Language engineering systems

- Statistical modelling
  - N-gram language models
  - Data generation
- Applications
  - Question-answering,
  - Dialogue systems
  - Information extraction
  - Machine translation
  - Reading and writing aids

Statistical modelling

- Uses
  - Disambiguation / selection
    - Structural ambiguities (e.g. PCFG)
    - Word sense disambiguation
    - Word prediction
  - Data generation
    - Translation data (e.g. bilingual dictionaries)
    - Collocations
- Acquisition
  - Probability (parameter) estimation on training data

Simple models

- A priori models
  - Choose the event with the highest probability
- With knowledge of related event O
  - Choose $E^* = \arg\max p(E|O)$
  - Examples:
    - Bigram language models
      $O$ is previous word and $w^* = \arg\max p(w | w_{i-1})$
    - Trigram model probabilities
      $O$ is two previous words and $w^* = \arg\max p(w | w_{i-2} w_{i-1})$

Specifying a language model

- Collect / Obtain a corpus
  - Separate parts for training and testing
- Create n-gram models, where n depends on the size of the corpus
  - smoothing, back-off
  - test on training corpus first, trying various techniques
- Test on test corpus

N-gram models

- Data sparseness
  - All n-grams of interest cannot be found in the training corpus.
- Solutions
  - “Smoothing” of probability mass onto all n-grams;
  - Rely on shorter n-grams when data is unavailable (“backoff”)

The noisy channel model

- Bayes’ rule
  $p(E|O) = p(E)p(O|E) / p(O)$
  $E^* = \arg\max p(E|O) = \arg\max p(E)p(O|E)$
  $p(E)$ is the a priori model
  (usually an n-gram model if E is a string)
  $p(O|E)$ is the channel model
Bigram model as automaton

- An n-gram model is a statistical automaton (Markov model) where states are associated with symbol sequences of length n-1.
- Bigram models (character sequences):

```
V 0.27 0.73 0.51 0.53 0.20
C 0.38 0.11
```

- \( P(V|O) = 0.27, \ P(V|C) = 0.53, \ P(V|V) = 0.11 \) etc.

Hidden Markov Models

- A HMM is a weighted automaton with a probability distribution over symbols in every state (rather than just a single symbol).
- The Viterbi algorithm
  - Given an observed symbol sequence, \( O \), determines the most likely state sequence to produce \( O \).

Entropy and perplexity

- Both measures "uncertainty in prediction"
  - Entropy: \( H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \)
  - Perplexity: \( 2^H \)
- Used for measuring how good a language model is

Co-occurrence statistics

- General idea
  - Two items that have a linguistic relation often also have a similar distribution in texts.
  - Examples
    - If \( w_1 \) and \( w_2 \) are collocates then \( w_2 \) is likely to appear in the vicinity of \( w_1 \) (e.g. komma och ihåg)
    - If \( w_1 \) is a translation of \( w_2 \), then \( w_1 \) is likely to appear in the same relative position in translations, as \( w_1 \) is in the originals (e.g. remember och ihåg).
  - So, by (cleverly) looking for items with similar distributions we are likely to find items with interesting relations

Co-occurrence statistics

- Dice koefficient
  \[
  \frac{2|X \cap Y|}{|X| + |Y|}
  \]

- t-variable
  \[
  \sqrt{\frac{|X \cap Y| - |X||Y|}{|X||Y|}}
  \]

Question Answering (Q&A)

- Collect an answer (a phrase, sentence or paragraph) from a corpus (open domain) to a (factual) question in natural language
- Main issues
  - Question interpretation (yes/no, "why", indirect, commands)
  - Answer extraction (length, multiple sources)
  - Presentation
- Approaches
  - Shallow, words + poss. syntactic analysis
  - Deep, semantics + poss. Logic
Question Answering (Q&A)

- Techniques, algorithms etc
  - Question type hierarchies/Taxonomies (Wordnet)
  - Semantic interpretation
  - Document retrieval (Indexing, semantic interpretations)
  - Named entity recognition (Regular expressions/finite state rule-based, ML/HMM)

- Articles
  - Harabagiu et al. FALCON – architecture, knowledge sources, processing, evaluation
  - Mann. Algorithm – Statistical approach

Dialogue systems

- Natural language dialogue is used to achieve a task (problem, information, teaching) through joint effort of user and system
- Main issues
  - Utterance interpretations (utterance segmentation, questions, answers)
  - User modelling
  - Dialogue management (Anaphora, ellipsis, turn-taking, clarifications, common ground)
  - Domain knowledge management
  - Generation in dialogue
Dialogue systems: phase architecture (NLPLab)

- Interpreter
- Dialogue Tree
- Domain Knowledge Manager
- Structured Information Source
- Lexicons & Grammars
- Dialogue Grammar
- Domain Knowledge Source
- Domain Ontology

