Natural Language Processing and Applications
Dr. Mark Lee
mgl@cs.bham.ac.uk

http://www.cs.bham.ac.uk/~mgl/nlpa
Topics to be covered today

- Overview of Module
  - Topic Schedule
  - Assessment
- What is NLP?
- Applications

How many Languages are there anyway?
Module Structure

Introduction (1 week)

Words (2 weeks)
  - Morphology & Tagging

Syntax & Parsing (2 weeks)
  - FUG & Lexically driven parsing
  - Statistical Natural Language Learning

Semantics & Meaning (2 weeks)
  - Logic & Meaning
  - Discourse Representation Theory

Discourse (1 week)
  - Co-reference resolution

Applications & Resources (2 weeks)
  - Information Retrieval & Text Mining

Wrap Up
Assessment

- Coursework
  A 2000-4000 word essay based on a topic from the lectures (20%) or
  Some software?
  (Possible titles will be suggested in Week 5)
- Exam
  90 minute exam in May
What is Natural Language Processing?

• The Processing of Natural Language (stupid!)
• But what is language?
  - Human intuition
  - Linguistic rules
  - Data
• The Variety of Language
  - 5000 - 10,000 different current natural languages
  - Dialects, pidgins & creoles (& dead languages)
  - Novel words & figurative language
• Could you spot a language if you saw one?
The Voynich manuscript
What is Computational Linguistics?

- **Same problems as theoretical linguistics**
  - How does language function as a medium for communication?
  - How do children (or adults) learn language?
  - How and why do languages change over time?

- **But with a concern for**
  - Performance (Precision/Recall)
  - Engineering (what applications can we build)

- **And insisting**
  - Any theory should be as well specified as a computer programme
  - (and ideally have an useful application)
Computational Linguistics as Science?

- **Computing as a source of inspiration**
  - Scheduling/procedures
  - Parallelism

- **Top-Down versus Bottom-up**

  A: “It's Sunday afternoon, and I'm bored”
  B: “Today, most shops are open in many cities”

  **Pragmatics** “Let's go the shops!”

  **Semantics** today(a0), MOST(s0, s1, s2), MANY(s2, s3), city(s2, c), shop(s2s), open(s3, s), past(s3)

  **Syntax** [cat=s mod = today subj=[cat=np det=most num=plur head=[cat=n pred=shop]]....]

  **Morphology** Today, most shop+s be+past open in many city+s

  **Phonology** “Today, most shops are open in many cities.”

- **What gets done when?**
  - The Autonomy of Syntax hypothesis (Chomsky)
Computational Linguistics as Engineering

Computing as a source of power

Previously many theories in linguistics could not be tested due to

- Lack of time
- Lack of data
- Lack of statistical foundations

This has changed.

- Computers are much faster/more powerful
- We know more about language & statistics

Data exists due to electronic publishing

- Machine Readable Dictionaries (LDOCE, COBUILD)
- Corpora (BNC, The Bank of English)
Applications

Traditional Applications
- Machine Translation (Systran)
- Human-Computer Interfaces

Current Applications
- Text mining
- Information Extraction
- Mobile Phone Text Prediction
- Spell Correction & Style Checkers
- Email Spam Filters etc. etc.
How should we study Language?

- **Traditional Linguistics**
  - Linguistics should be about what might be said rather than what has been said
  - Statistics can't tell us much about language - only about words.
  - It is difficult to balance a corpus. ("Bash")
  - Chomsky & Generative Linguistics

- **Corpus Linguistics**
  - Intuitions can be wrong. ("Total" as a verb, "the hoi polloi")
  - Science requires evidence.
  - We now have the data & computing power
  - Probable versus the Possible

- **Natural Language Processing**
  - Robust processing of the probable versus
  - Imperfect processing of all possibilities
Chomsky & Language Learning

- **Chomsky**
  - Children learn from a small number of examples with essentially no counter examples.
  - Language Learning in Children is fast!
  - (Formally) NL grammar is impossible to learn with only positive evidence

- **There are two things needed to learn a language**
  - Principles (Universal to all natural languages)
  - Parameters (Which are language specific)

