Learning Information Extraction Rules: An Inductive Logic Programming Approach

Stuart Aitken
Artificial Intelligence Applications Institute
Informatics
University of Edinburgh

Outline

- Framing the Information Extraction problem
- Ontology
  - The domain: Global Warming
  - Mark-up
- Learning rules
  - Text representation
  - Learning biases
- Experiment and Results
- Conclusions

The IE Problem

- To extract knowledge from unstructured or semi-structured sources
- To learn IE rules from a small set of marked-up texts
  - Encode content in an ontology
  - Using an existing ILP system: FOIL
  - Investigate:
    - Alternate text representations
    - The use of background theories

The IE Problem

- To extract knowledge from unstructured or semi-structured sources; knowledge is:
  - Named entities
  - Text fragments
  - Slots filled with named entities/text
  - Text categorisation
  - Ground relation instances **R(a,b)**

The IE Problem

To extract ground relation instances **R(a,b)** from sentences.

- A sentence may represent several relations.
- The constants **a b** will be drawn from an ontology
  - May be a finite set, e.g. attribute values
  - May be a constrained set, e.g. positive real numbers less than **N**
  - May be named entities

Domain Ontology

- The domain is ‘Global Warming’
- Texts were taken from Nature and New Scientist
- An ontology was created for:
  - Central concepts: GreenhouseGas, CarbonDioxide, Ozone,…
  - Qualitative and quantitative **attributes**:
    - level of concentration, emission
    - changes in concentration

Ontology and Mark-up

The burning of fossil fuels, cement production and changes in land use have led to an increase of the atmospheric concentration of CO2 by almost 30% over the pre-industrial level of 280 ppmv (ref 14)

<target name="changeInAtmosphericConcentration(CarbonDioxide,increase)"></target>
<target name="atmosphericConcentration(CarbonDioxide,280)"></target>.

Carbon dioxide, and other gases such as CH4 and N2O, are the most important greenhouse gases after water vapour and decrease the outgoing longwave radiative flux at the top of the tropopause if their concentrations increase

<target name="changeInAtmosphericConcentration(CarbonDioxide,increase)"></target>
<target name="changeInAtmosphericConcentration(Methane,increase)"></target>
<target name="changeInAtmosphericConcentration(NitrousOxide,increase)"></target>
<target name="changeInAtmosphericConcentration(WaterVapour,increase)"></target>
<target name="changeInAtmosphericConcentration(GreenhouseGas,increase)"></target>.

Ontology and Mark-up

The burning of fossil fuels, cement production and changes in land use have led to an increase of the atmospheric concentration of CO2 by almost 30% over the pre-industrial level of 280 ppmv (ref 14)

<target

changeInAtmosphericConcentration(CarbonDioxide,increase) >

</target>

<target

atmosphericConcentration(CarbonDioxide,280) >

</target>.

Ontology and Mark-up

From:

But in the June issue of Paleoceanography, Pagani, Arthur and Freeman 1 present evidence for surprisingly low CO2 levels of about 180 - 290 parts per million by volume (ppmv) throughout the early to late Miocene (9 - 25 Myr).

To (ideally):

But in the June issue of Paleoceanography, Pagani, Arthur and Freeman 1 present evidence for surprisingly low CO2 levels of about 180 - 290 parts per million by volume (ppmv) throughout the early to late Miocene (9 - 25 Myr)

```
<RDF-MARK-UP  atmosphericConcentration_Qual(CarbonDioxide,low) />
<RDF-MARK-UP  atmosphericConcentration(CarbonDioxide,180) />
<RDF-MARK-UP  atmosphericConcentration(CarbonDioxide,290) />.
```

Ontology and Mark-up

Evaluation of consistency of mark-up

- A set of 6 texts was randomly selected for double-marking
  - 46 and 54 relations were found respectively
  - In 44 cases the same mark-up was created
- Some differences in assumptions of scope were apparent
- Some differences in assessment of the degree of interpretation required

