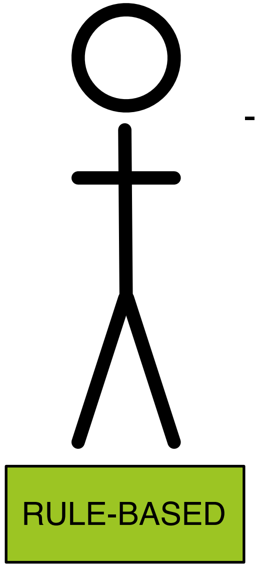
Statistical Machine Translation and Hybrid Machine Translation

Philipp Koehn
University of Edinburgh

11 August 2006

The Discussion

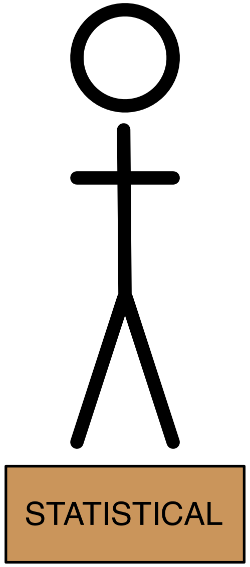


Rationalist Paradigm

- understand basic principles of language and translation
- encode this knowledge in representation and rules

Empiricist Paradigm

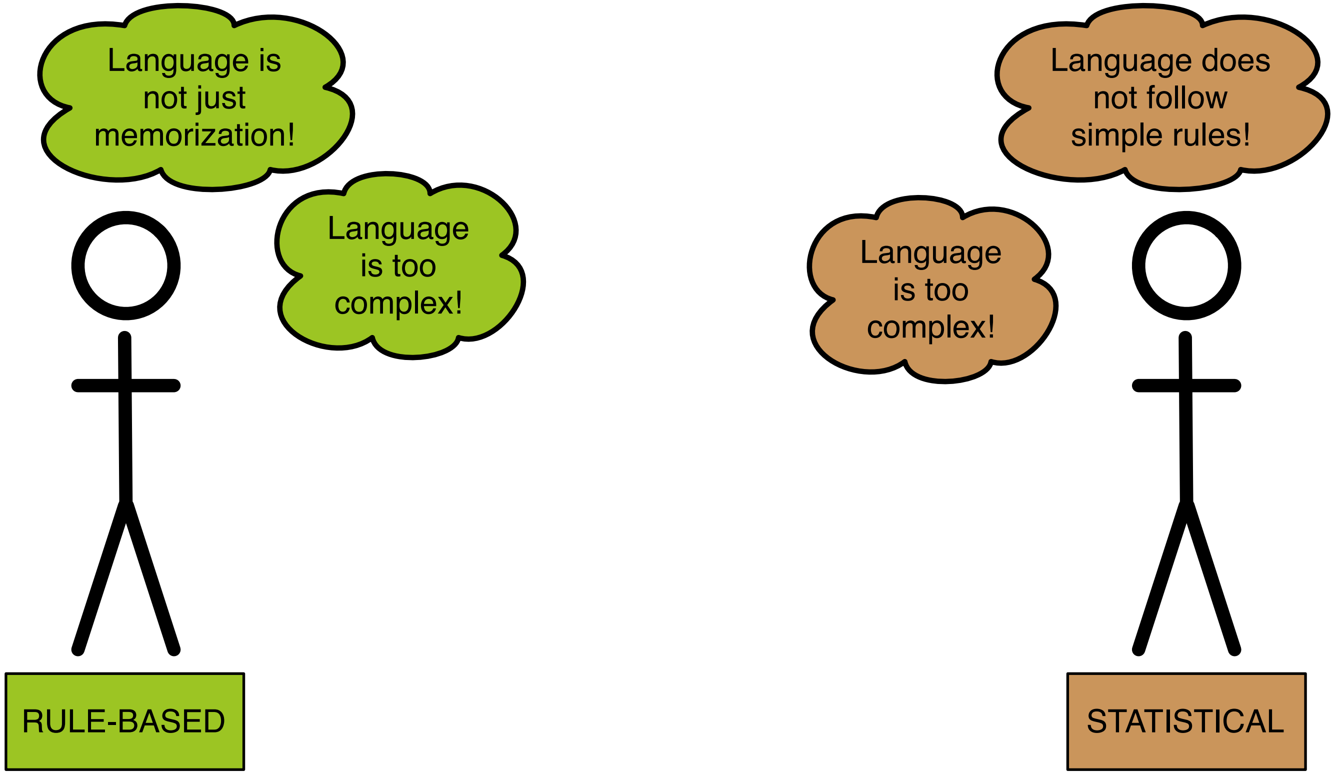
- automatically analyze large amounts of translated text
- build models that learn from this data



