# Introduction to Natural Language Processing: L100a

## Lecture 1: Introduction

Ann Copestake

Computer Laboratory
University of Cambridge

October 2012

# Outline of today's lecture

Some NLP applications

NLP and linguistics

Course structure

Examples of analysis

To do list

# Question answering

IBM Watson: Jeopardy!

Category: US cities. Its largest airport was named for a World War II hero; its second largest, for a World War II battle.

Answer: Chicago

# Question answering

IBM Watson: Jeopardy!

Category: US cities. Its largest airport was named for a World War II hero; its second largest, for a World War II battle.

Answer: Chicago

# Analysis of scientific text

The objective of this research was to evaluate techniques for the rapid detection of chromosomal alterations occurring in humans exposed to butadiene.

e.g., CRAB project, Korhonen.

## Assessing and teaching English

- Prompt: *Ricky stores his toys in the closet. Where are Ricky's toys?*.
- Child's task: create a grammatically correct answer from a set of words provided (organised by part of speech).
- System's task: check that the answer is grammatically correct and plausible, and, if possible, provide feedback on mistakes.

Flickinger: EPGY, Stanford, *Language Arts and Writing for Grades 2–6*

# Language learning examples

Some correct answers:

Ricky’s toys are in the closet
Ricky’s toys are in his closet
they are in his closet
the toys are in Ricky’s closet
Ricky’s toys are in Ricky’s closet

Some incorrect answers:

Ricky’s are his toys in his closet
in in in in Ricky’s in Ricky’s Ricky’s
Ricky’s toys are in the garden

# Simplifying text

Example from Siddharthan (2011):

The original police inquiry, which led to Mulcaire being jailed in 2007, also discovered evidence that he has successfully intercepted voicemail messages belonging to Rebekah Brooks, who was editor of the Sun when Mulcaire was working exclusively for its Sunday stablemate.

The original police inquiry led to Mulcaire being jailed in 2007. The police inquiry also discovered evidence that he has successfully intercepted voicemail messages belonging to Rebekah Brooks. Rebekah Brooks was editor of the Sun. This was when Mulcaire was working exclusively for its Sunday stablemate.

Advaith Siddharthan. Text Simplification using Typed Dependencies: A Comparision of the Robustness of Different

# Simplifying text

Example from Siddharthan (2011):

The original police inquiry, which led to Mulcaire being jailed in 2007, also discovered evidence that he has successfully intercepted voicemail messages belonging to Rebekah Brooks, who was editor of the Sun when Mulcaire was working exclusively for its Sunday stablemate.

The original police inquiry led to Mulcaire being jailed in 2007. The police inquiry also discovered evidence that he has successfully intercepted voicemail messages belonging to Rebekah Brooks. Rebekah Brooks was editor of the Sun. This was when Mulcaire was working exclusively for its Sunday stablemate.

Advaith Siddharthan. Text Simplification using Typed Dependencies: A Comparision of the Robustness of Different

# Analysis of text

- Analysis of text is the core of NLP applications.
- Level and type of analysis attempted depends on application:
  - Depth of analysis? Meaning?
  - Explicit linguistic analysis vs surface-based techniques?
  - Precision? e.g., testing for grammaticality vs processing noisy data.
- NLP is moving away from pipelined models: syntactic and semantic analysis used along with surface analysis techniques.
- Also work on language generation and regeneration.

# NLP and linguistics

The computational modelling of human language (NLP/computational linguistics/CL)

1. Morphology — the structure of words.
2. Syntax — the way words are used to form phrases.
3. Semantics
   - Compositional semantics — the construction of meaning based on syntax.
   - Lexical semantics — the meaning of individual words.
4. Pragmatics — meaning in context.

# Language and Speech courses in the ACS

- L100: Introduction to NLP (L100a and b)
- L106: Spoken Language Processing
- L113: Word meaning and discourse understanding
- R211: Biomedical informations (half module)
- R207: Language and Concepts
- L101: Machine learning for language processing (L106 is prerequisite)
- L102: Statistical machine translation
- L107: Syntax and semantics of natural languages (NB: logic required)

# Syllabus for L100

- Linguistics for NLP — morphology, syntax, semantics and pragmatics [6 lectures, AAC]
- Finite-State/Markovian Techniques — lemmatisation, part-of-speech (PoS) tagging, phrase chunking and named entity recognition (NER) [4 lectures, SC]
- Parsing — grammars, treebanks, representations and evaluation, statistical parse ranking [4 lectures, SC]
- Interpretation — compositional semantics and entailment, pragmatic inference [2 lectures, AAC]
- Two practicals.

# Resources for L100a

On website:

- slides (on website after lecture).
- ejb's handouts: assigned reading before each lecture (do exercises!)
- introductory logic worksheet
- assessed exercises
- Possibly useful background: NLP Part II lecture notes.

Recommended Book: Jurafsky and Martin (2009)
See also suggestions at the end of the handouts.

# Assessment for L100

- Four ‘ticked’ assignments: first due at 10am October 15 (hand in to student admin). Each worth 5%.
- Practical reports, due next term.

# Plan for L100a

1. Introduction (this lecture).
2. Morphology and lexical categories (this wednesday).
3. Syntax 1
   - Review of assignment 1.
   - Syntax.
4. Syntax 2
   - Review of assignment 2.
   - Using context free grammars.
5. Syntax 3
   - Review of assignment 3.
   - Grammatical relations.
6. Compositional semantics
   - Review of assignment 4.
   - Semantics.
7. Semantics and pragmatics.
8. Topics arising . . .

# Examples of analysis

Chromosomal alterations were detected in humans exposed to butadiene.

- Morphological analysis and POS assignment
- Noun phrase bracketing
- Syntax tree
- Grammatical relations
- Logical form

# To do

- Before the next lecture (Wednesday): read sections 1–4 of ‘Introduction to Linguistics’ handout and do the associated exercises.
- Start logic worksheet.