Learning IE Rules

- Learning from examples can replace human-authored rules
  - Requires the creation of a data set
  - Cannot ignore the potential need to review rules
- Inductive logic programming (ILP) is appropriate
  - Natural representation of relations
  - Can use background theories

Learning IE Rules

- ILP has similar existing applications:
  - IE, text classification, sentence interpretation (queries to a geographical database)
- Features used:
  - Token-based **has_*word***
  - Structural **next_token**
  - Features may be used as arguments **every(feature,value)**

Sentence Representation

- A sentence-based relational representation is used:

  hasWord(SentenceID,POS,Word)

- Sentences are:
  - Tagged (Brill 92)
  - The tag set is reduced
  - Morphologically analysed (Minnen 00)
  - Frequency filtered

Sentence Representation

- Structural information:

  context(SentenceID,Word1,Word2)
- Named entities

  hasWord(SentenceID,ne,Concept)

Background Theories

- Ontology

  isa(Class,Class)
- Textual indicator

  txtform1(Class,Word)

  txtform2(Class,Word,Word)
- Bias relations – generated from the target relations to create a sentence where concepts co-occur

  hasWord(NewID,ne,CarbonDioxide)

  hasWord(NewID,j,increase)

Learning from Language Data

Language data creates problems for ILP:

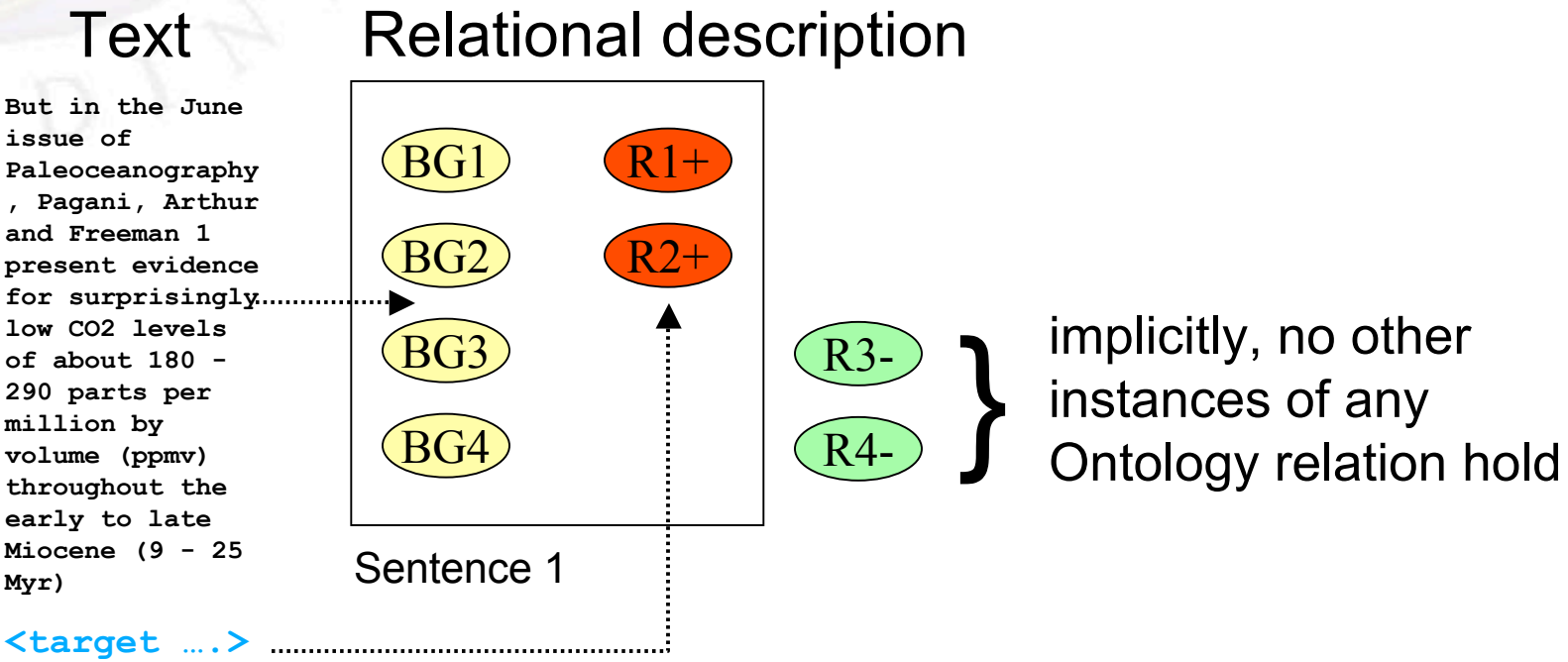
- Small sets are unrepresentative of texts as a whole, but
- large, accurate, data sets are expensive to create
- Scale
- Noise
- Consists of positive examples only

