Concept Chunking

Sander Canisius
Text Mining
February 22, 2005

Overview

- Introduction
- Techniques
- Applications

Introduction

Example

Apologies, as always, for any cross-postings...

CALL FOR PAPERS

THE CHALLENGE OF IMAGE RETRIEVAL

A Workshop on Content-Based Image & Video Retrieval

February 5, 1998, University of Northumbria at Newcastle, UK

IMPORTANT DATES:

Deadline for Submission: 24 November 1997

Notification of Acceptance: 8 December 1997

Camera-ready Papers due: 23 January 1998

This one-day Research Workshop forms the first day of a two-day conference on image retrieval to be held in Newcastle upon Tyne on 5-6 February 1998. It aims to provide a forum for presenting new research

...

Slot types

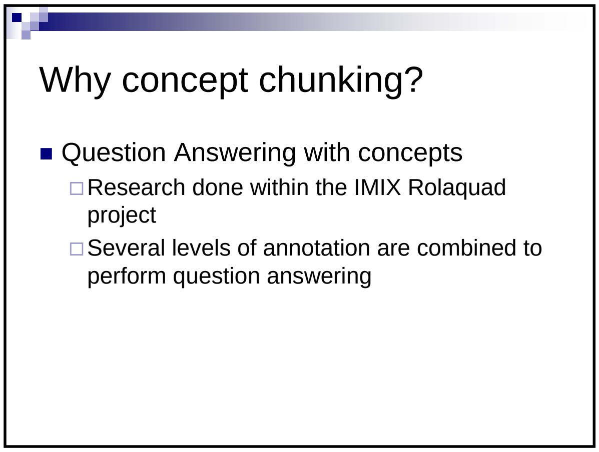
conferenceacronym • conferencehomepage • conferencename • workshopacronym • workshopcamerareadycopydate • workshopdate • workshophomepage • workshoplocation • workshopname • workshopnotificationofacceptancedate • workshoppapersubmissiondate

What is a concept?

- Textual information unit that is an instance of one of several domain-specific types
  - □ For example, disease names in a medical domain, submission dates in call for papers, etc.
- Some standard concept chunking / information extraction domains
  - □ Seminar announcements
  - □ Molecular biology
  - □ News reports on terrorist attacks

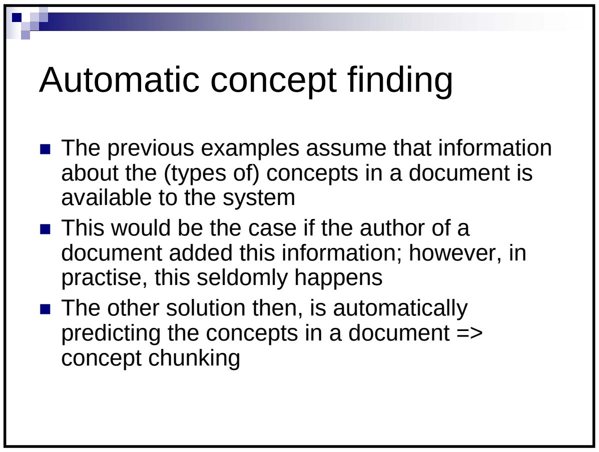
Why concept chunking?

- An example for information retrieval
  - □ Standard information retrieval engines treat tokens in a document as atomic units that are equal to terms in a search query if and only if their byte strings match exactly
  - □ In case of a google query “Java”, all top-ranked matches are about the programming language, which is of no use to someone searching information about the island
  - □ One way of overcoming this issue is if you could somehow tell google that you are looking for the island Java rather than for the programming language
    - For example, “island(Java)”



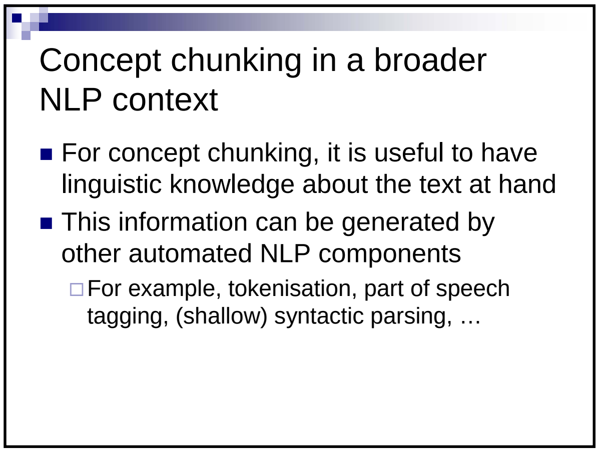
Why concept chunking?