- **You learn a (first) language by acquiring the correct set of parameters and slotting them into your set of principles**

- **This is why second language learning is so hard.**
Chomsky (II)

- **External Language (E-Language)**
  - Samples of language (performance). Describe features of the same through structures.
  - Social Convention
  - Behavior
  - Pragmatic Performance

- **Internal Language (I-Language)**
  - Invented sentences. Describe aspects of the mind with principles.
  - Mental reality
  - Grammatical Competence

- **Competence versus Performance**
  - Linguistic Intuition versus Statistical Occurrence in a Corpus
  - Intuitions about rare data are more interesting than about common data
  - “Colourless green ideas sleep furiously”
Radical Empiricism

• We should build language models by applying standard techniques directly to the data. (throw the computer at it)
  – Stochastic Methods
  – Connectionism
  – Example-based methods
  – Genetic Algorithms

• Central Claim (& Hope)
  – Radical Empiricism was good for speech
  – It should be equally good for higher levels of language
Hidden Markov Models

• Hidden Markov Models (HMM)
  – A finite-state machine with
  – Probabilities assigned to transitions, with
  – Sets of outputs associated with states, and
  – Probabilities assigned to outputs

• n-gram models
  – “The cat sat on the mat” (versus)
  – “The dog slept on the rug”
HMMs are Trainable

- Model Type (bigram etc.) gives basic geometry
- Training assigns probabilities (Verterbi, Forward Backward algorithm etc.)

Applications
  - Speech Recognition
  - Part of Speech Tagging
  - Spell checkers
  - Thesaurus Grouping and Concept Formation
  - Text prediction

Limitations
  - HMMs are finite state models.
  - Can deal with local phenomena (Phonology/POS tagging)
  - Syntax/semantics/pragmatics phenomena are recursive (and global)
  - tri-grams tend to be at the limit in terms of computer power/amount of data
Language is situated

- How much is there to be learned from text?
  - Interpretation is determined by context & world knowledge

- Examples
  - The police refused the students a permit because they feared violence
  - The police refused the students a permit because they advocated violence
  - I just got back from Saudi Arabia. I had forgotten how good beer tastes.
  - I just got back from the Czech Republic. I had forgotten how good beer tastes.
  - (Ich hatte vergeBen wie gut(es) Bier schmeckt.)

- Counter Claim: The facts are not in the text.
Machine Translation

- Machine Translation
  - The first task NLP tried to do (1950s)
  - Was aimed for scientists understanding other scientists
  - Glossing was enough
  - Systran

“Little Johnny was looking for his toy box. Finally he found it. The box was in the pen. Johnny was very happy” (Bar-hillel, 1960)

- Interest re-emerged in the 1980s
  - Computers were faster/resources were available/our understanding was better..

- Present day
  - Verbmobil
  - (Speech to speech translation between German/English/Japanese in real time)
What is Translation?

- It's what translators do. (of course)
- The original
  
  Il ya une dizaine d’annéees, on croyait que les pays industrialises était débarasses des risques lies aux maladies infectieuses et que la sante publique n’était menacee que par des maladies comme le cancer, les troubles cardiaques et les anomalies genetiques.
  
- A word for word translation
  
  Ten years ago, it was believed that industrialized countries were free from risks connected to infectious diseases and that public health was menaced only by illnesses like cancer, cardiac problems and genetic anomalies.
  
- The actual translation
  
  As recently as a decade ago it was widely believed that infectious disease was no longer much of a threat to the developed world. The remaining challenges to public health there, it was thought, stemmed from noninfectious conditions such as cancer, heart disease and degenerative diseases.
Statistical MT

• “The most effective method IBM has found for reducing error rates is to fire a linguist”

• “Better data is more data”

  (Attributed to Bob Mercer, IBM)
IBM Project in the late 1990s
“Pure” Statistics-based Machine Translation
Basic Idea
- Use Canadian Hansard
- Learn sentence/clause fragments of English/French
- Assign frequencies for Eng. $W_1...W_n \rightarrow$ French $W_1...W_n$
- Use Bayesian probabilities to pick most likely translation
Small development cost & performance slightly worse than Systran
Issues

