

Title	Semantic Parsing: transforming sentences to logical forms using machine learning models
Author(s)	Nguyen, Minh Le
Citation	
Issue Date	2007-03-07
Type	Presentation
Text version	publisher
URL	<a href="http://hdl.handle.net/10119/8297">http://hdl.handle.net/10119/8297</a>
Rights	
Description	4th VERITE : JAIST/TRUST-AIST/CVS joint workshop on VERification TEchnologyでの発表資料, 開催 : 2007年3月6日 ~ 3月7日, 開催場所 : 北陸先端科学技術大学院大学・知識講義棟 2 階中講義室

## Semantic Parsing: transforming sentences to logical forms using machine learning models

Minh Le Nguyen

School of Information Science  
Japan Advanced Institute of Science and Technology

COE '07

1

## Syntactic and Semantic Natural Language Learning

- Most computational research in natural-language learning has addressed "low-level" syntactic processing.
  - Morphology (e.g. past-tense generation)
  - Part-of-speech tagging
  - Chunking
  - Syntactic parsing
- Learning for semantic analysis has been restricted to relatively "shallow" meaning representations.
  - Word sense disambiguation (e.g. SENSEVAL)
  - Semantic role assignment (determining agent, patient, instrument, etc., e.g. FrameNet, PropBank)
  - Information extraction

COE '07

2

## Semantic Parsing

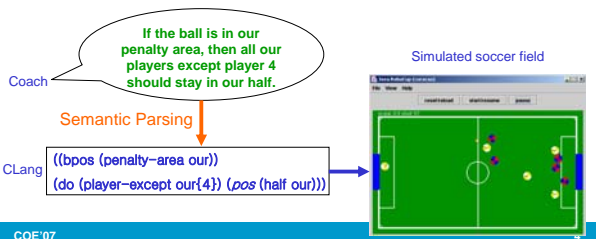
- Semantic parsing is the process of mapping a natural-language sentence to a complete, detailed semantic representation: **logical form** or **meaning representation (MR)**.
- For many applications, the desired output is immediately executable by another program.
- Application domains:
  - CLang: RoboCup Coach Language
  - GeoQuery: A Database Query Application
  - Legal domain

COE '07

3

## CLang: RoboCup Coach Language

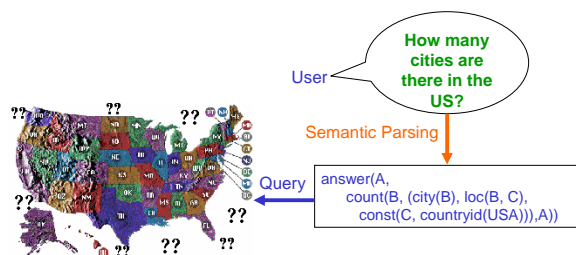
- In RoboCup Coach competition teams compete to coach simulated players
- The coaching instructions are given in a formal language called CLang



COE '07

## GeoQuery: A Database Query Application

- Query application for U.S. geography database containing about 800 facts [Zelle & Mooney, 1996]