- Question Answering with concepts
  - □ Research done within the IMIX Rolaquad project
  - □ Several levels of annotation are combined to perform question answering



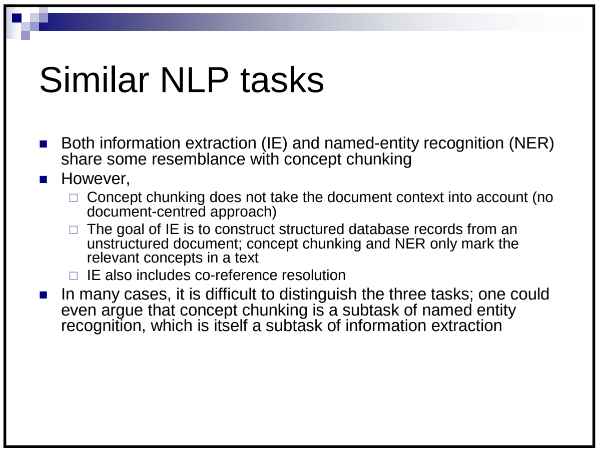
Automatic concept finding

- The previous examples assume that information about the (types of) concepts in a document is available to the system
- This would be the case if the author of a document added this information; however, in practise, this seldomly happens
- The other solution then, is automatically predicting the concepts in a document => concept chunking



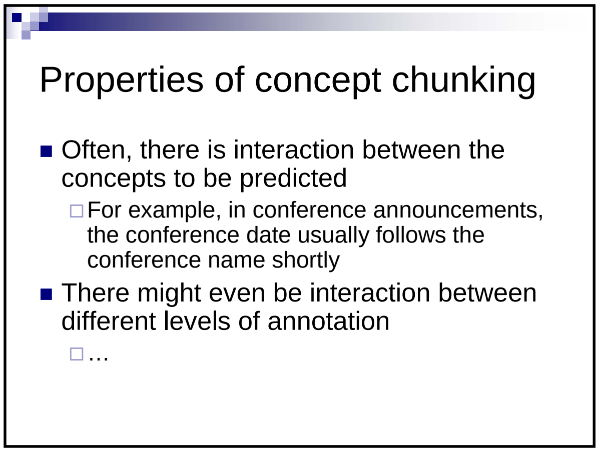
Concept chunking in a broader NLP context

- For concept chunking, it is useful to have linguistic knowledge about the text at hand
- This information can be generated by other automated NLP components
  - □ For example, tokenisation, part of speech tagging, (shallow) syntactic parsing, …



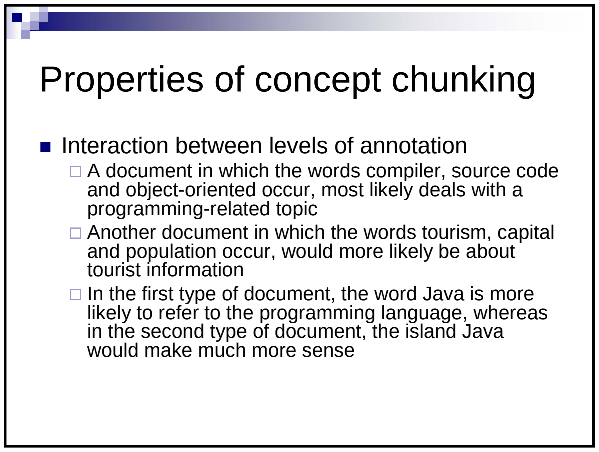
Similar NLP tasks

- Both information extraction (IE) and named-entity recognition (NER) share some resemblance with concept chunking
- However,
  - □ Concept chunking does not take the document context into account (no document-centred approach)
  - □ The goal of IE is to construct structured database records from an unstructured document; concept chunking and NER only mark the relevant concepts in a text
  - □ IE also includes co-reference resolution
- In many cases, it is difficult to distinguish the three tasks; one could even argue that concept chunking is a subtask of named entity recognition, which is itself a subtask of information extraction



