Special Topics in Computer Science

NLP in a Nutshell

CS492B Spring Semester 2009

Jong C. Park

Computer Science Department

Korea Advanced Institute of Science and Technology

INTRODUCTION

Objectives

- Natural Language Processing (NLP) is a discipline that takes a semester or two to appreciate the full spectrum of its implications.
- In this course, however, we will spend a minimal amount of time for the introductory materials and jump into real-world applications,
- focusing on text mining and human-robot interface, for the students to get a hands-on experience as early as possible and with meaningful results by the end of the semester.

Administrative Details

- Instructor
  - Jong C. Park
  - Email: park@cs.kaist.ac.kr
  - Office: CS Bldg. Room 2406
  - Phone: x3541
- Teaching Assistants
  - Seung-Cheol Baek and Hee-Jin Lee
  - Email: cs492@nlp.kaist.ac.kr
  - Phone: x3581

Administrative Details

- Homepage
  - http://nlp.kaist.ac.kr/~cs492
- Time and Place
  - Tuesdays and Thursdays
  - 1pm~2:20pm
  - CS Bldg. Room 2443

Course Plan

- Textbooks
- Textbook Materials
- Software Review
- Project
- Feedback Plan
- Weekly Schedule

Textbooks

- Primary
  - An Introduction to Language Processing with Perl and Prolog: An Outline of Theories, Implementation, and Application with Special Consideration of English, French, and German, Pierre M. Nugues, Springer, 2006. http://www.cs.lth.se/~pierre/ilppp/
- Secondary
  - Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Daniel Jurafsky & James H. Martin, Pearson, 2nd edition, 2009.

Textbook Materials [chapters]

- An Overview of Language Processing
- Corpus Processing Tools
- Encoding, Entropy, and Annotation Schemes
- Counting Words
- Words, Parts of Speech, and Morphology
- Part-of-Speech Tagging Using Rules
- Part-of-Speech Tagging Using Stochastic Techniques

Textbook Materials [chapters]

- Phrase-Structure Grammars in Prolog
- Partial Parsing
- Syntactic Formalisms
- Parsing Techniques
- Semantics and Predicate Logic
- Lexical Semantics
- Discourse
- Dialogue

Textbook Materials [software]

- An Overview of Language Processing
  - The German Institute for Artificial Intelligence Research http://registry.dfki.de
  - The Oxford Text Archive http://ota.ox.ac.uk
  - The Linguistic Data Consortium of the University of Pennsylvania http://www.ldc.upenn.edu
  - The European Language Resources Association http://www.elra.info
  - Peedy at the Microsoft Research Web site http://www.research.microsoft.com

Textbook Materials [software]

- Corpus Processing Tools
  - Finite-State Automata in Prolog (p29~30)
  - Concordances in Prolog (p46~47)
  - Concordances in Perl (p48~50)
  - The Minimum Edit Distance in Perl (p53~54)
  - Searching Edits in Prolog (p54)
  - The FSA Utilities
    http://odur.let.rug.nl/~vannoord/Fsa
  - The FSM Library
    http://www.research.att.com/sw/tools/fsm

Textbook Materials [software]

- Encoding, Entropy, and Annotation Schemes
  - IBM's Library of Unicode Components in Java and C++ http://www.ibm.com/software/globalization/icu

Textbook Materials [software]

- Counting Words
  - Tokenizing Texts in Prolog (p89~91)
  - Tokenizing Texts in Perl (p91)
  - Counting Unigrams in Prolog (p93)
  - Counting Unigrams and Bigrams with Perl (p93~95)
  - Extracting Collocations with Perl (p108~109)
  - The SRI Language Modeling Collection http://www.speech.sri.com
  - The CMU-Cambridge Statistical Language Modeling Toolkit http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html

Textbook Materials [software]