Dialogue systems: hub architecture

DARPA Communicator

- Hub
- Language Generation
- Dialogue Management
- Text-to-Speech Conversion
- Speech Recognition
- Frame Constructor
- Audio Server
- Application Backend
- Context Tracking

Dialogue systems

- Techniques, algorithms etc
  - Interpretation (partial/full, statistical/rule-based, syntactic/semantic)
  - Dialogue management (frame, grammar, plan)
  - Domain knowledge management (ontology)
  - Generation (templates)

Dialogue systems

- Evaluation
  - User satisfaction
    - Time
    - Understandability
    - Learnability
  - Task completion cost
    - Time, turns or seconds
    - Number of queries
    - Number of turns for error correction
    - Inappropriateness of questions, answers, clarifications

Dialogue systems

- Trends
  - Portability, frameworks
  - Multi agents
  - Multimodality

- Articles
  - Johansson et al. Development
  - Walker et al. PRADISE – evaluation

Information Extraction (IE)

- Identification of relevant objects and events to be extracted from texts and represented in a structured format
- Main issues
  - Text categorisation
  - Identification of relevant information (NER, …)
  - Merging of information (coreference, …)
- Approaches
  - Hand-coded rules and patterns, syntactic or semantic
  - Statistical, training on tagged data
Information Extraction (IE) architecture

- Tokenisation
- Word segmentation
- Word sense tagging
- POS tagging
- Morphological and syntactic analysis
- Syntactic analysis
- Domain analysis
- Coreference
- Merging partial results

Information Extraction (IE)

- Techniques, algorithms etc
  - POS tagging
  - Partial (domain-oriented) parsing
  - Named entity recognition
  - Co-reference resolution (taxonomies/ontology)
- Evaluation
  - Precision and Recall, F-measure
  - Components
    - Named entities F=95% (<100%)
    - Co-reference F=50-60% (80%)
    - Template elements F=80% (93%)

Information Extraction (IE)

- Trends
  - Adaptability/Portability
  - Interactive/Supervised learning
- Articles
  - Gaizauskas et al. Multi-lingual IE – architecture, knowledge sources
  - Hobbs et al. FASTUS – architecture, development, evaluation

Summarisation

- Interpret and understand text to generate a summary of the information
- Main issues
  - Text interpretation and representation
  - Extraction and generation of summary
- Approaches
  - Shallow – cut and paste
  - Deep – understand, extract, generate

Summarisation architecture (SUMMONS)

- IE templates
- Content planner
- Combiner
- Paragraph planner
- Linguistic generator
  - Ontologizer
  - Lexical choser
  - Sentence generator
  - Operators
    - Addition
    - Trend
    - Agreement

Evaluation

- Comparison of text and summary
- Questions
- Comparison with summaries made by humans
- Function/Usability

Techniques, algorithms etc

- IE + NL generator
- POS tagging and Partial parsing/Phrase recognition
- Named entity recognition
- Co-reference
- Statistic methods for comparison and merging
Summarisation

- Trends
  - Domain/genre specific
  - Multi document summarisation
- Articles
  - Shiffman et al. Biographical summarisation
    - architecture, processing, evaluation
  - Mani. Evaluation

Machine Translation (MT)

- Automatic or semiautomatic translation of texts
- Main issues
  - Interpretation, transfer
  - Data acquisition
- Approaches
  - Rule-based
    - Direct, transfer, Interlingua
  - Data-driven
    - Example-based, Statistical

Machine Translation (MT)

- Techniques, algorithms etc
  - Full parsing transfer models
  - Statistical machine translation (SMT)
  - Example-based machine translation (EBMT)
- Evaluation
  - Comparison with other target texts
  - Human judges
    - Fluency
    - Fidelity
    - Usability

Spelling and Grammar

- Detect faults in spelling and grammar, and suggest corrections
- Challenges
  - Non-word detection
  - Isolated word error detection
  - Context dependent word error detection
  - Grammatical error detection
  - Detecting combinations of spelling and grammar errors
  - Generating and presenting correction alternatives

Machine Translation (MT)

- Trends
  - Assisted MT
- Articles
  - Senellart et al. Systran – architecture, knowledge sources
  - R. Brown: Example-Based Machine Translation in the Pangloss System
  - Papineni et al. BLEU - evaluation

Spelling and Grammar

- Approaches
  - Rule-based
  - Probabilistic models
- Techniques, algorithms etc
  - Minimum edit distance (insertion, deletion, substitution, transposition)
  - Noisy channel
Spelling and Grammar

- Evaluation
  - Precision and recall
  - Usability
- Articles
  - Carlberger et al. GRANSKA – architecture, evaluation