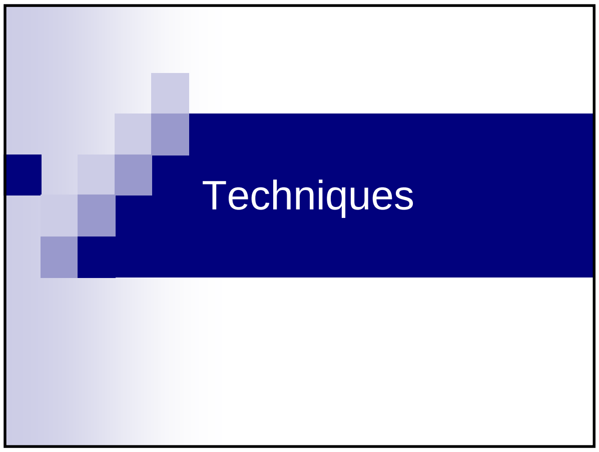
Properties of concept chunking

- Often, there is interaction between the concepts to be predicted
  - □ For example, in conference announcements, the conference date usually follows the conference name shortly
- There might even be interaction between different levels of annotation
  - □ …

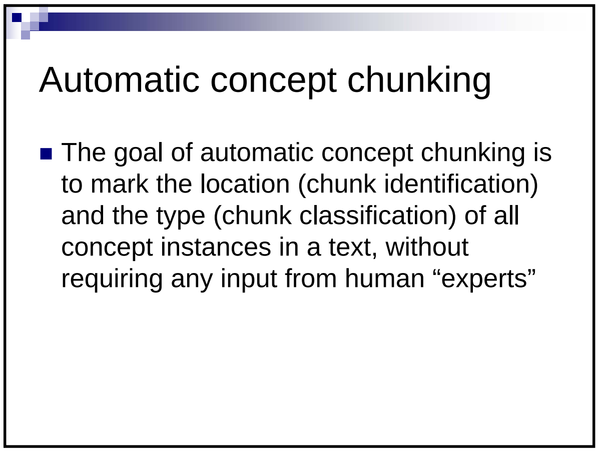


Properties of concept chunking

- Interaction between levels of annotation
  - □ A document in which the words compiler, source code and object-oriented occur, most likely deals with a programming-related topic
  - □ Another document in which the words tourism, capital and population occur, would more likely be about tourist information
  - □ In the first type of document, the word Java is more likely to refer to the programming language, whereas in the second type of document, the island Java would make much more sense

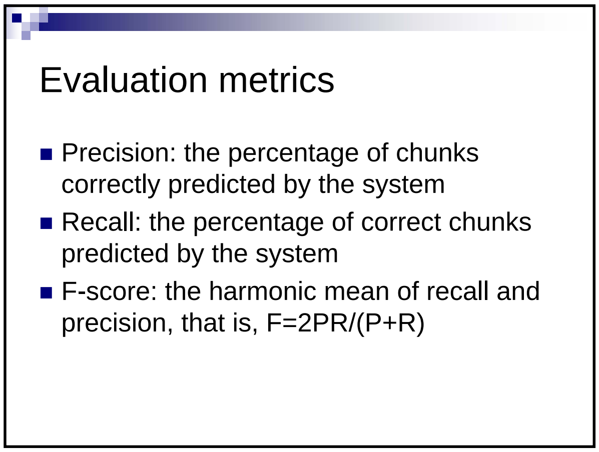


Techniques



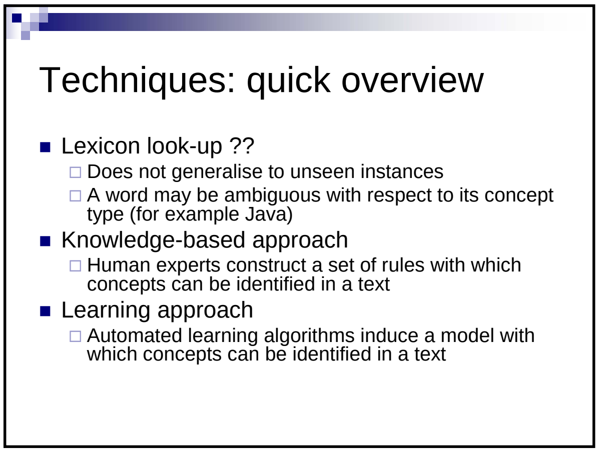
Automatic concept chunking

- The goal of automatic concept chunking is to mark the location (chunk identification) and the type (chunk classification) of all concept instances in a text, without requiring any input from human "experts"



