Combining Interlingua with SMT

Stephanie Seneff

CSAIL, MIT

Panel Discussion

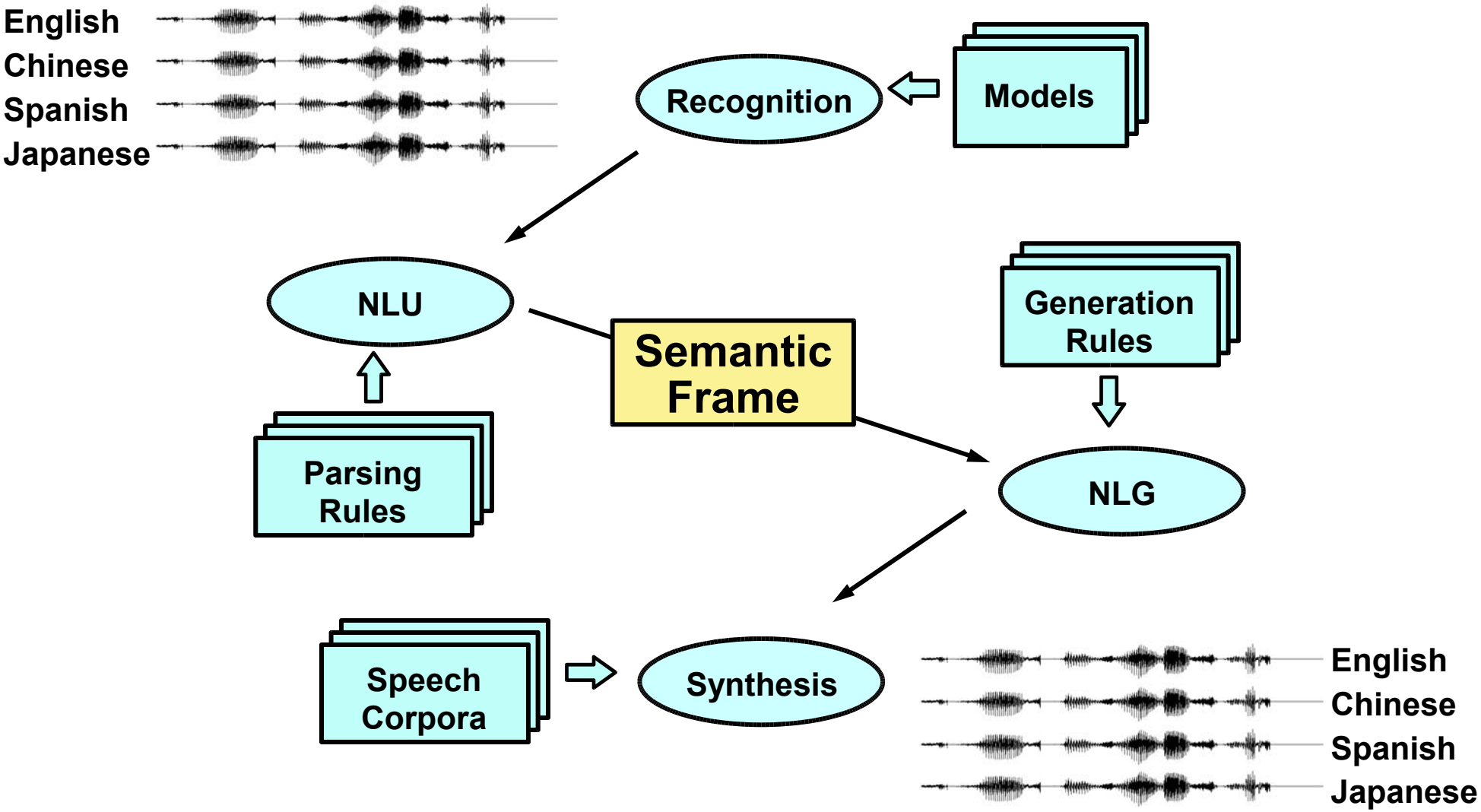
August 11, 2006

Introduction

- Our group at MIT has been developing multilingual spoken dialogue systems since the late 1980's
  - Common meaning representation for dialogue manager
  - Many different languages: English, French, Spanish, German, Italian, Chinese, Japanese, Korean
  - Many different domains: flights, weather, restaurants, hotels, calendar management, etc.
- Recent research has focused on dialogue interaction to learn a second language
- Spoken language translation can assist student:
  - Translation quality must be near perfect
  - Narrow domain makes it feasible
- This is a very different problem from general language text translation

Multilingual Spoken Translation Framework

Common meaning representation: *semantic frame*



Semantic Frame as Interlingua

- Our representation captures syntactic structure but discards temporal word order
- Decompose syntax-directed translation into two steps
  - Syntactic order to hierarchy
  - Hierarchy to syntactic order
- This greatly reduces the rule space

Some Thoughts

- Interlingual approach appears daunting (out of reach?) for general language text translation task
- Speech-based translation is necessarily domain-restricted due to speech recognition constraints
- Multilingual spoken dialogue systems are symbiotic with interlingual translation
  - Map all language inputs to common meaning representation
  - Provide large corpora of spoken utterances for training
- Proposal:
  - Pursue interlingual approach within restricted domains of existing conversational systems
  - Construct common grammar and generation rules for all domains
  - Seek generalities wherever feasible to reduce required expertise

Some Requirements for the Parsing and Generation Components

- Manually created lexicalized grammar capturing syntactic structure, developed by linguists
  - Include mechanism to address movement
  - Strong probability model
  - Automatic training methods
  - Simple rules to map to hierarchical semantic frame
- Generation system capable of producing multiple hypotheses and/or selecting for word senses based on context-conditioned probability model
  - Context conditions specified through hierarchical locality

Ways to Combine Linguistic and Statistical Methods

- Use statistical *n*-grams to post-select from multiple generation hypotheses
- Use SMT-assisted interactive tools to support grammar and lexicon development
- Use a statistical parser to post-select from multiple SMT outputs
- Use SMT as a back-up method upon parse failure
- Combine statistical alignment with parsing to seed translation lexicon
  (and to seed grammar for new language??)