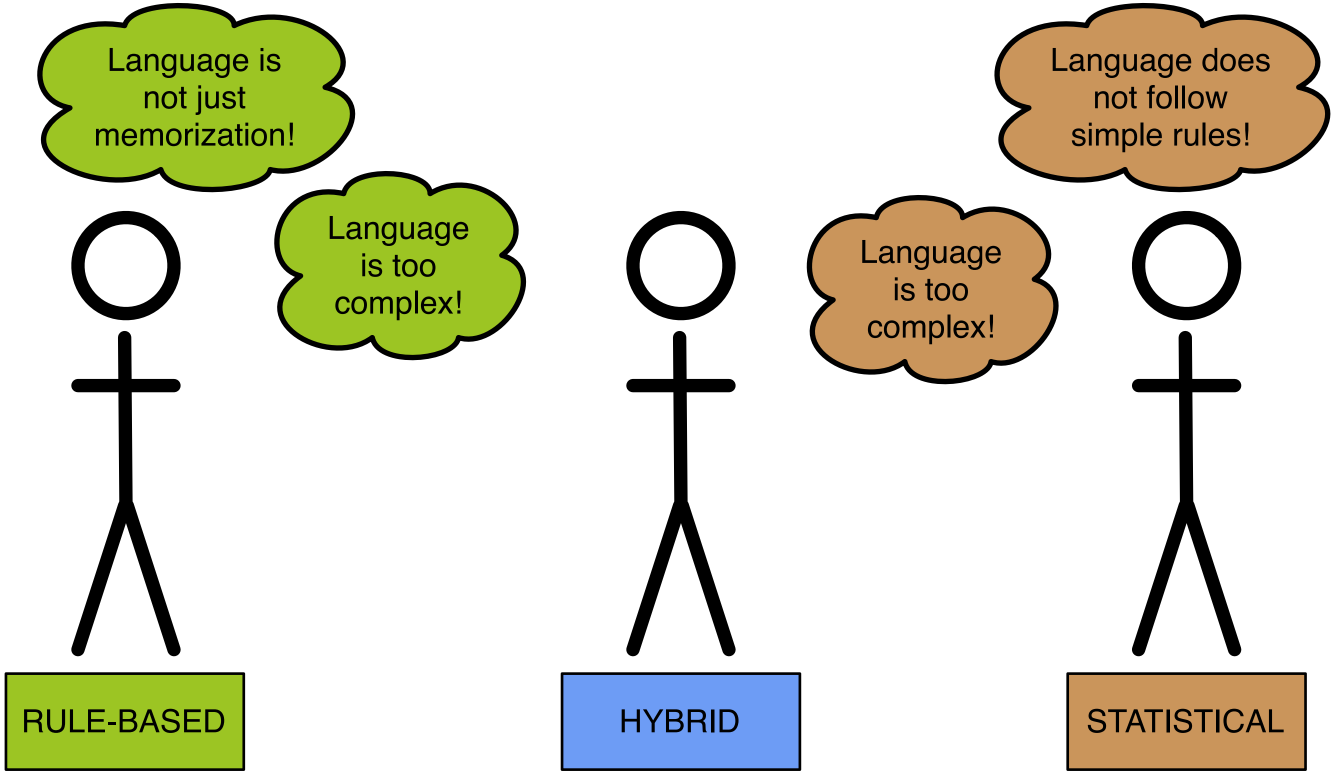
The Discussion

Language is not just memorization!
Language is too complex!
RULE-BASED
STATISTICAL