Learning from Language Data

- ILP algorithms work in two ways:
  - Bottom-up: GOLEM forms the relative least general generalisation of two positive examples, then tests the resulting rule on positive and negative examples
  - Top-down: FOIL specialises clauses, beginning with the empty clause, testing potential new clauses using an information gain metric
  - Negative examples are easiest to create from a type-restricted set: FOIL

Learning from Language Data

- The problem: only positive examples arise naturally



Experiment

- A baseline experiment, repeated with the addition of one of 5 theories was performed. This was followed by experiments with combinations of theories.
- 30 texts; 205 sentences; 5862 words
- 2/3 training – 1/3 testing; repeated 10 times

Baseline Experiment

- Example: the target relation **changeInAtmosphericConcentration** (cAC)
  - Training: 137 sentences 16 occurrences
  - Testing: 68 sentences 7 occurrences
- Experiments:
  - Baseline 1288.4; txtform 10.0; ne 93.6; isa 9.0; bias 9.0; context 17.7

Baseline Experiment

Scoring

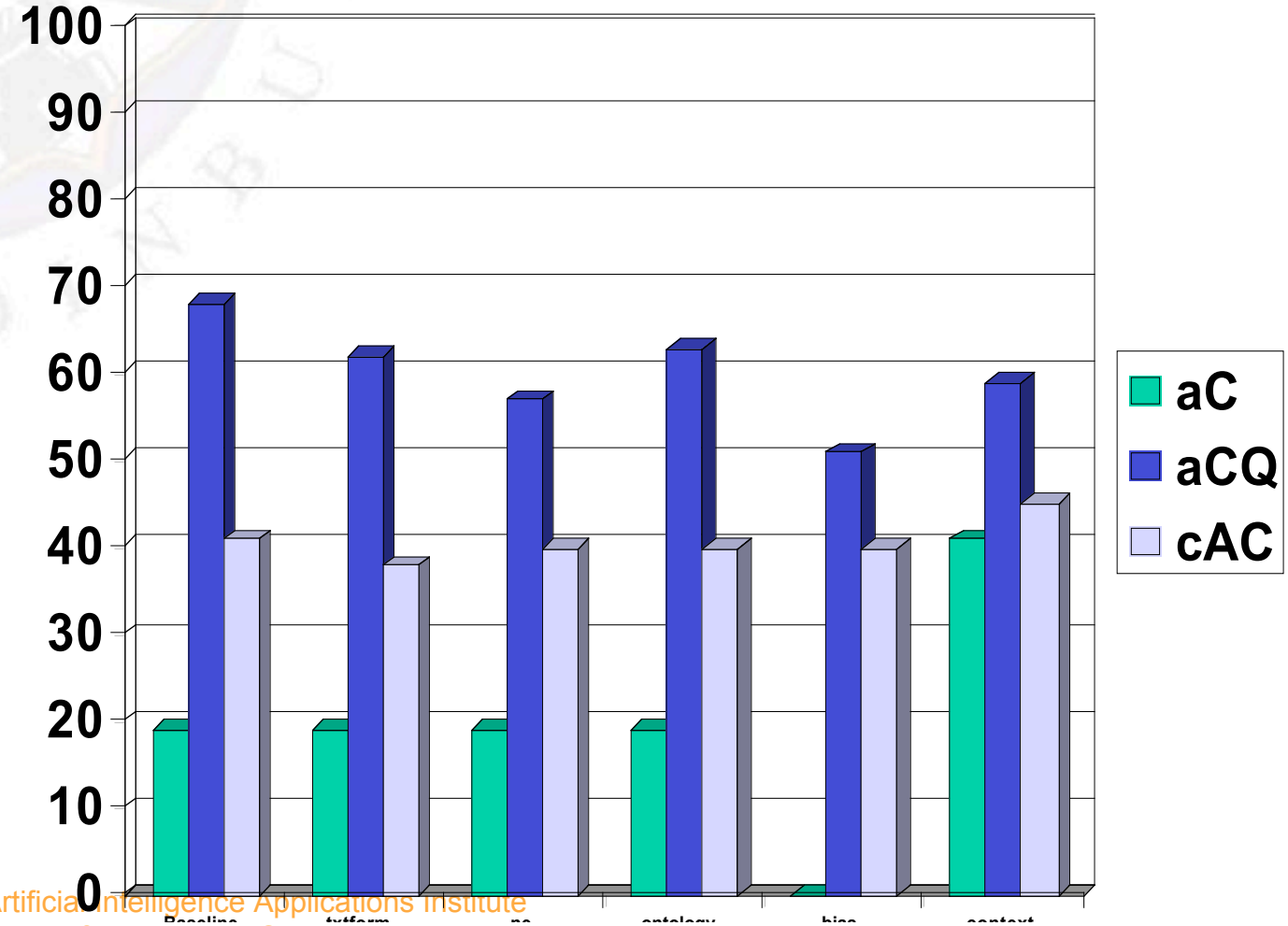
- Precision = $\left|R_{derived} \cap R_{correct}\right| \Big/ \left|R_{derived}\right|$

- Recall = $\left|R_{derived} \cap R_{correct}\right| \Big/ \left|R_{correct}\right|$

