Using Multiple Ontologies in Information Extraction

Daya C. Wimalasuriya (dayacw@cs.uoregon.edu)
Advisor: Dejing Dou (dou@cs.uoregon.edu)
Advanced Integration and Mining (AIM) Lab, University of Oregon

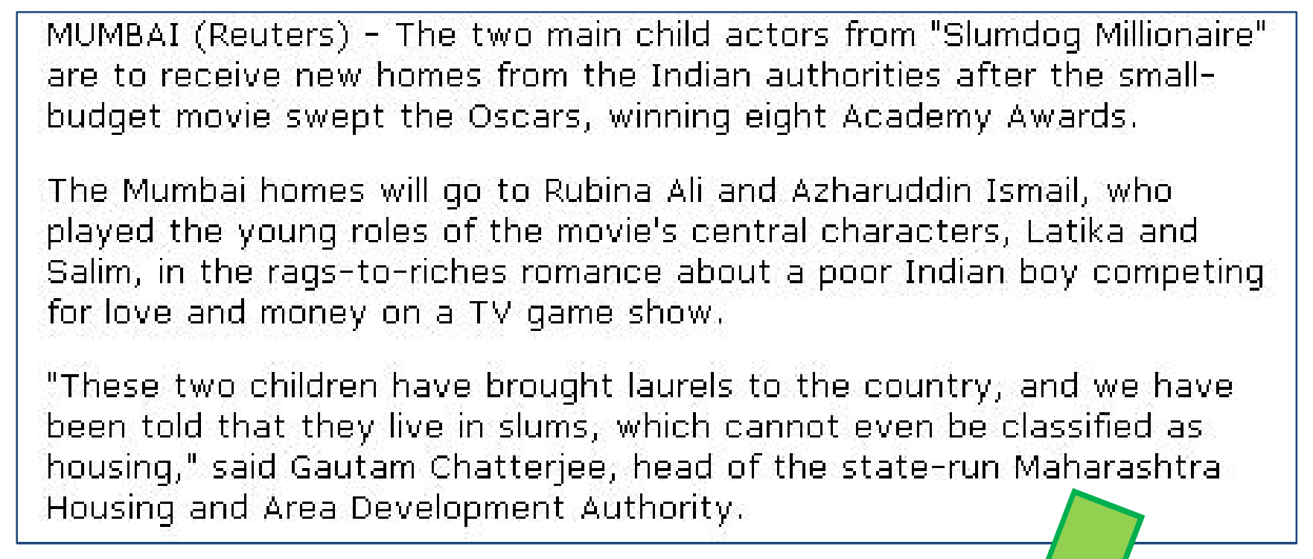
Introduction

Information Extraction (IE) systems recognize and extract *certain types of information* from text.

An Ontology is a formal and explicit *specification* of a shared *conceptualization*.

Ontology-Based Information Extraction (OBIE) has recently emerged as a subfield of information extraction. Here ontologies are used to *guide* information extraction.