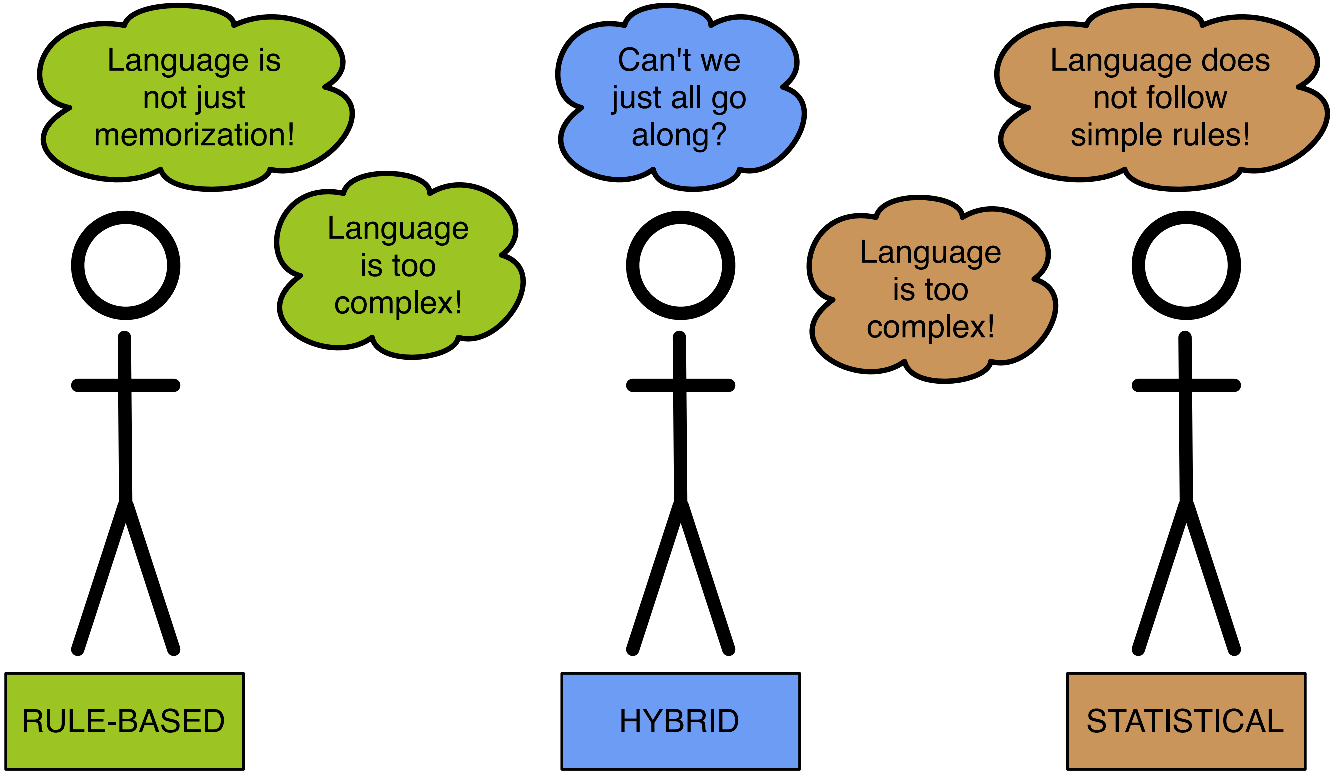
The Discussion


Language is not just memorization!
Language is too complex!
RULE-BASED
Language does not follow simple rules!
Language is too complex!
STATISTICAL

The Discussion



The Discussion


Language is not just memorization!
Language is too complex!
Can't we just all go along?
Language is too complex!
Language does not follow simple rules!
RULE-BASED
HYBRID
STATISTICAL

The case for knowlegde

- Consider this sentence:
  - German: Ich bin gestern von Baltimore nach Boston geflogen.
  - Gloss: I am yesterday from Baltimore to Boston flown.
- Reordering required
  - group verbal components together: bin ... geflogen
  - put the at the right place in the input sentence (after subject)
- Hard to do with a system that has no notion of verb, subject, etc.

The case for statistics

- German Sicherheit translates either as safety or security
- It is very hard to define the difference between safety and security
  - even harder to come up with rules that automatically make this distinction
- Statistical language models do a great job at using context to resolve this