- How common are “Canadian Hansards”?
- Greater need for MT for resource poor languages (e.g. Somalian rather than French and English)
- What is the other stuff in the algorithm?
  - POS Tagging
  - Sentence splitting
  - Stemming
  - Dictionary look up
  - Grammar parsing
- Is this statistical MT with some NLP or NLP with some statistics?
Evaluations

- Evaluation Competitions are a current trend
  - Message Understanding Conference (DARPA)
    - Information Extraction from Text
    - Fleet operations from military reports
    - Terrorist activity in newspaper text
    - etc.
- Currently there are evaluation competitions for most applications of NLP
  1) Problem Definition
  2) Development period on original data
  3) Evaluation on more original (but similar) data
  4) Conference to share results & techniques
Sentiment Analysis

- Sentiment Analysis is the automatic classification of qualitative and emotional language.

- Value judgements can be as fine grained as required
  - Positive versus Negative
  - Anger/Surprise/Sadness/Happiness etc.

- **Sentiment Analysis versus Opinion Mining**
  - SA -> Analysis of user submitted data about a given product/service
  - OM -> Search and analysis of relevant data about a product etc.
Industrial Relevance

- Highly relevant to Industry
  - Product reviews
  - Advertising campaigns
  - Customer feedback

- Rapide & Rant 'n' Rave
  - Automatic Sentiment Analysis of Customer Feedback
  - Mobile Phone Feedback
  - Current contracts with National Express, Audi etc. etc.
  - In house Sentiment Analysis Engine (NOT the one I'll describe today)
Sentiment Analysis

- Consider an application
  - Internet Spider retrieves reviews of your product
- Automatic classification of
  - Positive comments
  - Negative comments
  - (and reasons why?)
- How would you implement this?
“We were looking forward to this hotel as we have been to Egypt several times and looked forward to see the temples etc. recreated but it turned out to be more of a tomb. Never in my life have I found it so difficult to move around a hotel. We were in a room miles from any elevator on the 6th floor, to my knowledge only 2 lifts went there. Trying to find an exit was a nightmare but I was told this is to keep you in the casino. I wouldn’t mind but they don’t have a 5 card stud table. The swimming pool didn’t open till 10 am while we were there and the other one was closed. also closed was the imax cinema, the fusion restraunt and various other places under construction.”

(Luxor Hotel, LV review from www.tripadvisor.com)
A first step at a naïve approach

- Score words according to their “affective quality”
  - Good = + 50
  - Bad = - 50
  - Etc.

- Given a document of $W_1 \ldots W_n$
  - affective score of document = $\sum w_1 \ldots w_n$

- Use word frequencies in a corpus to estimate Bayesian probabilities
Limitations

- In extended text, it's not clear that repeated mention indicates emotional strength.

- Negation and emphasis are ignored
  - “The hotel was not bad ...”
  - “The hotel was very bad”

- “Nigeria hostage feared dead is freed.”
  - Three negative terms
  - One positive term
  - Overall positive interpretation

- Options for further Information Extraction are limited
  - Why does the writer hold this judgement?
Sentiment Analysis Competition

- Affective Text Task (Strapparava and Mihalcea, 2007)
- Held as part of SemEval conference on Word Sense Disambiguation
- Development data set of 250 newspaper headlines
- Evaluation data set of 1000 newspaper headlines
- Human annotation of emotional valence
  - 6 human annotators
  - Fine grain: [-100,100]
  - Coarse grain: Negative [-100,-50], Neutral [-50,50], Positive [50,100]
  - High degree of human interannotator agreement
Some examples from the corpus

- Mortar assault leaves at least 18 dead
- Goal delight for Sheva
- Bombers kill shoppers
- We were 'arrogant and stupid' over Iraq, says US diplomat
- Bush Insists Troops Stay in Iraq, Predicts Midterm Victory
- How to Prevent Iraq From Getting Even Worse
- Tumor Types May Explain Survival Rates for Cancer
- Marathon victory comes at heavy cost
- Nigeria hostage feared dead is freed
- Etc.
Our Approach