An article on a movie

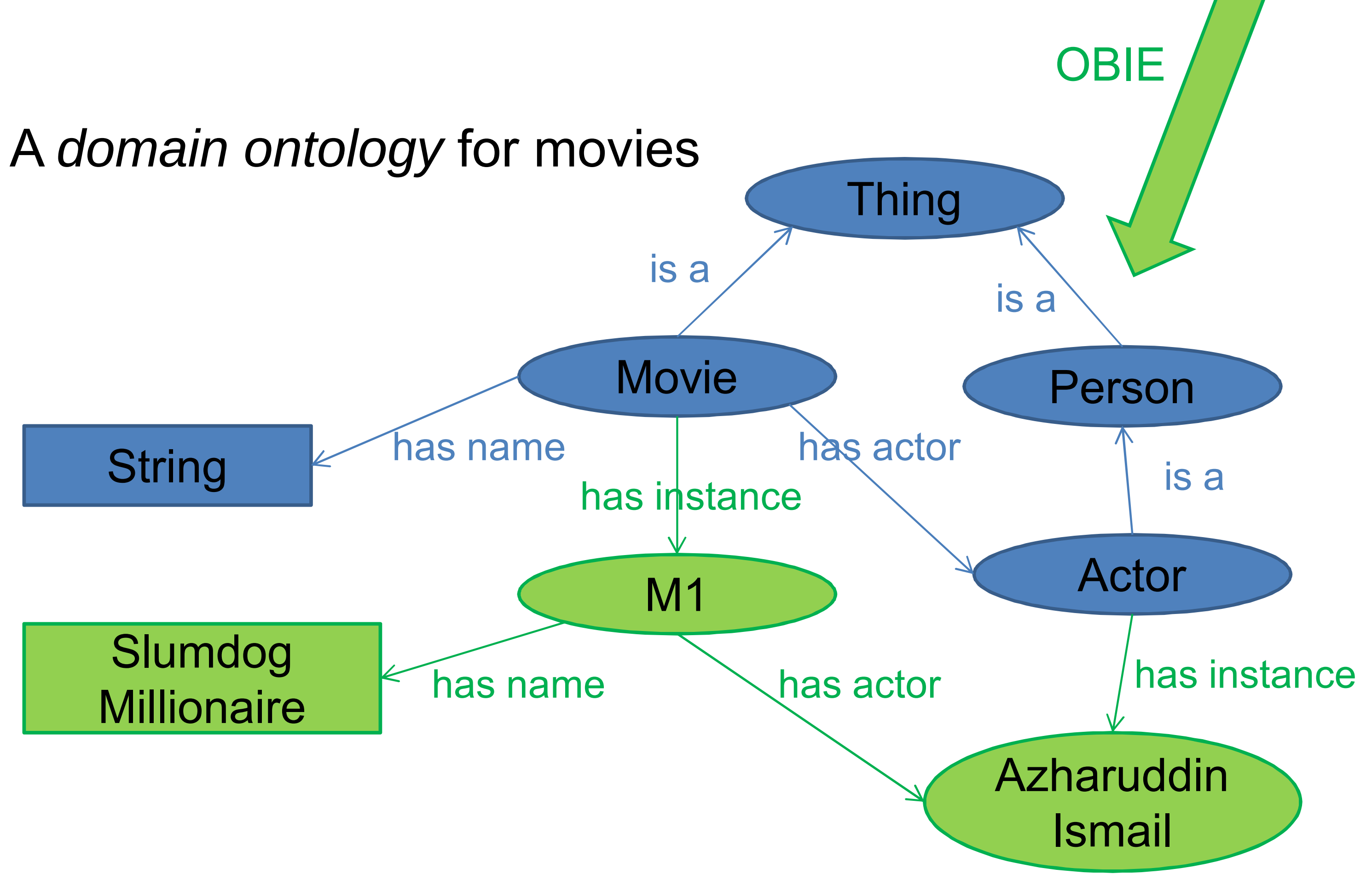


MUMBAI (Reuters) - The two main child actors from "Slumdog Millionaire" are to receive new homes from the Indian authorities after the small-budget movie swept the Oscars, winning eight Academy Awards.

The Mumbai homes will go to Rubina Ali and Azharuddin Ismail, who played the young roles of the movie's central characters, Latika and Salim, in the rags-to-riches romance about a poor Indian boy competing for love and money on a TV game show.

"These two children have brought laurels to the country, and we have been told that they live in slums, which cannot even be classified as housing," said Gautam Chatterjee, head of the state-run Maharashtra Housing and Area Development Authority.

A *domain ontology* for movies



All previous OBIE systems use *a single ontology*.