Recent Developments

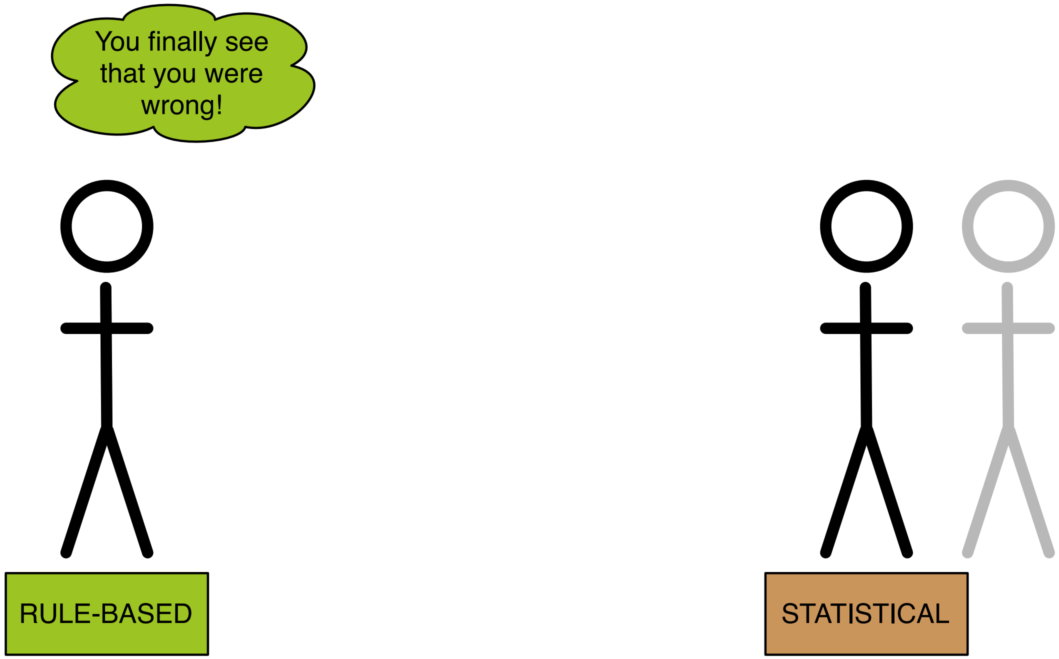
Recent Developments

RULE-BASED

STATISTICAL

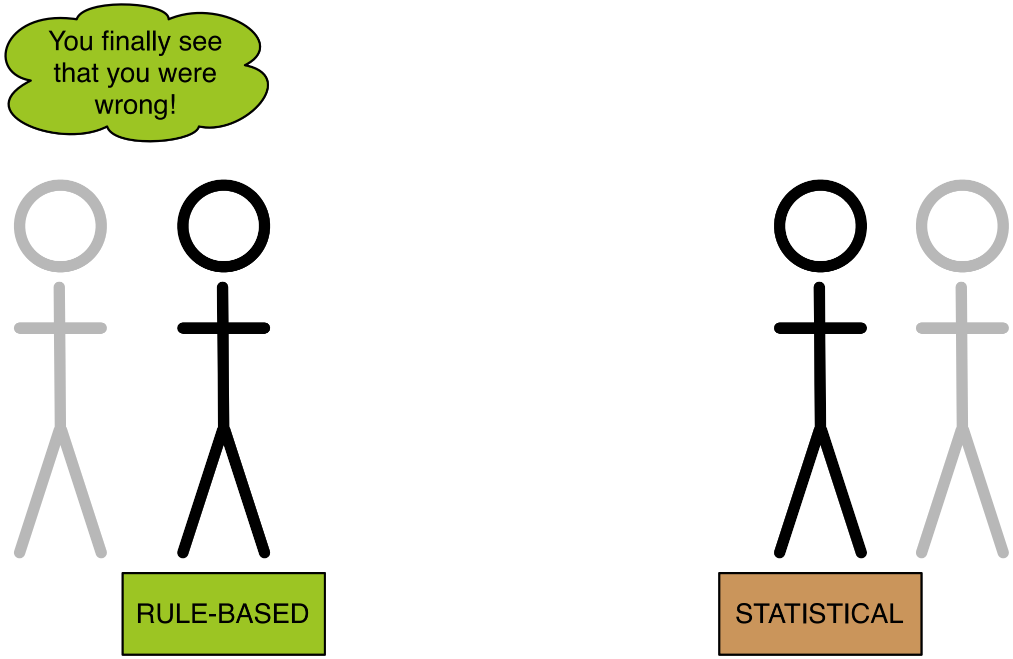
morphological analysis as pre-processing
syntactic reordering
tree-based and syntax-based models

Recent Developments



morphological analysis as pre-processing
syntactic reordering
tree-based and syntax-based models

