Knowledge Extraction by using an Ontology-based Annotation Tool

Maria Vargas-Vera, E.Motta, J. Domingue, S. Buckingham Shum and M. Lanzoni

Knowledge Media Institute(KMi)

The Open University

Milton Keynes, MK7 6AA

October 2001

Outline

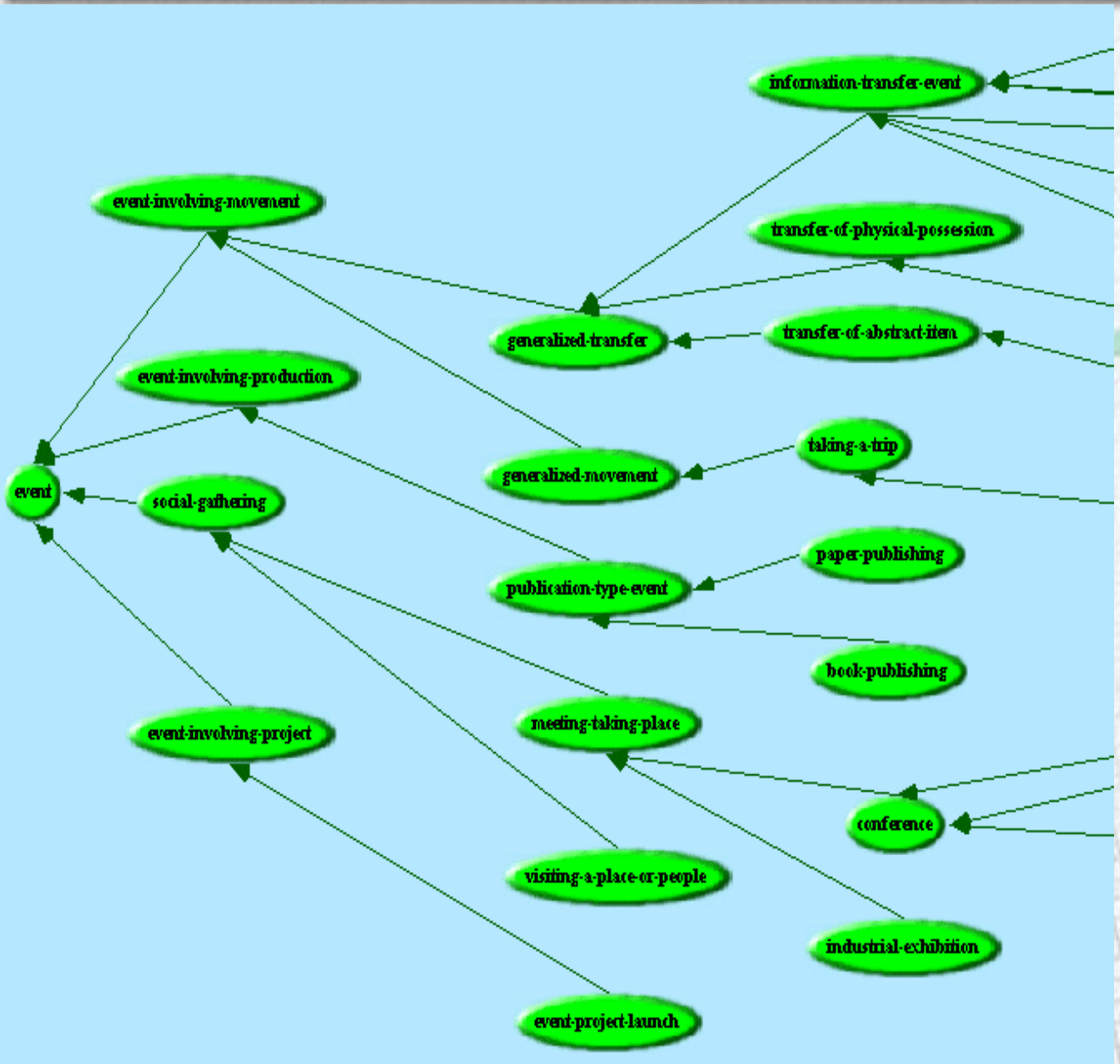
- Motivation
  - Extraction of knowledge structures from web pages
  - Final goal -Ontology population
- Approaches to semantic annotation of web pages (SAW)
  - OntoAnnotate [Stab, et al]
  - SHOE [Hendler et al]
- Our solution to SAW problem
  - Ontology driven annotation
- Work so far - we had tried with two different domains (KMi stories and Rental adverts)
- Conclusions and Future work