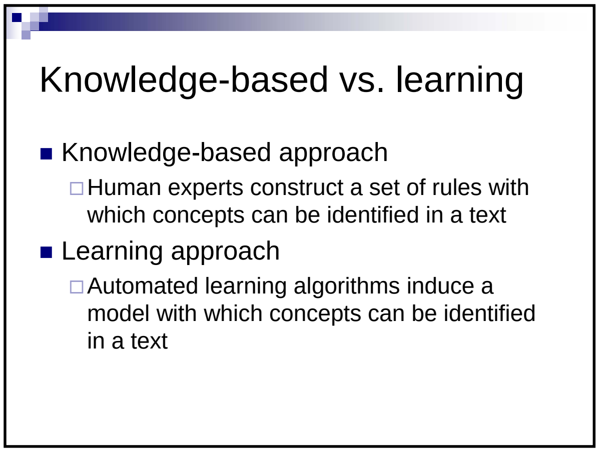
Evaluation metrics

- Precision: the percentage of chunks correctly predicted by the system
- Recall: the percentage of correct chunks predicted by the system
- F-score: the harmonic mean of recall and precision, that is, F=2PR/(P+R)



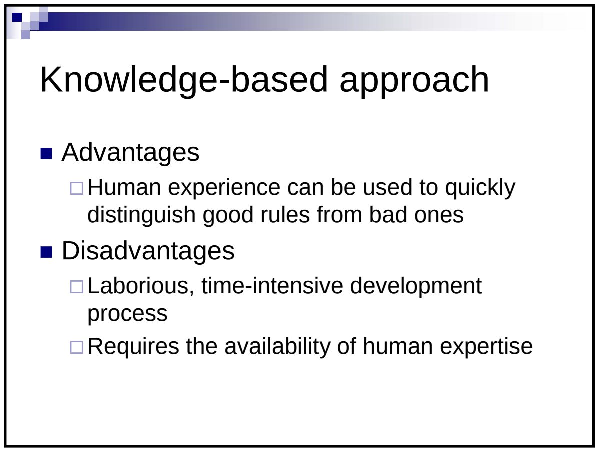
Techniques: quick overview

- Lexicon look-up ??
  - □ Does not generalise to unseen instances
  - □ A word may be ambiguous with respect to its concept type (for example Java)
- Knowledge-based approach
  - □ Human experts construct a set of rules with which concepts can be identified in a text
- Learning approach
  - □ Automated learning algorithms induce a model with which concepts can be identified in a text



Knowledge-based vs. learning

- Knowledge-based approach
  - □ Human experts construct a set of rules with which concepts can be identified in a text
- Learning approach
  - □ Automated learning algorithms induce a model with which concepts can be identified in a text



Knowledge-based approach

- Advantages
  - □ Human experience can be used to quickly distinguish good rules from bad ones
- Disadvantages
  - □ Laborious, time-intensive development process
  - □ Requires the availability of human expertise

Learning approach

- Advantages
  - □ There is no need for human experts
  - □ Techniques are largely domain independent
  - □ Exceptions are not likely to be overlooked
- Disadvantages
  - □ (Large amounts of) example data are required to train most common machine learning algorithms
    - Knowledge acquisition bottleneck vs. data acquisition bottleneck
  - □ Resulting model might not be easily understandable by human observer

Creating a rule for the concept <programming-language>

- Observation: a programming language concept is always a proper noun
- Context predicates:
  - □ “written” “in” <NOUN>
  - □ <NOUN> “compiler”
- But what to do with:
  - □ Java is a beautiful programming language
  - □ Java is a beautiful island
    - Long distance dependencies can be problematic
- Rules can be created by a human expert or automatically by a rule-induction algorithm
- Other machine learning algorithms may use more abstract models than rules

Identification & classification: parallel vs. sequential

- Parallel:
  - □ One classifier performs chunk identification and classification at the same time
- Sequential:
  - □ One classifier performs chunk identification; another performs classification
    - Might be useful if there are several similar but different concept types (workshop data, submission data, camera-ready date)
    - In that case identification can focus on correctly identifying a higher-level concept (date), and classification can focus on disambiguating already identified phrases