Recent Developments



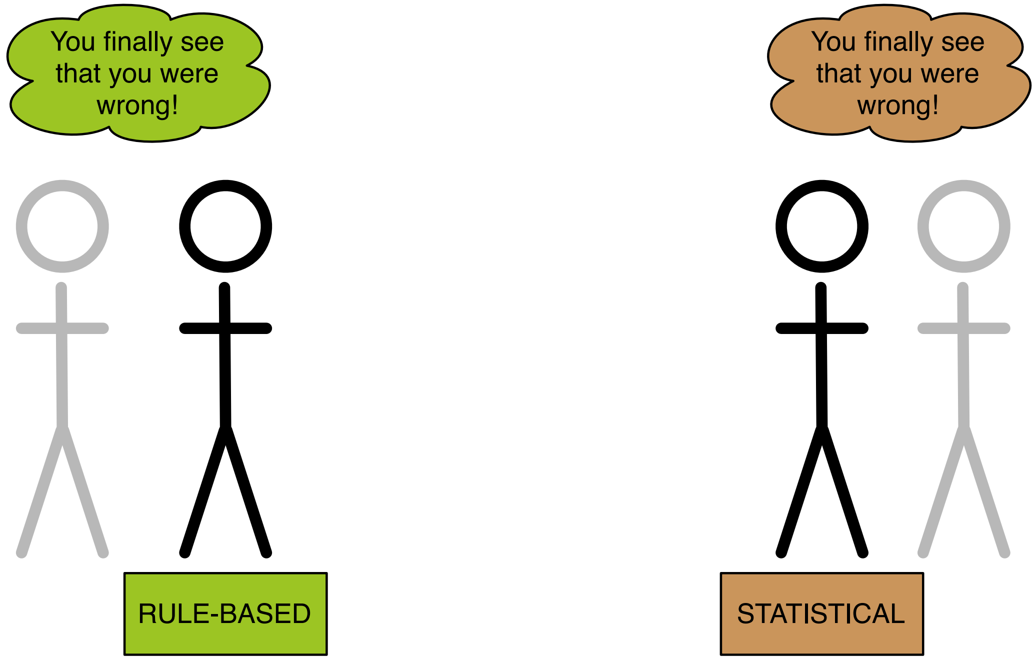
RULE-BASED

STATISTICAL

corpora for terminology acquisition
language models for disambiguation

morphological analysis as pre-processing
syntactic reordering
tree-based and syntax-based models

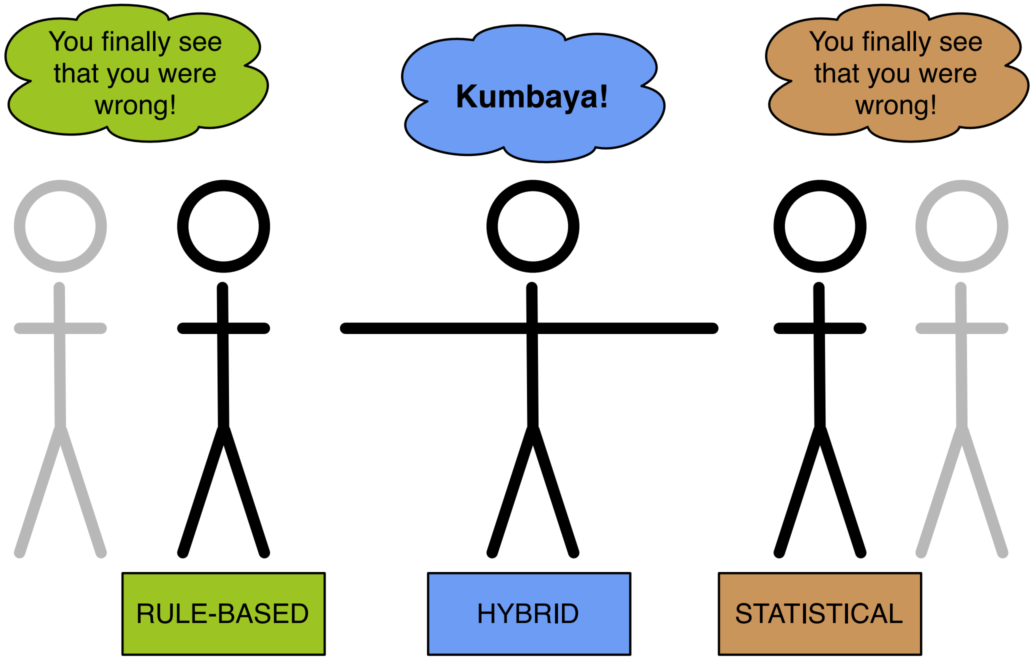
Recent Developments


You finally see that you were wrong!
You finally see that you were wrong!
RULE-BASED
STATISTICAL

corpora for terminology acquisition
language models for disambiguation

morphological analysis as pre-processing
syntactic reordering
tree-based and syntax-based models

