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 too complex!
Language is not just memorization!

RULE-BASED

STATISTICAL
The Discussion

Language is too complex!
Language does not follow simple rules!
Language is not just memorization!
Language is too complex!

RULE-BASED

STATISTICAL
The Discussion

Language is not just memorization!
Language is too complex!
Language does not follow simple rules!

RULE-BASED
HYBRID
STATISTICAL
The Discussion

Language is too complex!

Language does not follow simple rules!

Language is not just memorization!

Can't we just all go along?

Language is too complex!

RULE-BASED

HYBRID

STATISTICAL

Philipp Koehn, University of Edinburgh       AMTA Hybrid MT Panel 2006
11 August 2006
The case for knowledge

- 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

RULE-BASED

STATISTICAL

You finally see that you were wrong!
Recent Developments

**RULE-BASED**

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

**STATISTICAL**
Recent Developments

You finally see that you were wrong!

- Rule-based
- Statistical

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

You finally see that you were wrong!

RULE-BASED

corpora for terminology acquisition
language models for disambiguation

STATISTICAL

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

- Rule-based corpora for terminology acquisition
- Language models for disambiguation
- Morphological analysis as pre-processing
- Syntactic reordering
- Tree-based and syntax-based models

You finally see that you were wrong!
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

- What is hybrid?
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

- Reorder German clause structure with **manual rules** [ACL 2005]
- Translate with SMT (positive results on German–English and Chinese–English)
Factored representation of words, breaking up the translation process into several mapping steps that translate or generate target factors.