Chunk identification

- Most common methods
  - □ Chunking-as-tagging (IOB tagging)
    - Each token is assigned a tag denoting whether a word is outside a chunk (O), inside a chunk (I), or inside a chunk that is different from the previous one (B)
    - Possible problem: discontinuous chunks
      - □ For example, Java/I programming/O language/I
  - □ Open/Close bracketing
    - Tokens that start or end a chunk are assigned a “[“ or “]” symbol respectively
    - Possible problem: unmatched brackets
      - □ For example, [Java programming] language]

Chunk classification

- Parallel identification and classification
  - □ Append concept type to identification tag
    - For example, I-programming_language, or [programming_language
- Sequential classification
  - □ Compress the chunk found in the identification step into a single unit
    - For example, concatenating all words (may lead to sparse data!!!)
    - Orthographic properties
  - □ Encode chunk as bag-of-words

Instances for concept chunking

- Granularity
  - □ Tokens, characters, syntactic chunks
- Sliding window
- Any feature that may be useful
  - □ POS tag, syntactic chunk, orthographic information, seed lists

Machine learning techniques

- General-purpose classifiers
  - □ K-NN, Maximum Entropy, SVM, …
- Sequence learners
  - □ Hidden Markov Models, Conditional Random Fields

Class skewedness

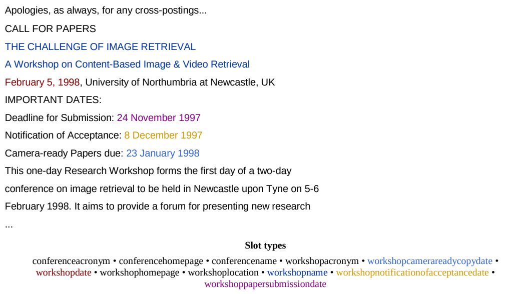
- In concept chunking, there is often an unbalanced class distribution, where the majority class is the negative class
- Standard machine learning techniques try to optimise towards accuracy
- As a result they may converge to always predicting the majority class
  - □ Leads to high precision, but low recall
- Sampling may be used for dealing with class skewedness
  - □ Up-sampling: add copies of positive instances
  - □ Down-sampling: remove negative instances

Applications

Two concept chunking domains

- Call for papers domain
- IMIX Rolaquad: medical concept finding

Call for papers domain



Apologies, as always, for any cross-postings...

CALL FOR PAPERS

THE CHALLENGE OF IMAGE RETRIEVAL

A Workshop on Content-Based Image & Video Retrieval

February 5, 1998, University of Northumbria at Newcastle, UK

IMPORTANT DATES:

Deadline for Submission: 24 November 1997

Notification of Acceptance: 8 December 1997

Camera-ready Papers due: 23 January 1998

This one-day Research Workshop forms the first day of a two-day conference on image retrieval to be held in Newcastle upon Tyne on 5-6 February 1998. It aims to provide a forum for presenting new research

...

Slot types

conferenceacronym • conferencehomepage • conferencename • workshopacronym • workshopcamerareadycopydate • workshopdate • workshophomepage • workshoplocation • workshopname • workshopnotificationofacceptancedate • workshoppapersubmissiondate

CfP domain: approach

- Double classification for dealing with class skewedness
  - □ First select relevant sentences, then do concept chunking on the selected sentences
- Features:
  - □ POS tags, orthographic information, Named entities

CfP: results

- Precision: 66.5
- Recall: 40.9
- F-score: 50.6
- State-of-the-art performance
  - □ F-score: 73.5
  - □ However, uses document-centred approach

Medical concepts

POKKEN

of variola major, een besmettelijke, door het variola virus verwekte ziekte. De ziekte is door het intensieve wereldwijde `eradicatieprogramma' van de Wereldgezondheidsorganisatie (WHO), officieel sinds 8 mei 1980 volledig uitgeroeid. Het pokkenvirus wordt nu nog slechts in een aantal laboratoria bewaard.

Slot types

disease • disease_feature • disease_symptom • method_of_diagnosis • person • person_feature • body_part • bodily_function • treatment • advice • micro-organism • duration

Medical concepts: approach

- Relatively new project
- Only simple tagging applied so far

Medical concepts: results

- Precision: 69.09
- Recall: 66.03
- F-score: 67.53