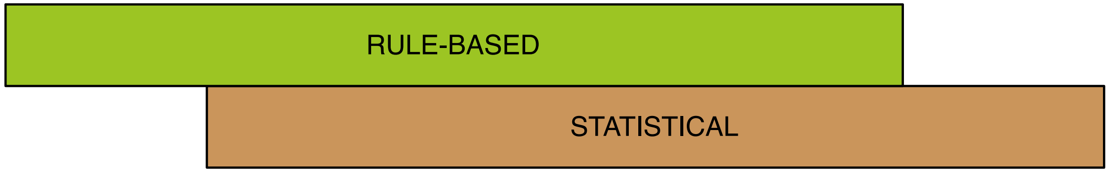
Recent Developments


You finally see that you were wrong!
Kumbaya!
You finally see that you were wrong!
RULE-BASED
HYBRID
STATISTICAL

corpora for terminology acquisition
language models for disambiguation

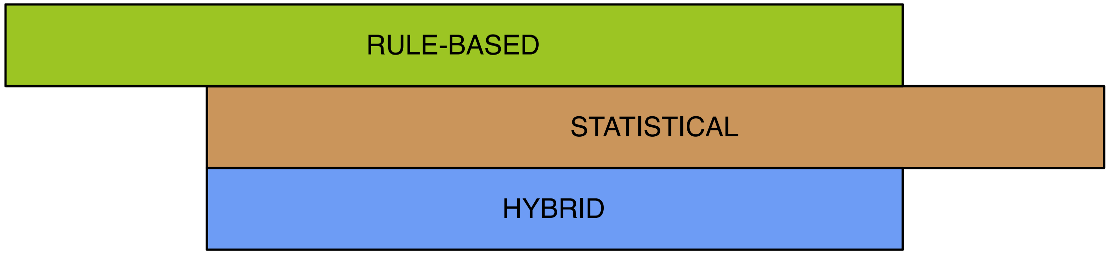
morphological analysis as pre-processing
syntactic reordering
tree-based and syntax-based models

Range of approaches



- What is statistical?
- What is rule-based?

Range of approaches



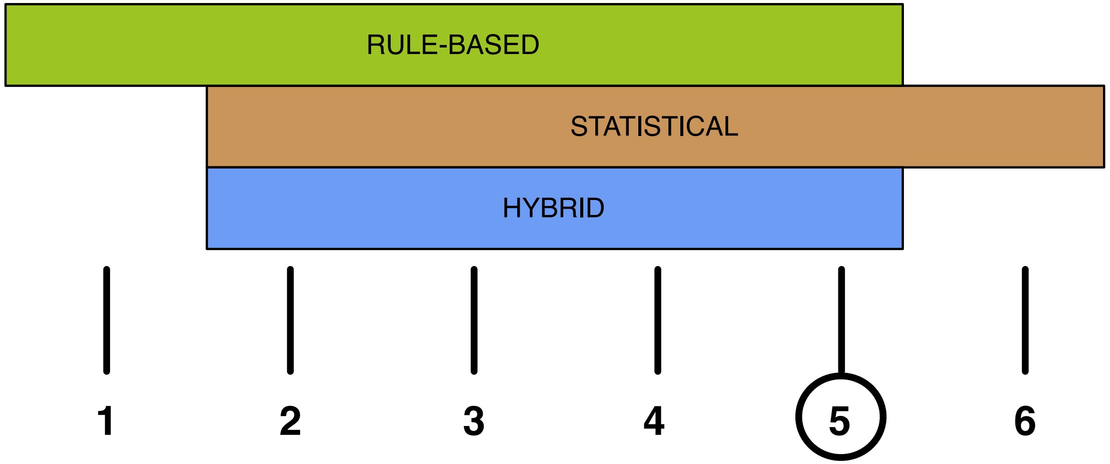
RULE-BASED

STATISTICAL

HYBRID

- What is hybrid?

