Statistical NLP at First Glance

Detlef Prescher
University of Heidelberg
prescher@cl.uni-heidelberg.de

April 17, 2007

Applications of NLP

Natural language and NLP play a central role in systems that

- **augment textual or spoken data with information** (e.g. automatic transcription of speech signals, part-of-speech tagging, named-entity recognition, parsing/chunking, word-sense disambiguation)
- **transform textual or spoken data** (e.g. text-to-speech, speech-to-text, spelling correction, text summarization, machine translation)
- **extract information from textual or spoken data** (e.g. information retrieval, question answering, information extraction, data mining)
- **communicate with people** (dialog systems)

The Elements of NLP

Phonetics/Phonology: map acoustic signals to phoneme and/or grapheme sequences and vice versa (speech recognition/synthesis)

Morphology: analyze the structure of words (morphological analysis)

Syntax: identify the category of words (POS tagging), analyze the structure of sentences (parsing/generation)

Semantics: calculate the meaning of words/sentences (lexical/compositional semantics)

Discourse: analyze the structure of dialog or text (discourse representation)

Pragmatics: incorporate world knowledge, cultural conventior a specific use of language.

The Aim of NLP

Scientific: Build models reflecting the human use of language and speech.

Technological: Build models that serve in technological applications.

The main NLP questions are:

1. What are the kind of things that people say and write?
2. What do these things mean?
3. How to incorporate the knowledge about these things into algorithms?

How to build models of NLP?

Traditional View: Competence (Chomsky, ~ 1960)
Grammaticality of sentences in a language is defined via a set membership test:

- A *sentence* is a sequence of words,
- A *language* is a set of sentences,
- A *formal grammar* is a device defining the language,

Modern View: Performance (~ 1990)
Given a specific NLP task and a specific domain of language use, the human language-behavior is modeled by a

(black-box) function: input $\longrightarrow$ **output**,

the output that humans perceive as the most plausible for a given input.

What changed NLP?

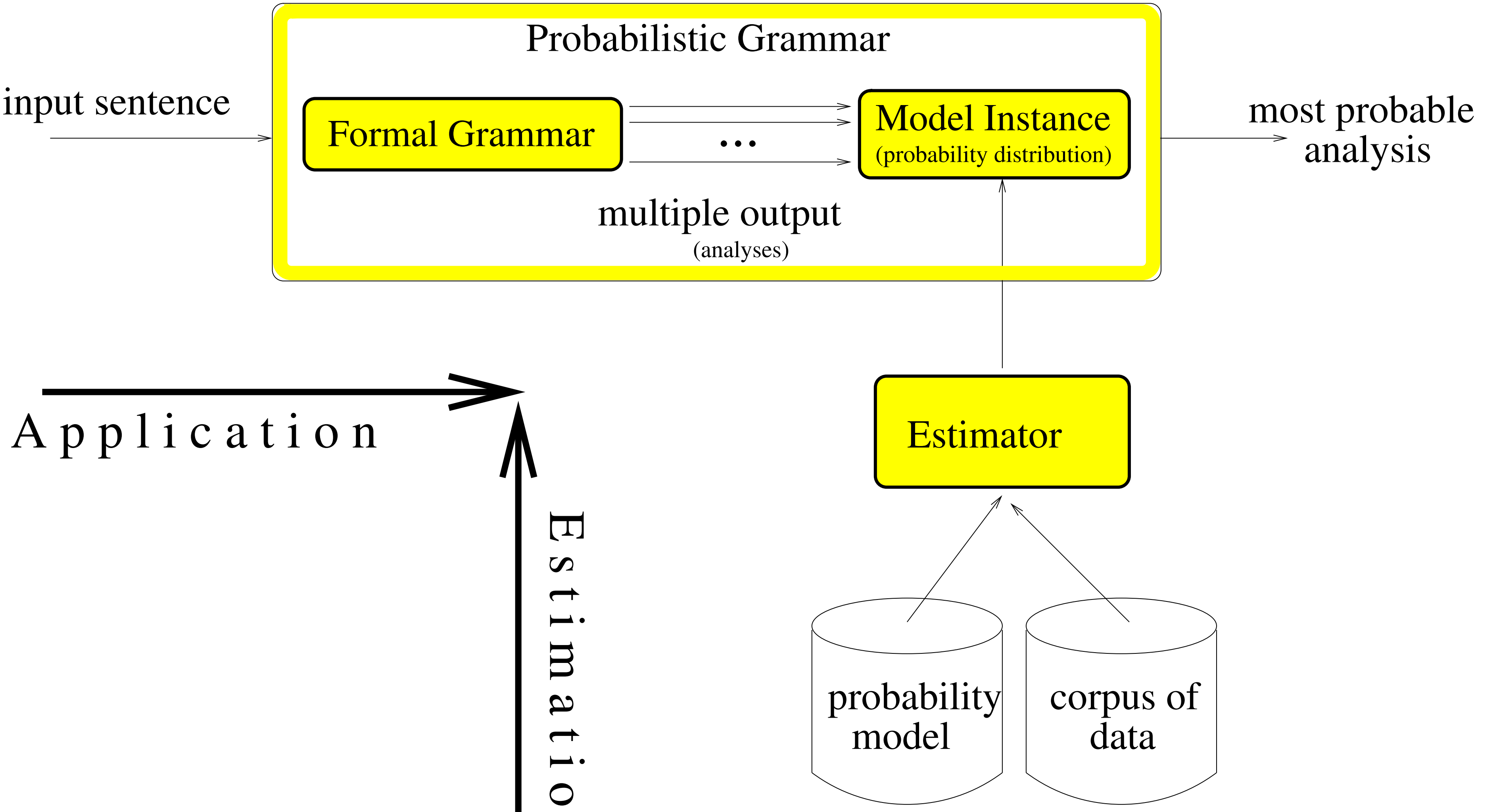
Competence Models: In contrast to people, the linguistic view of language as a set does not care about problems caused by ambiguity. Competence models

- **cannot resolve multiple output** ☹
- **cannot handle multiple input (noisy utterances)** ☹
- **cannot express multiple levels of grammaticality** ☹

Performance Models: Mimic people's language behavior and are specifically designed to resolve ambiguity. They

- handle uncertainty with Probability Theory and Statistics ☺
- utilise competence models as components ☺
- have even the potential to model extra-linguistic factors ☺

Building Models of NLP



Probabilistic Grammar

input sentence

Formal Grammar

...

Model Instance
(probability distribution)

most probable
analysis

multiple output
(analyses)

Application

Estimation

Estimator

probability
model

corpus of
data

– either on sentences or analyses (un/supervised learning)

Example: Grammar Estimation