# Learning from Uncontrolled Corpora

Katharina Probst

May 21, 2003

# Introduction

Problem Definition:

- Given: parallel corpus, word alignments, a bilingual dictionary
- Goal: learn syntactic rules from English into a TL (as before)

Notes:

- The following slides represent what I tentatively plan to complete before the proposal, excluding the slide 'Future work'.
- The adjusted algorithms grew out of the circumstances of dealing with uncontrolled corpora. They will equally be applied to a controlled corpus.

## Challenges when learning from uncontrolled corpora

- No feature-tagged TL dictionary
- No compositional structure of training data
- Long sentences
- Possibly a very big transfer dictionary

## Tagging the TL dictionary

- Use English POS-tagger and alignments
- Penn treebank tagset is rich enough to tag some features
- Extract likely instances of plural nouns (/past tense verbs) in TL
- Assume that feature is marked at the beginning and/or end of the word
- Learn a classifier that can tag unseen instances

# Training data extraction

- Parse English sentence (Charniak parser)

- Traverse parse from top down, extract training example for the chunk rooted at each node

- For each training example, record

  1. features from POS tags
  2. corresponding TL chunk
  3. partial parse
  4. alignments

## Examples of extracted training instances

See file on avenue://usr0/kathrin/RuleLearning/
VersionIndepFiles/IOFiles/InFiles/EFTrainingFromUnc.rulelearning.txt

## Seed Generation revisted

- Introduce all possible constraints

- Lexical constraints only make sense if marked in the lexicon, e.g. if we introduce a constraint ((Y1 NUM) = (X2 NUM)), NUM must be marked in the TL lexicon

- Introduce head-passing constraints in SL, see head rule table in avenue://usr0/kathrin/RuleLearning/HeadFinding/EnglishHeadTable.txt

# Seed Generation revisted (II)

- Examples of extracted training examples with English heads marked: avenue://usr0/kathrin/RuleLearning /VersionIndepFiles/IOFiles/InFiles /EFTrainingFromUnc.slheads.rulelearning.txt

- Try to infer head in TL phrases, also pass up features up from head (may be future work)

- Also pass up features from non-heads, e.g. definiteness from determiners to NPs (future work)

## Compositionality revisited

- Compositionality has been moved to be exclusively context-free $\rightarrow$ constraints are ignored
- This allows the system to give the transfer engine only the lexical entries it needs for a phrase/sentence
- This step now learns the context-free backbone of the transfer grammar, esp. important if few features are given
- Examples of compositional (context-free) rules: avenue://usr0/kathrin/RuleLearning/VersionIndepFiles/IOFiles/OutFiles/FinalGrammar.txt

## Seeded Version Space learning revisited

- With constraints introduced during Seed Generation, SVS starts with more specific rules

- Ex. of a non-safe generalization: deletion of constraint

- Ex. of a safe generalization: merging of two value constraints to an agreement constraint: ((Y3 NUM) = PL), ((Y4 NUM) = PL) $\rightarrow$ ((Y3 NUM) = (Y4 NUM))

- Only allow for 'safe' generalizations, non-safe generalizations are moved to the run-time system (future work)

## Run-time system

- English parse is traversed top-down
- At each node, translation is attempted
- Empty translation ($\rightarrow$ skipping) allowed
- Partial translations are combined
- Currently no way to distinguish between translations (future work)

## Future work

- TL dictionary feature tagging expanded
- Run-time system revisited:
  - Incorporate rule scores, lexical translation probabilities
  - Incorporate probabilities of zero translations
  - Violation of constraints with penalty (transfer engine?)
- Evaluation of algorithms

## Conclusion

- Learning from uncontrolled corpus as important extension to system

- Automatically inferred TL feature values

- Stronger burden on run-time system