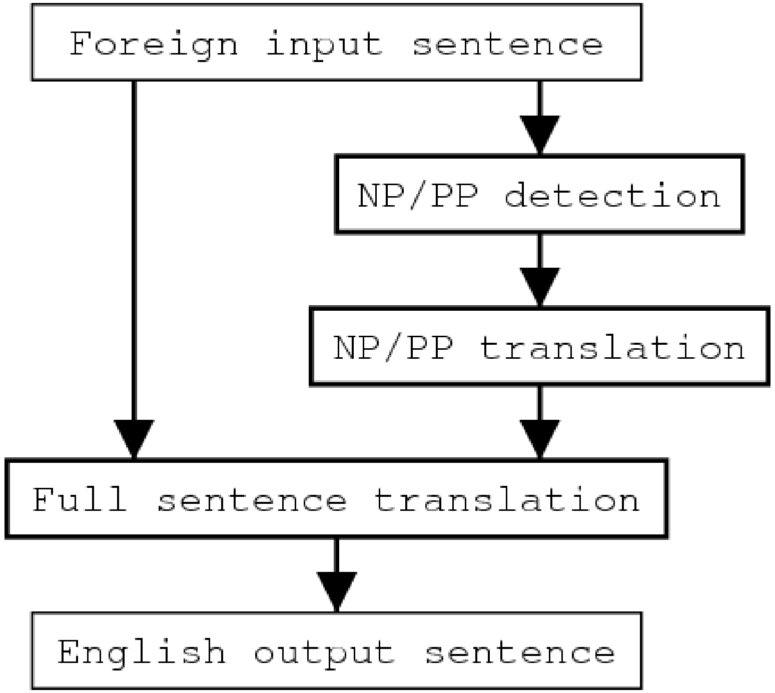
Hybrid scale



- Hybrid scale — I put myself on ”5”
  - starting with a statistical approach
  - linguistic concepts are useful