Our system

- Our system consists of 4 phases:
  - Browse
    - browser selection
  - Mark-up phase (mark-up text in training set)
  - Learning phase (learns rules from training set)
  - Extraction phase (extracts information from a document)

Mark-up phase

- **Ontology-based Mark-up**
  - **The user is presented with a set of tags (taken from ontology)**
  - **user selects slots-names for tagging.**
  - **Instances are tagged by the user**



EVENT 1:

visiting-a-place-or-people

visitor (list of person(s))

people-or-organisation-being-visited (list of person(s) or organisation)

has-duration (duration)

start-time (time-point)

end-time (time-point)

has-location (a place)

other agents-involved (list of person (s))

main-agent (list of person (s))

Learning phase

- Learning phase was Implemented using Marmot and Crystal.
- Mark-up all instances in the training set
- Marmot performs segmentation of a sentence: noun phrases,verbs and prepositional phrases.
- Example: “David Brown, the Chairman of the University for Industry Design and Implementation Advisory Group and Chairman of Motorola, visited the OU”.
- Marmot output:
- SUBJ: DAVID BROWN %comma% THE CHAIRMAN OF THE UNIVERSITY
- PP: FOR INDUSTRY DESIGN AND IMPLEMENTATION ADVISORY GROUP AND CHAIRMAN OF MOTOROLA
- PUNC: %COMMA%
- VB: VISITED
- OBJ: THE OU

Learning phase (cont)

- Crystal derives a set of patterns from a training corpus.
- Example of Rule generated using Crystal.
  - Conceptual Node for visiting-a-place-or-people event:
    - Verb: visited (active verb) (trigger word)
    - Visitor: V (person)
    - Has-location: P (place)
    - Start-time: ST (time-point)
    - End-time: ET (time-point)
- Example of patterns:
  - `X visited Y on the date Z`
  - `X has been awarded Y money from Z`

Extraction phase

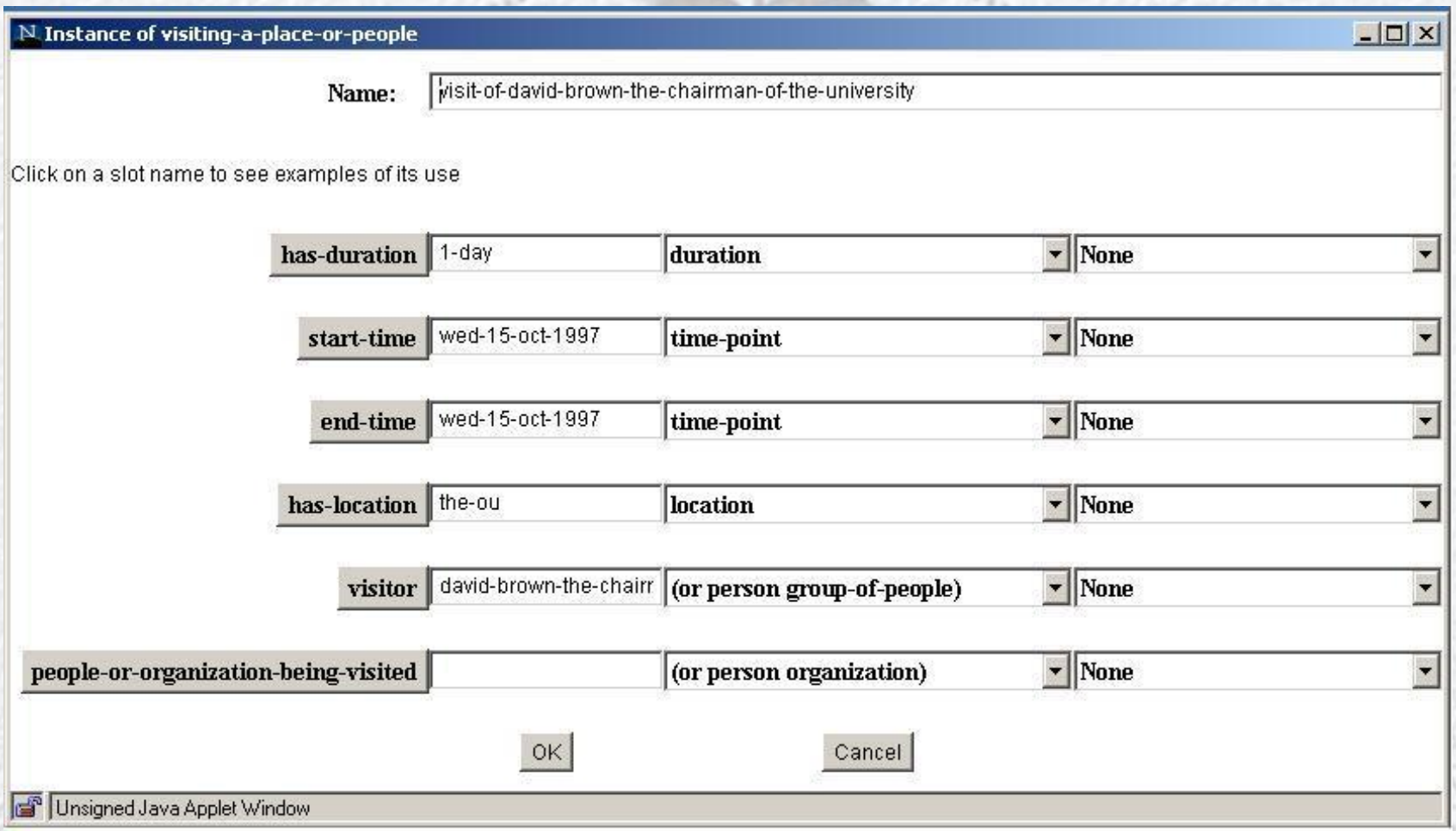
- Badger makes instantiation of templates.
- In our example (David's Brown story), Badger instanciates the following slots of a Event -1 frame:
  - Type: visiting-a-pace-or-people
  - Place: The OU
  - Visitor: David Brown

