Information Extraction
ICS 482 Natural Language Processing

Lecture 23: Information Extraction

Husni Al-Muhtaseb

بسم الله الرحمن الرحيم

ICS 482 Natural Language Processing

Lecture 23: Information Extraction

Husni Al-Muhtaseb

NLP Credits and Acknowledgment

These slides were adapted from presentations of the Authors of the book

SPEECH and LANGUAGE PROCESSING:

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

and some modifications from presentations found in the WEB by several scholars including the following

NLP Credits and Acknowledgment

If your name is missing please contact me
muhtaseb
At
Kfupm.
Edu.
sa

NLP Credits and Acknowledgment

Husni Al-Muhtaseb
James Martin
Jim Martin
Dan Jurafsky
Sandiway Fong
Song young in
Paula Matuszek
Mary-Angela Papalaskari
Dick Crouch
Tracy Kin
L. Venkata Subramaniam
Martin Volk
Bruce R. Maxim
Jan Hajič
Srinath Srinivasa
Simeon Ntafos
Paolo Pirjanian
Ricardo Vilalta
Tom Lenaerts

Heshaam Feili
Björn Gambäck
Christian Korthals
Thomas G. Dietterich
Devika Subramanian
Duminda Wijesekera
Lee McCluskey
David J. Kriegman
Kathleen McKeown
Michael J. Ciaraldi
David Finkel
Min-Yen Kan
Andreas Geyer-Schulz
Franz J. Kurfess
Tim Finin
Nadjet Bouayad
Kathy McCoy
Hans Uszkoreit
Azadeh Maghsoodi

Khurshid Ahmad
Staffan Larsson
Robert Wilensky
Feiyu Xu
Jakub Piskorski
Rohini Srihari
Mark Sanderson
Andrew Elks
Marc Davis
Ray Larson
Jimmy Lin
Marti Hearst
Andrew McCallum
Nick Kushmerick
Mark Craven
Chia-Hui Chang
Diana Maynard
James Allan

Martha Palmer
julia hirschberg
Elaine Rich
Christof Monz
Bonnie J. Dorr
Nizar Habash
Massimo Poesio
David Goss-Grubbs
Thomas K Harris
John Hutchins
Alexandros Potamianos
Mike Rosner
Latifa Al-Sulaiti
Giorgio Satta
Jerry R. Hobbs
Christopher Manning
Hinrich Schütze
Alexander Gelbukh
Gina-Anne Levow
Guitao Gao
Qing Ma
Zeynep Altan

Previous Lectures

- Introduction and Phases of an NLP system
- NLP Applications - Chatting with Alice
- Finite State Automata & Regular Expressions & languages
- Morphology: Inflectional & Derivational
- Parsing and Finite State Transducers, Porter Stemmer
- Statistical NLP – Language Modeling
- N Grams, Smoothing
- Parts of Speech - Arabic Parts of Speech
- Syntax: Context Free Grammar (CFG) & Parsing
- Parsing: Earley's Algorithm
- Probabilistic Parsing
- Probabilistic CYK - Dependency Grammar
- Semantics: Representing meaning - FOPC
- Lexicons and Morphology – invited lecture
- Semantics: Representing meaning
- Semantic Analysis: Syntactic-Driven Semantic Analysis

Today's Lecture

- Semantic Grammars
- Information Extraction Techniques
- A Problem to Solve
- First Presentation
  - Saleh Al-Zaid - Language Model Based Arabic Word Segmentation

Semantic Grammars

- An alternative to taking syntactic grammars and trying to map them to semantic representations is defining grammars specifically in terms of the semantic information we want to extract
  - Domain specific: Rules correspond directly to entities and activities in the domain

  I want to go from Dammam to Jeddah on Tuesday, May 2nd 2006
  - TripRequest → Need-spec travel-verb from City to City on Date
  - …

Predicting User Input

- Semantic grammars rely upon knowledge of the task and (sometimes) constraints on what the user can do when
  - Allows them to handle very sophisticated phenomena

I want to go to Jeddah on Tuesday.

I want to leave from there on Tuesday for Riyadh.

TripRequest → Need-spec travel-verb from City on Date for City

Drawbacks of Semantic Grammars

- Lack of generality
  - A new one for each application
  - Large cost in development time
- Can be very large, depending on how much coverage you want it to have
- If users go outside the grammar, things may break disastrously

  I want to go shopping.

  I want to leave from my house.

Information Extraction

- Idea is to ‘extract’ particular types of information from arbitrary text or transcribed speech
- Examples:
  - Names entities: people, places, organization
  - Telephone numbers
  - Dates
- Many uses:
  - Question answering systems, fisting of news or mail…
  - Job ads, financial information, terrorist attacks

Information Extraction

- Appropriate where Semantic Grammars and Syntactic Parsers are Not
  - Input too complex and far-ranging to build semantic grammars
  - But complete syntactic parsers are impractical
    - Too much ambiguity for arbitrary text
    - 50 parses or none at all
    - Too slow for real-time applications

Information Extraction Techniques

- Often use a set of simple templates or frames with slots to be filled in from input text
  - Ignore everything else
  - Husni's number is 966-3-860-2624.
  - The inventor of the First plane was Abbas ibnu Fernas
  - The British King died in March of 1932.
- Context (neighboring words, capitalization, punctuation) provides cues to help fill in the appropriate slots

The IE Process

- Given a corpus and a target set of items to be extracted:
  - Clean up the corpus
  - Tokenize it
  - Do some hand labeling of target items
  - Extract some simple features
    - POS tags
    - Phrase Chunks …
  - Do some machine learning to associate features with target items or derive this associate by intuition
  - Use e.g. FSTs, simple or cascaded to iteratively annotate the input, eventually identifying the slot fillers

A Problem to Solve

- Given a list of links to English newspapers/ sites, find all pages that are talking about Saudi Arabia
- Group as teams and suggest a high level procedure to solve this problem in 7 minutes
- Let Us Discuss it

Students Presentations

- Evaluation at WebCT
- First Presentation

Thank you

- السلام عليكم ورحمة الله