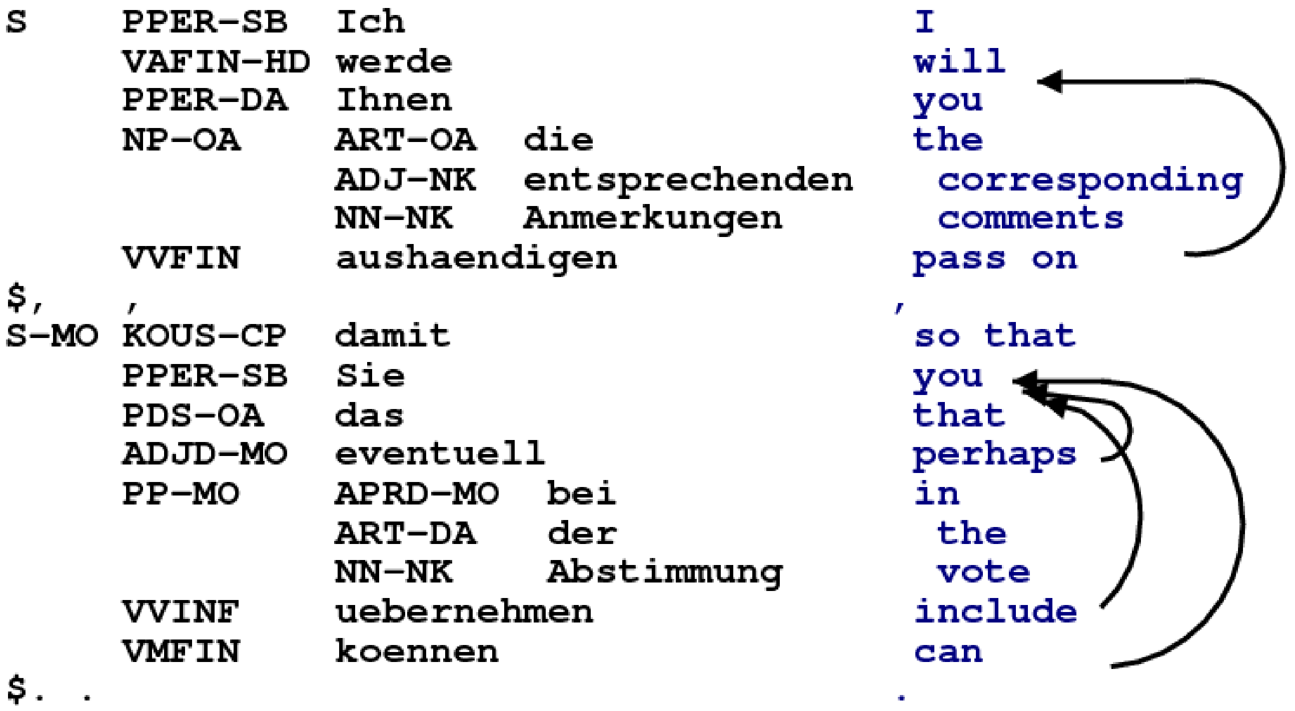
→ learn them from (annotated) data

Noun Phrase Translation



- Translate noun phrases and prepositional phrases in isolation [ACL 2003]
- Integrate **special features** (compound splitting, case agreement, etc.)

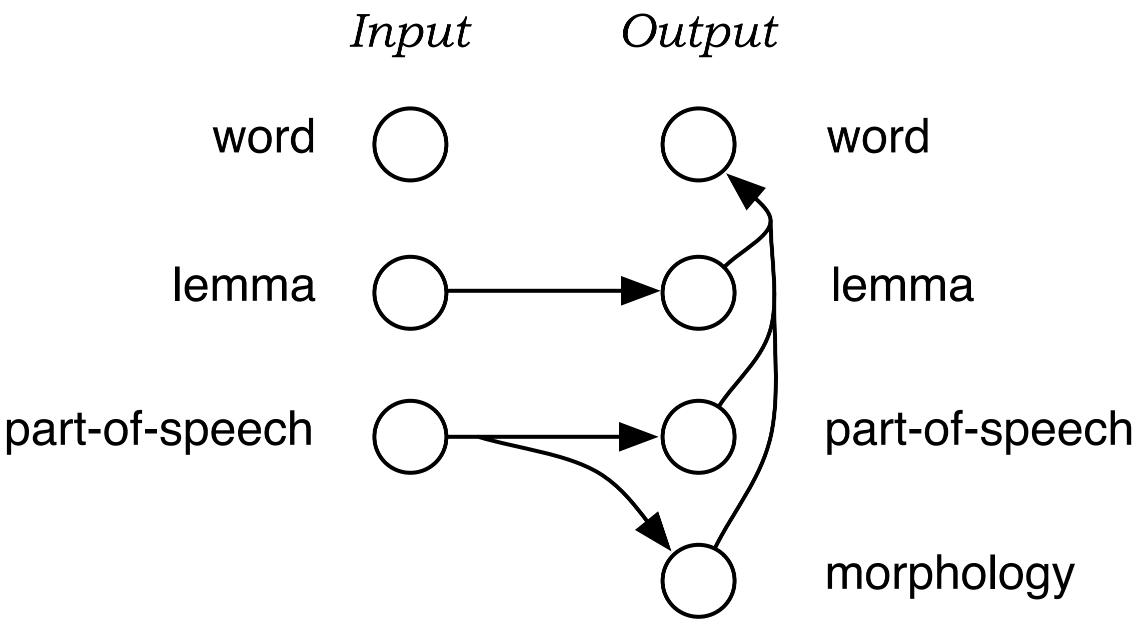
Clause restructuring



```
S     PPER-SB  Ich                                  I
      VAFIN-HD werde                                will
      PPER-DA  Ihnen                                you
      NP-OA    ART-OA  die                          the
               ADJ-NK  entsprechenden                corresponding
               NN-NK   Anmerkungen                   comments
      VVFIN    aushaendigen                         pass on
$,    ,                                            ,
S-MO  KOUS-CP  damit                                so that
      PPER-SB  Sie                                  you
      PDS-OA   das                                  that
      ADJD-MO  eventuell                            perhaps
      PP-MO    APRD-MO  bei                         in
               ART-DA   der                          the
               NN-NK    Abstimmung                   vote
      VVINF    uebernehmen                          include
      VMFIN    koennen                              can
$.  .                                              .
```

- Reorder German clause structure with **manual rules** [ACL 2005]
- Translate with SMT (positive results on German–English and Chinese–English)

Factored translation models



- Factored representation of words, breaking up the translation process into several **mapping** steps that **translate** or **generate** target factors
- JHU Summer Workshop 2006, available in **open source SMT toolkit Moses**