- Words, Parts of Speech, and Morphology
  - Building a Trie in Prolog (p121~123)
  - Finding a Word in a Trie in Prolog (p123)
  - Finite-State Transducers in Prolog (p134~136)
  - PC-KIMMO 2 http://www.sil.org
  - General-Purpose Finite-State Transducer Toolkits http://www-igm.univ-mlv.fr/~unitex
  - Xerox XFST http://www.xrce.xerox.com/competencies/content-analysis/fst/

Textbook Materials [software]

- Part-of-Speech Tagging Using Rules
  - Tagging in Prolog (p151~153)
- Part-of-Speech Tagging Using Stochastic Techniques
  - GIZA++ http://www.fjoch.com/GIZA++.html

Textbook Materials [software]

- Phrase-Structure Grammars in Prolog
  - Tokenizing Texts Using DCG Rules (p200~202)
- Partial Parsing
  - (Simplified) ELIZA in Prolog (p215~216)
  - Multiword Detection using DCG Rules (p219~221)
  - Detecting Noun Groups using DCG Rules (p223~224)
  - Detecting Verb Groups using DCG Rules (p225~227)
  - Detecting Prepositional Groups using DCG Rules (p232~p234)

Textbook Materials [software]

- Syntactic Formalisms
  - Unification of Feature Structures in Prolog (p263)
- Parsing Techniques
  - Shift-Reduce Parsing in Prolog (p279~281)
  - Earley Parsing in Prolog (p288~294)
  - CYK Parsing in Prolog (p298~300)
  - Nivre's Parser in Prolog (p306~309)
  - Constraint Handling Rules (available) in SWI Prolog

Textbook Materials [software]

- Semantics and Predicate Logic
- Lexical Semantics
- Discourse
- Dialogue

Software Review

- Students will make presentations in groups.
- Topics for student presentations will be on:
  - Software modules from the primary textbook
  - Software packages from the Internet
- Each presentation should include:
  - Basic concepts
  - Usage of software modules or packages
  - Pros and cons of such modules or packages

Project

- Students may form groups to propose and work on projects related to…
  - Human-Robot Interaction
  - Text Mining
- You can choose your own programming language.
  - Use Prolog (or Perl) and enhance the functions of the software modules in the textbook.
  - Use other languages (such as Python or Java) and combine modules from the available packages.

Feedback Plan

- Evaluation
  - Attendance: 15%
  - Presentation: 30%
  - Programming Assignments: 15%
  - Project: 40%

Weekly Schedule (1/3)

Week	Lecture Materials	Software Review	Homework
1	Introduction: An Overview of Language Processing		Read: Intro to Prolog [pdf]
2	Corpus Processing Tools; Encoding, …	An Overview of Language Processing	Prolog Programming Assignment 1
3	Counting Words; Words, POS, …	Corpus Processing Tools; Encoding, …	
4	POS Tagging …; PSGs in Prolog	Counting Words; Words, POS, …	Prolog Programming Assignment 2
5	Partial Parsing	POS Tagging …; PSGs in Prolog	
6	Syntactic Formalisms; Parsing Techniques	Partial Parsing	Project Ideas (1 page)

Weekly Schedule (2/3)

Week	Lecture Materials	Software Review	Homework
7		Syntactic Formalisms; Parsing Techniques	Project Proposal
8	Midterm Week		[no exam]
9	Text Mining Applications: Resources	[Project Proposal : Presentation]	
10	Text Mining Applications: Terminology and NER	Text Mining Applications: Resources	
11	Text Mining Applications: Extraction and Annotation	Text Mining Applications: Terminology and NER	Project Interim Demo
12	[Project Interim Demo]	Text Mining Applications: Extraction and Annotation	

Weekly Schedule (3/3)

Week	Lecture Materials	Software Review	Homework
13	HRI Applications: Speech Synthesis		
14	HRI Applications: Robots with Emotion	HRI Applications: Speech Synthesis	Project Final Demo
15	[Project Final Demo]	HRI Applications: Robots with Emotion	Final Report
16	Final Exam Week		[no exam]

Reading Assignment

- An Introduction to Prolog
  - PDF file available at http://www.cs.lth.se/~pierre/ilppp
- SWI Prolog
  - http://www.swi-prolog.org