Three Main Components

- An electronic dictionary of affective words
- Combinatory Categorical Grammar Parser
- Contextual Valence Shifters
Wordnet 3.0

• Electronic Dictionary organised as a semantic network
• Each node is a set of synonyms (synset)
• Links between synsets are hyponyms, meronyms, “related to” etc.
• 150,000 words organized in over 115,000 synsets
• Verbs/Nouns/Adverbs/Adj only
Building an affective dictionary: Adjectives and Adverbs

- Manually constructed seed set of positive and negative terms
  - \{good, beautiful, happy, pleasant, clean, etc.\}
  - \{bad, hideous, sad, unpleasant, dirty, etc.\}
- All seedset terms assigned a valence of 100 or -100
- Expand each seedset
  - Follow links in wordnet to a depth of 11
  - Any word beyond this depth is “neutral”
- 4200 Affective adverbs and adjectives
Manual seedsets of positive and negative terms constructed

- \{celebrate, pleasure, protect, cure, etc.\}
- \{catastrophe, misfortune, disrespect, pain, etc.\}

- Manual estimation of Affective Valence between [-100,100]
- Expand each seed set
  - Follow links in wordnet to a depth of 11
  - Any word beyond this depth is “neutral”
- Seedsets were 100 terms expanded to 5700 synsets and 3900 words
Combinatory Categorial Grammar

- Based on combinatory logic
- All constituents as assigned categories based on their syntactic behaviour
- Grammar is lexicalized
  - Most grammatical knowledge is in the dictionary
  - Small number of application rules (in our case 2)
- Clear mapping from syntax to semantics
From Syntax to Semantics

\[
\begin{align*}
\text{John} & \quad \text{has} & \text{very} & \text{little} & \text{money} \\
N & \quad (S\backslash NP)/NP & (N/N)/(N/N) & (N/N) & N \\
NP & \quad (N/N) & N/N & N & N \\
& \quad NP & \quad S\backslash NP & S
\end{align*}
\]
Contextual Valence Shifters

• Certain words can modify (or flip) the valence expressed by other terms
  - “He is not bright”
  - “That is extremely good”
  - “The food was cheap however it tasted awful.”

• Negatives, Intensifiers and Diminishers
  - Re-calculate affected phrase by valence of shifter
  - “not” = -1 (so it flips valence)
  - “extremely” = 2 (so it intensifies valence)

• Connectors
  - “while”, “but”, “however” etc.
  - In such cases, only calculate valence for main clause of sentence.
Classification

- Pre-Processing (expand contractions etc.)
- Partially Parse Text
- Dictionary lookup
  - Assign values for “sentiment”
- Full parse
- Resulting semantics is then used to calculate the overall sentiment score.
## Results

<table>
<thead>
<tr>
<th>Dictionary in Use</th>
<th>Accur.</th>
<th>Precis.</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>63.20</td>
<td>53.21</td>
<td>50.48</td>
<td>51.81</td>
</tr>
<tr>
<td>Adj &amp; Adv only</td>
<td>58.80</td>
<td>43.48</td>
<td>2.44</td>
<td>4.62</td>
</tr>
<tr>
<td>Nouns &amp; Verbs only</td>
<td>62.30</td>
<td>52.00</td>
<td>50.73</td>
<td>51.36</td>
</tr>
</tbody>
</table>

### SemEval2007 Systems*

<table>
<thead>
<tr>
<th>System</th>
<th>Accur.</th>
<th>Precis.</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClaC</td>
<td>55.10</td>
<td>61.42</td>
<td>9.20</td>
<td>16.00</td>
</tr>
<tr>
<td>ClaC-NB</td>
<td>31.20</td>
<td>31.18</td>
<td>66.38</td>
<td>42.43</td>
</tr>
</tbody>
</table>

(*Andreevskaiia & Bergler, 2007)
Summary

- Module Structure & Assessment
- What is NLP?
- Linguistics & the study of Language
- Applications

- Claim: Linguistic knowledge is useful when processing Text!