H!

Sparse-Data: The Structure-Frequency Dilemma

Relative frequency estimator: $P(A|B) = \frac{Count(A,B)}{Count(B)}$

Probability estimates will be good only when corpus is large enough.

For all other corpora:

Words might not be in the corpus.

Rules might be missing from corpus grammar.

More complex models drag more history along their derivations. More history implies sparser statistics (for same corpus).

How can we use "more structured" language models?

Smoothing by Abstraction

- A derivation, breaks the parse into rules and their history $P(r_0, r_1, \ldots r_k) = P(r_0) \prod_{i=1}^{k} P(r_i | r_1, \ldots, r_{i-1})$
- Suppose we are rewriting rule r now
- History $h_0^n = \langle h_0, \ldots h_n \rangle$: $P(r|h_0^n)$
- As n grows, $Count(\langle r, h_0^n \rangle)$ becomes smaller

Smoothing by abstraction

Backoff When $Count(\langle r, h_0^n \rangle) = 0$: gradually forget h_j, starting $j = n$ down, until $Count(\langle r, h_0^i \rangle) \neq 0$

Interp. Interpolate all backoff models with current model for all events $\langle r, h_0^n \rangle$.

How much probability to reserve for unseen events? Good-Turing

Examples Abstraction and Independence

- The estimate of $P(w|v)$ is bad

Bkf In case $Count(w, v) \approx 0$
$\hat{P}(w|v) \approx \alpha(v)P(w)$
$\alpha(v)$ determined by Good-Turing and normalization

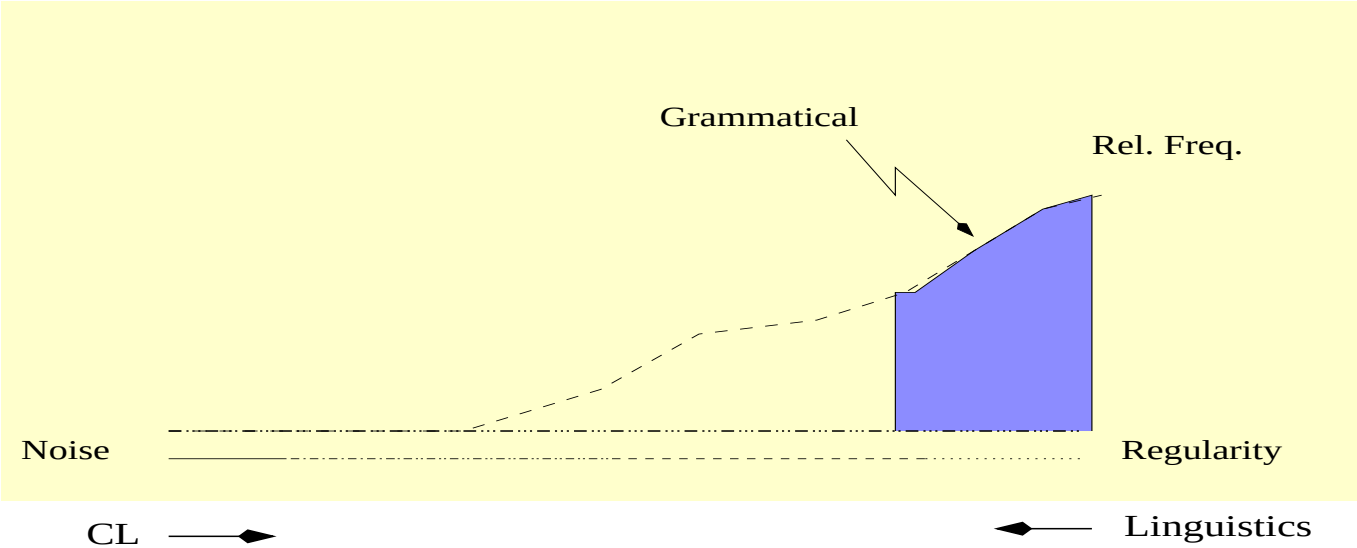
- The estimate of $P(A \rightarrow B\ C|A)$ is bad
- $P(A \rightarrow B\ C|A) = P(A \rightarrow B\ A_R|A)P(A_R \rightarrow C|A \rightarrow B\ A_R)$

Bkf In case $Count(A \rightarrow B\ C) \approx 0$
$\hat{P}(A \rightarrow B\ C|A) \approx P(A \rightarrow B\ A_R|A)P(A_R \rightarrow C|A_R)$

Smoothing by similarity (Zavrel+Daelemans'97; Dagan++'99)

Grammaticality or Noise?