OCML code (definition of an instance of class visiting-a-place-or-people)

```
(Def-instance visit-of-david-brown-the-chairman-of-the-university
              visiting-a-place-or-people
((start-time wed-15-oct-1997)
  (end-time wed-15-oct-1997)
   (has-location the-ou)
   (visitor david-brown-the-chairman-of-the-university)
)
)
```

Populating the ontology

- **David Brown's story output after the OCML code is sent to Webonto.**



Library of IE Methods

- **Currently our library contains methods for learning:**
  - **Crystal (bottom-up learning algorithm)**
  - **Whisk (top-down learning algorithm)**

- **We plan to extend the library with other methods besides Crystal and Whisk.**

Whisk (second tool for learning)

- Whisk: learns information extraction rules
  - can be applied to semi-structured text (text is un-gramatical, telegraphic).
  - can be applied to free text (syntactically parsed text).
- It uses a top-down induction algorithm seeded by a specific training example.
- Whisk has been used:
  - CNN weather forecast in HTML
  - BigBook addresses in HTML
  - Rental ads in HTML (our second domain)
  - Seminar announcements
  - job posting
  - Management succession text from MUC-6

Sample Rule from Rental domain

- **Domain Rental Adverts:**
- **Ballard - 2 Br/2 Ba, top flr, d/w 1000 sf, $820. (206) 782-2843.**
- **Rule expressed as regular expression:**
- **ID 26 Pattern:: * (Nghbr) * (<digit>) 'Br' * '$' (<number>).**
- **Output:: Rental{Neighbourhood $1} {Bedrooms $2} {Price $3}**

Whisk example (continuation)

- Items in green colour are semantic word classes.
- Nghbr :: Ballard | Belltown| …
- digit :: 1|2|…|9
- number :: (0-9)*

- Complexity : restricted wild card therefore, time is not exponential.

Conclusions and Future Work

- We had built a tool which extracts knowledge using and Ontology, IE component and OCML pre-processor.
- We had worked with 2 different domains (KMi stories and Rental adverts)
  - first domain
    - Precision over 95%
  - second domain
    - Precision: 86% - 94%
    - Recall: 85% - 90%
- We will integrate more IE methods in our system.
- To extend our system in order to produce XML output, RDFS,…
- to integrate visualisation capabilities