Information Extraction

- Definition and explanation
- Performance measures
- Example algorithm
- State of the art

What is IE?

- Field of natural language processing
- Analyse unrestricted text to extract information about pre−specified types of events, entities or relationships.
- Not information retrieval

information extraction gets facts out of documents and you analyse the facts

information retrieval gets sets of relevant documents and you analyse the documents

What is IE? cont.

<DOC?><TXT> New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business−side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent. </TXT><DOC>

ORGANIZATION : "New York Times„

POST : "president"

WHO_IS_IN : "Russell T. Lewis"

WHO_IS_OUT : "Lance R. Primis"

Why is IE difficult?

- There are many ways of expressing the same fact.
  - BNC Holdings Inc named Ms G Torretta as its new chairman.
  - Nicholas Andrews was succeeded by Gina Torretta as chairman of BNC Holdings Inc.
  - Ms. Gina Torretta took the helm at BNC Holdings Inc.

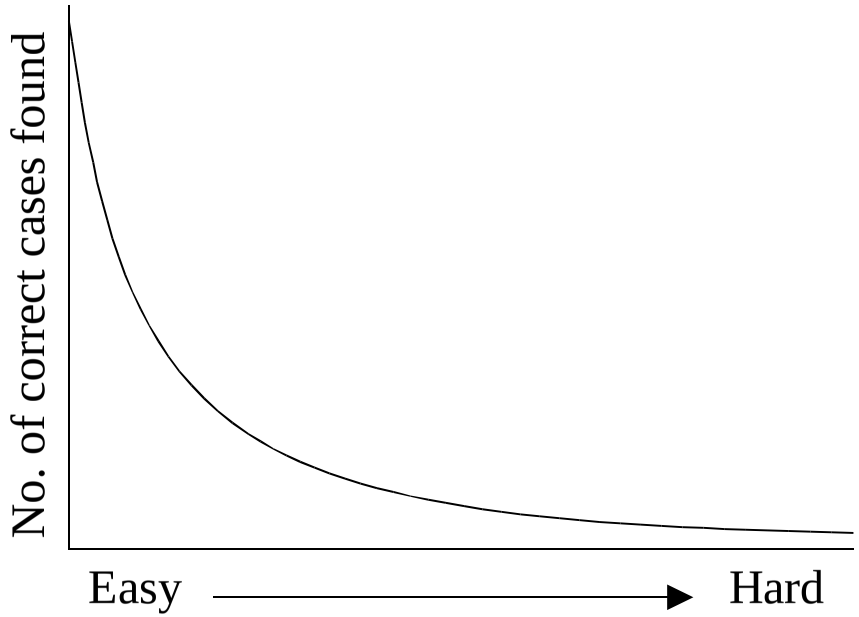
- Combine information from several sentences:
  - After a long boardroom struggle, Mr Andrews stepped down as chairman of BNC Holdings Inc. He was succeeded by Ms Torretta.

Examples of IE applications

- Competitive intelligence
- Health insurance
- Intelligence gathering
- There are a lot of potential applications but not many actual ones. Why?

Why is there not widespread use of IE technologies?

- Custom systems
- Machine learning IE fairly new field
- Generic systems don´t perform well
  - It is easy to extract many patterns



Measuring performance

- Recall
  - items correctly extracted / target items
- Precision
  - items correctly extracted / total items extraced
- Trade off
- F−Measure
  - harmonic mean of recall and precision
  - F = 2 * (recall * precision) / (recall + precision)

How to IE ...

- Match input text against patterns
- Machine learning used to find these patterns
  - Works for simple relation / event extraction
    - Easy pattern:

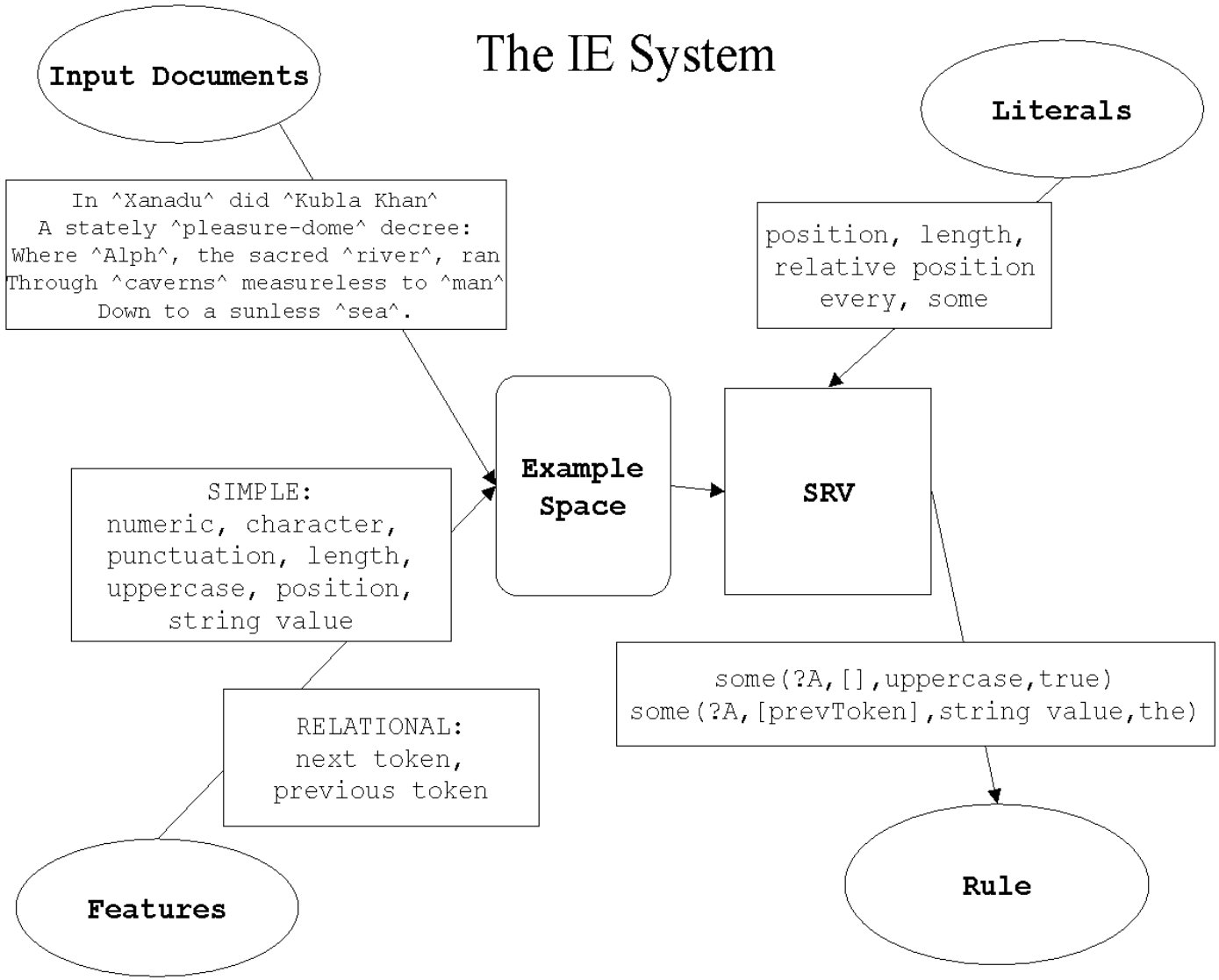
      `Subject: coffee`
    - Hard Pattern:

      `Want to meet me at the gallery?`
  - Hand coded rules still work better for more complex tasks
  - Patterns have a skewed distribution

SRV – an example IE algorithm

- Relational inference model
  - Limit context search space
  - Top down symbolic rule induction
- Generic domain
- Abstraction
  - Document characterisation
  - Symbolic rule representation
- Linguistic processing optional

The IE System



Inside SRV

- Process
  - Look for rules which match positive examples
  - Initial rule null – specialise by adding literals
  - Greedy algorithm – use gain function to choose best literal

    $\text{Information}(r) = -\log_2[\text{positive}(r) / (\text{positive}(r) + \text{negative}(r))]$

    $\text{Gain} = \text{positive}(\text{new\_r})[\text{information}(r) - \text{information}(\text{new\_r})]$
  - Limited context search space
- Covering algorithm
- Recall vs precision

SRV results

- Good
  - Can use context when there is a strong identifying factor
    - Noun class ´the´
    - Subject line of emails ´Subject: ´
  - Strong rules when features and literals capture the target information
  - Relational – only looks at context with high entropy
- Bad
  - No semantic information – but easily extensible
  - What is the best document characterisation
    - What information in the document is used? (features)
    - What form do the rules based on the information take? (literals)

Hidden Markov Models

- Markov chain
  - Finite state machine with probabilities on the transitions
- HMM order k
  - The next state is dependant on the last k states
  - The current state is hidden – only see associated output
- State of the art
  - HMM with various extensions
  - F1 = 27.2 => 87.5 average F1 = 57.2
  - Outperforms SRV

From here on ....

- SRV is now being used for wrapper induction
  - This is learning simple extraction procedures for highly structured texts such as web pages – consider the HTML / XML tags
  - With a few improvements it returns results comparable with other state of the art algorithms
- Active learning might help to make systems more accessible to the user
  - People don´t know how to mark up examples which will be good examples for extraction unless they know the algorithm being used
- Systems which have the most success may combine a number of methods
  - For example active learning is one of the improvements that could be added to a system based on a HMM