Suppose[2] we have a scale of regularity/noise in sentences

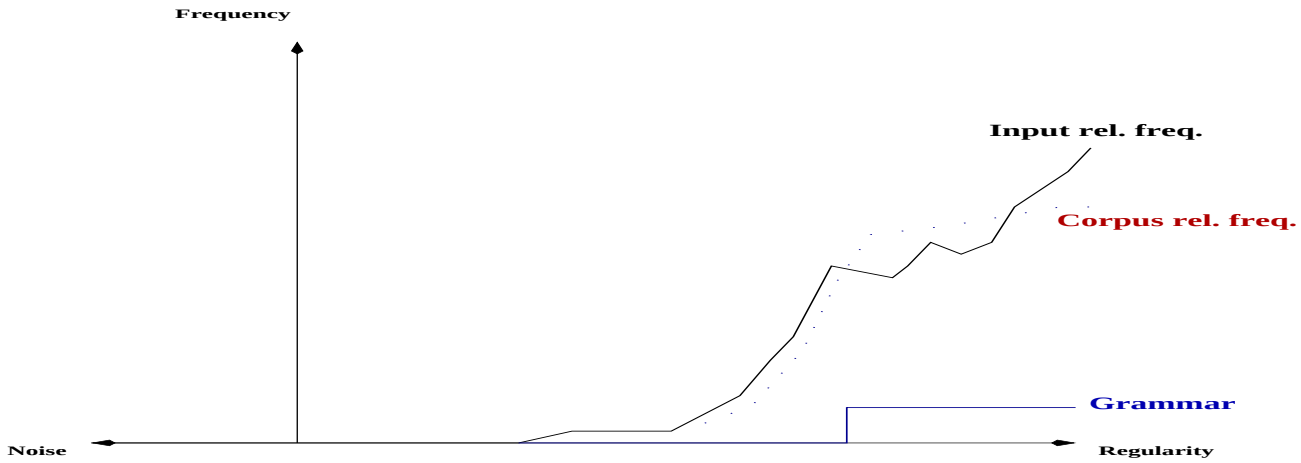


- Linguistics fixes grammaticality judgments
- CL looks for regularity within noise

[2] This presupposes a grammar already!

Corpora as Priors: Corpus Grammars

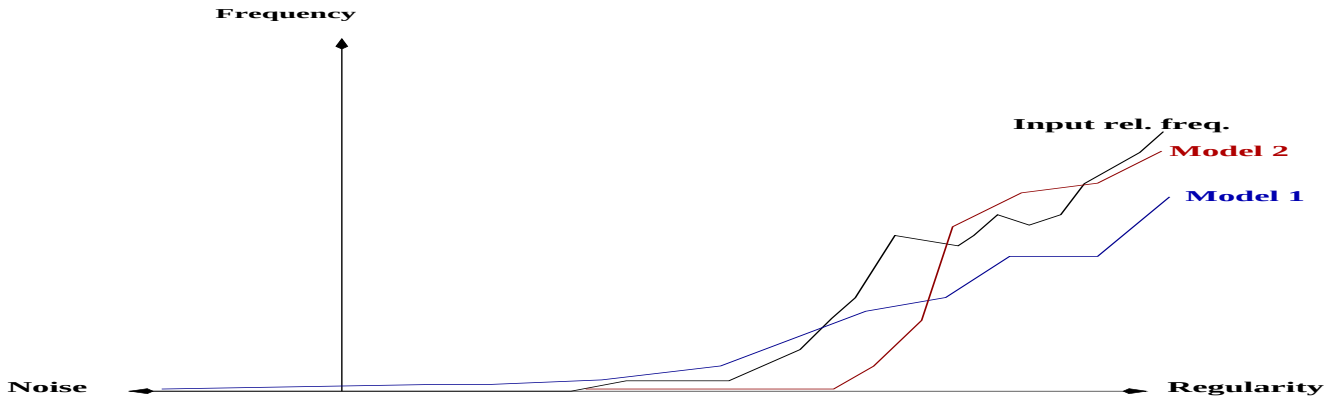
When a suitable[3] corpus is collected



- A corpus provides evidence of frequency and variation of expected future input
- Grammar only provides the judgments

[3] In the sense that it is a good sample.

Adding More Structure Demands Smoothing



Model 1: Little knowledge of grammar/structure
Spreads probability over inputs
Less accurate in expected frequency

Model 2: More knowledge of grammar/structure
Is more crisp: many inputs are ungrammatical
The more regular the input, the better the expected frequency

Interpolating the two models is unavoidable!

Corpus-Driven Models

Annotation can be any graph representations over sentences.

Rules can be any Bigrams or linguistic grammar productions are all events in the corpus with contextual evidence.

Models can be any Models of varying complexity for approximating $P(m)$, e.g. Markov models over words, Bilexical-Dependency over parses...

Already in 1995 Frederick Jelinek said ...

"put language back into language models".

Current Spots for Knowledge of Language

In current CL research:

- Corpus annotations and feature engineering.
- More suitable *kinds of* grammars:
  - Space of hypotheses for estimation.
  - Grammar that generates message-observation pairs.

"Future telling"

Short term More methodology frenzy, "Internet-driven" research, applications ...

Long term Induction of structure from raw corpus?

Current Spots for Knowledge of Language

In current CL research:

- Corpus annotations and feature engineering.
- More suitable *kinds of* grammars:
  - Space of hypotheses for estimation.
  - Grammar that generates message-observation pairs.

"Future telling"

Short term More methodology frenzy, "Internet-driven" research, applications ...

Long term Induction of structure from raw corpus?

Current Spots for Knowledge of Language

In current CL research:

- Corpus annotations and feature engineering.
- More suitable *kinds of* grammars:
  - Space of hypotheses for estimation.
  - Grammar that generates message-observation pairs.

"Future telling"

Short term	More methodology frenzy, "Internet-driven" research, applications ...
Long term	Induction of structure from raw corpus?

Towards Language Engineering?

Engineering demands fixed formulae!

Current situation:

- Collect a suitable corpus
- Conduct "feature-detection" on corpus (linguist)
- Create an "annotated" training corpus using features
- Train a statistical method on corpus.
- Evaluate the system empirically

- To improve, use more data and find new features.

The more "fixed" this formulae the better!

Finally...

Technology leads to new questions and often even to new fields of science e.g. the steam engine brought Thermodynamics; telegraphy brought Communication and Information Theory.

About building models:

> *Our task is not to penetrate into the essence of things, the meaning of which we don't know anyway, but rather to develop concepts which allow us to talk in a productive way about phenomena in nature. (Niels Bohr).*