COE '07

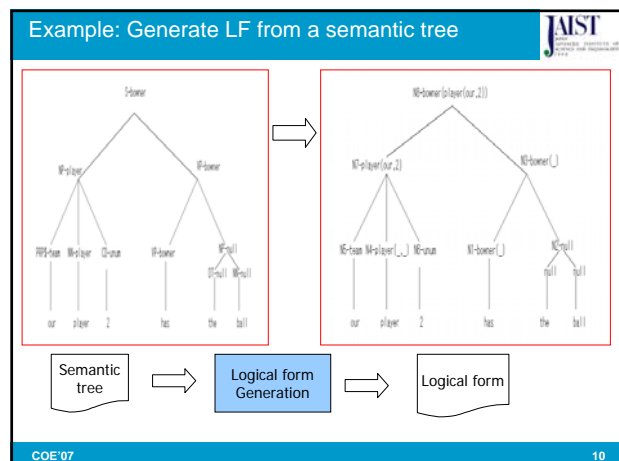
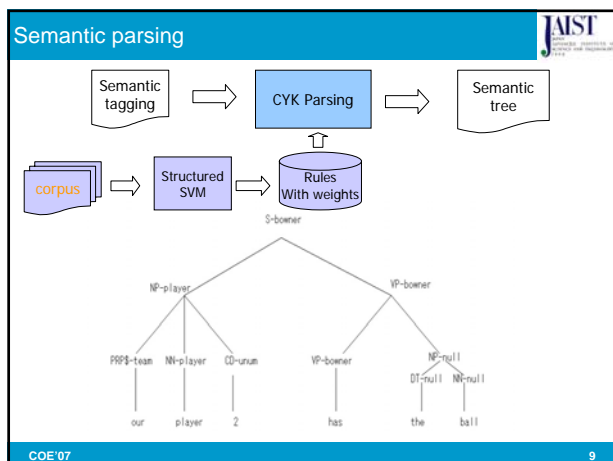
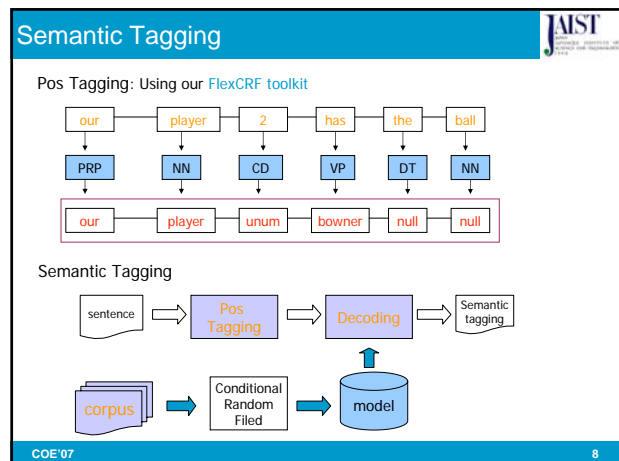
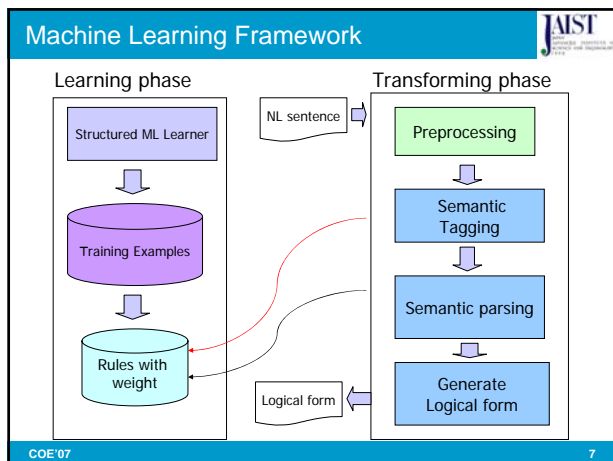
5

## Approach

- Applying ML to the transforming problem
- Motivations
  - Robustness, reduction of development rules
  - Treating ambiguity
  - Handling with the difficulty of consistent rules
- Current work
  - ML for query database language / robocup controlled language
  - ML for legal domain

COE '07

6



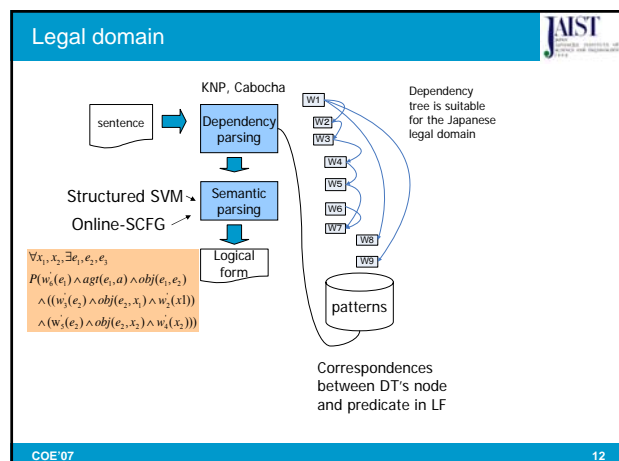
### Current Work

- English Data (CLANG)
  - Structured SVM (Robocup)
    - Precision: 85%
    - Recall: 74%
  - Maximum entropy model (DB query)
    - Precision 89%
    - Recall 51%
- Japanese Data
  - Splitting long sentences into a set of short sentences
  - Mapping NL Japanese sentence to logical form

$$precision = \frac{\#correct-representation}{\#completed-representation}$$

$$recall = \frac{\#correct-representation}{\#sentences}$$

COE '07 11



## Dependency Parsing

Sentence → Dependency Parsing → Dependency tree

corpus → Online Structured Learning → Dependency Model

Japanese unlabeled accuracy 93.9%  
Japanese unlabeled accuracy 91.6%

very good results in shared task CONLL-2006 including Japanese data

State of the art result in English data (Penn III)  
English unlabeled accuracy 91.6%  
English unlabeled accuracy 90.7%

Plan to participate CONLL-2007 task

COE'07 13

## Patterns learning

- Input: set of dependency trees and their logical forms
- Output: the correspondences of each node in DT with a predicate in LF
- Method: Using statistical machine translation to align a node in DT to a predicate in LF

Example:

W1 W2 W3 W4 W5 W6 W7 W8 W9

COE'07 14

## Splitting clauses in legal domain

- Using machine learning for splitting clauses

NL → splitting → mapping → LF → combining → LF

zenjou ni gaitou suru mono ga aru toki ha, kuchou ha, kore o kokuhatsu suru mono to suru  
前条に該当する者があるときは、区長は、これを告発するものとする。

前条に該当する者があるときは、区長はこれを告発するものとする。

- Collected 108 sentences and their logical form
- 9/10 for training and 1/10 for testing data
- The accuracy of the model is 93.10%!