$R_{derived}$: **provable** from data+rules learned+theory

$R_{correct}$: from mark-up

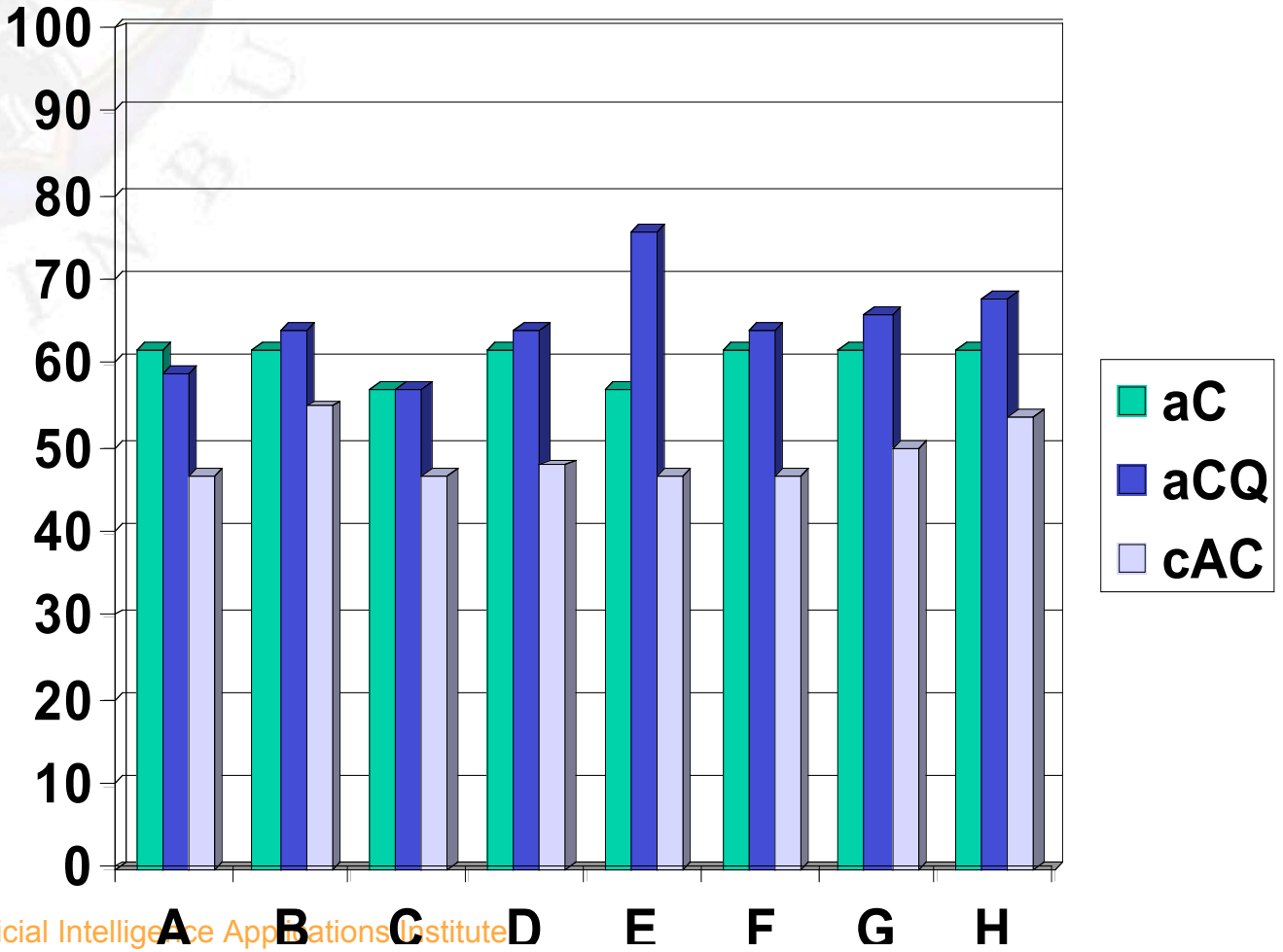
Baseline Experiment: F Scores



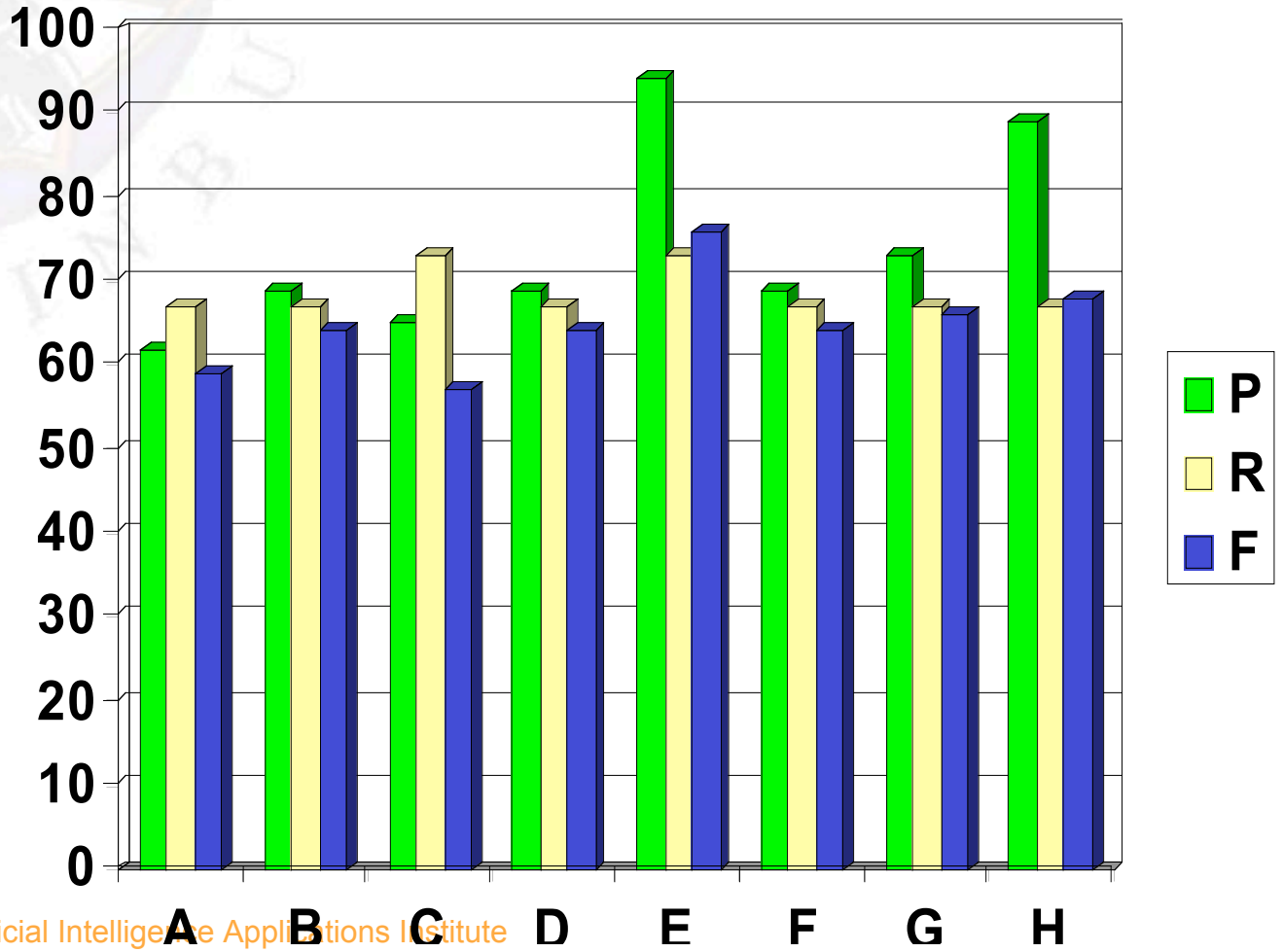
Combination Experiment

- A: bias + context
- B: + text mapping
- C: + named entity
- D: + ontology
- E: + text mapping + named entity
- F: + text mapping + ontology
- G: + named entity + ontology
- H: + text mapping + named entity + ontology

Combination Experiment: F Scores



Combination Experiment: aCQ



Results: Rules Learned

[B] cAC(A,'Methane',increase):-
hasWord(A,v,increase), txtform1('Methane',E),
hasWord(A,_,E).

[C] aCQ(A,'CarbonDioxide',C):-
hasWord(A,_,'CarbonDioxide'),
hasWord(A,_,C).

[F] cAC(A,B,increase):- hasWord(A,v,increase),
isa(B,E), txtform2(E,_,G), hasWord(A,_,G).

[G] cAC(A,B,increase):- hasWord(A,_,B),
hasWord(A,v,rise).

Analysis

- Bag-of-words + context is adequate for learning attribute relations
- More generally, grammatical information will be required
- Small data set:
  - results were repeated (but not improved) in an enlarged set of 371 sentences
  - 95% confidence interval was consistently reduced
- A knowledge engineer took 1hr 58min to produce a rule set for the same texts

Conclusions

- Attribute relations can be learned using ‘surface’ features of the text
- Small training sets can be used successfully
- Multiple knowledge sources are a benefit