COE'07 15

## Online-SCFG Methods

- Preprocessing
  - Generate a sequence of word tokens
  - Transform a logical form representation into a sequence of atomic logical form.
- Using GIZA++ to generate alignment between each word in NL to each token in LF
- Using synchronous grammar to estimate the model for generating logical form
- Using online structured prediction learning to estimate the SCFG grammar
- Some issues for Japanese data
  - Require a formal grammar representation for LF
  - In the case there is no formal grammar it becomes phrase based SMT models

COE'07 16

## Context-Free Semantic Grammar

QUERY → What is CITY

CITY → the capital CITY

CITY → of STATE

STATE → Ohio

QUERY  
What is CITY  
the capital CITY  
of STATE  
Ohio

COE'07 17

## Synchronous Context-Free Grammars (SCFG)

- Developed by Aho & Ullman (1972) as a theory of compilers that combines *syntax analysis* and *code generation* in a single phase
- Generates a pair of strings in a single derivation

COE'07 18

## Synchronous Context-Free Grammars

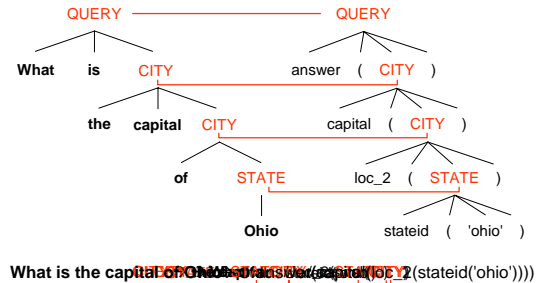
Developed by Aho & Ullman (1972) as a theory of compilers that combines *syntax analysis* and *code generation* in a single phase  
Generates a pair of strings in a single derivation

QUERY  $\rightarrow$  What is CITY / answer(CITY)

COE'07

19

## Synchronous Context-Free Grammars

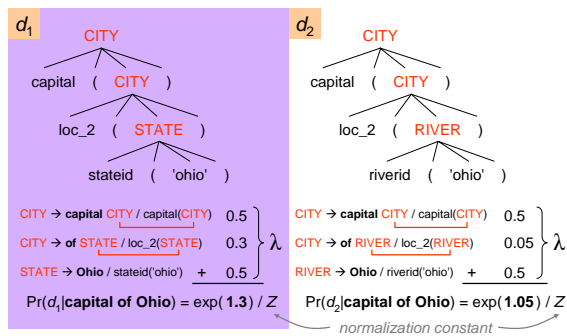


What is the capital of Ohio? / answer(capital(loc\_2(stateid('ohio'))))

COE'07

20

## Probabilistic Parsing Model



COE'07

21

## Parsing Model

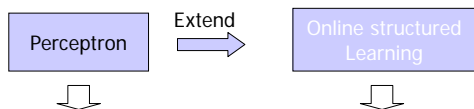
- $N$  (non-terminals) = {QUERY, CITY, STATE, ...}
  - $S$  (start symbol) = QUERY
  - $T_m$  (MRL terminals) = {answer, capital, loc\_2, (, ), ...}
  - $T_n$  (NL words) = {What, is, the, capital, of, Ohio, ...}
- QUERY  $\rightarrow$  What is CITY / answer(CITY)  
 CITY  $\rightarrow$  the capital CITY / capital(CITY)  
 CITY  $\rightarrow$  of STATE / loc\_2(STATE)  
 STATE  $\rightarrow$  Ohio / stateid('ohio')
- $\lambda$  (parameters of probabilistic model) = ?
- Online structured prediction learning

COE'07

22

## Online structured prediction learning

- Extend the traditional Perceptron learning



Two class labels learning problem    Many class labels, tree structure

COE'07

23

## Results

- English data (CLANG)
  - It is applicable for Robocup language
    - Precision 89.5 (best precision)
    - Recall 61.2
  - The result is good because we do not need fully semantic tree annotation
- Japanese data (110 sentences)
  - Because of spare data problem so the alignment of each word in NL sentence and each token in LF is not good.
  - It is need to verify this problem in detail for improving the accuracy of our model
    - Enlarge the training data ?

COE'07

24

## Conclusions



- Learning is applicable for transforming NL to logical form
- The number of training data should be enlarged to make sure the accuracy of the models
- The splitting result for legal Japanese data is attractive



We should integrate this model  
with the rule-based model

COE'07

25

## Future work



- Experiment on a larger corpus of Japanese data
- Integrate with the rule-based models
- Transforming a logical form representation to a NL sentence
- Semi-supervised learning models for semantic parsing

COE'07

26