Reasons for using multiple ontologies in OBIE:

1. Improving *recall*

$$\text{Recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|} \qquad \text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

2. Supporting different views represented by the different ontologies

Theory

An ontology consists of several components such as

- **classes** (e.g., *Actor*)
- **properties** (e.g., *has actor*)
- **individuals** (e.g., *M1*)
- **values** (e.g., *M1* has name *"Slumdog Millionaire"*)

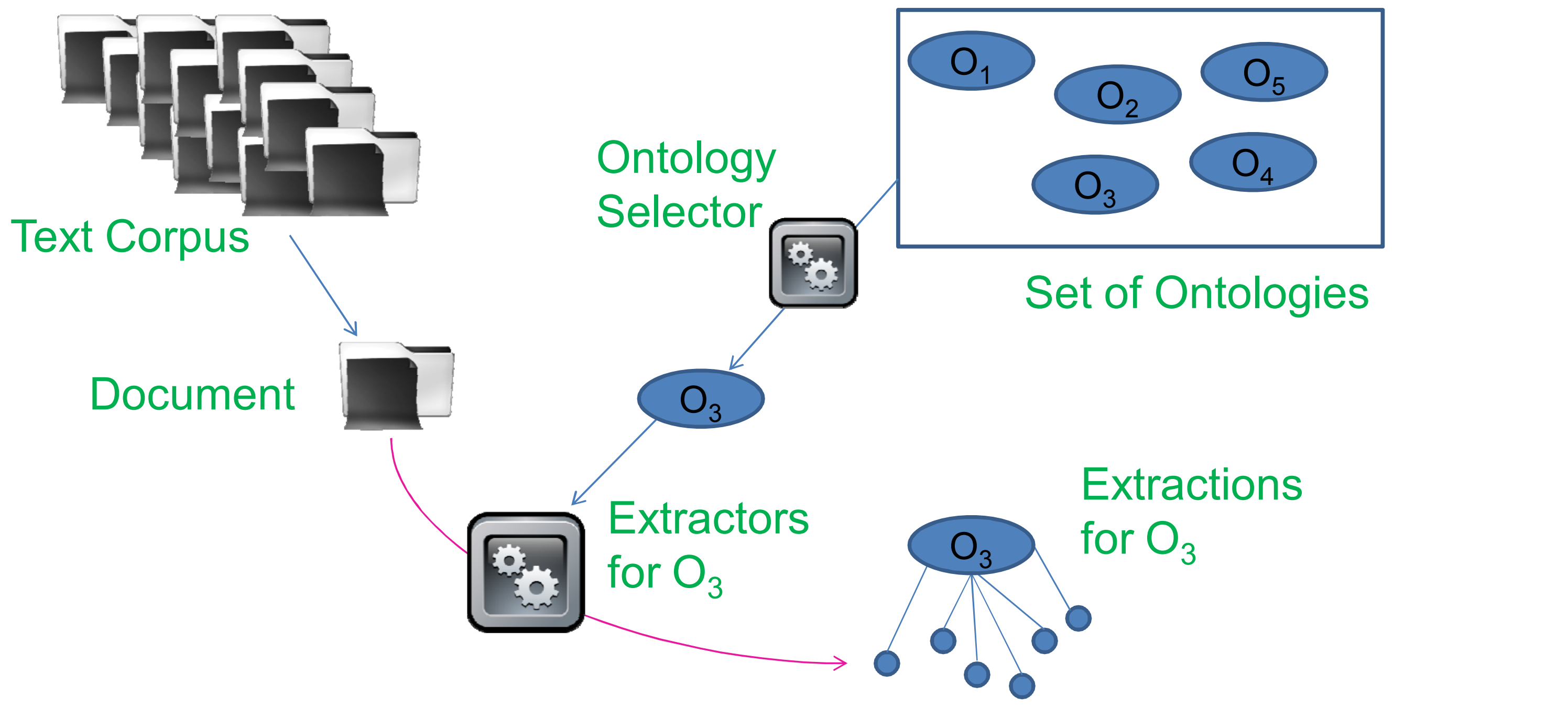
We define an OBIE system as a **set of information extractors**, each identifying

- *individuals for a class* or
- *values for a property* of the ontology in concern.

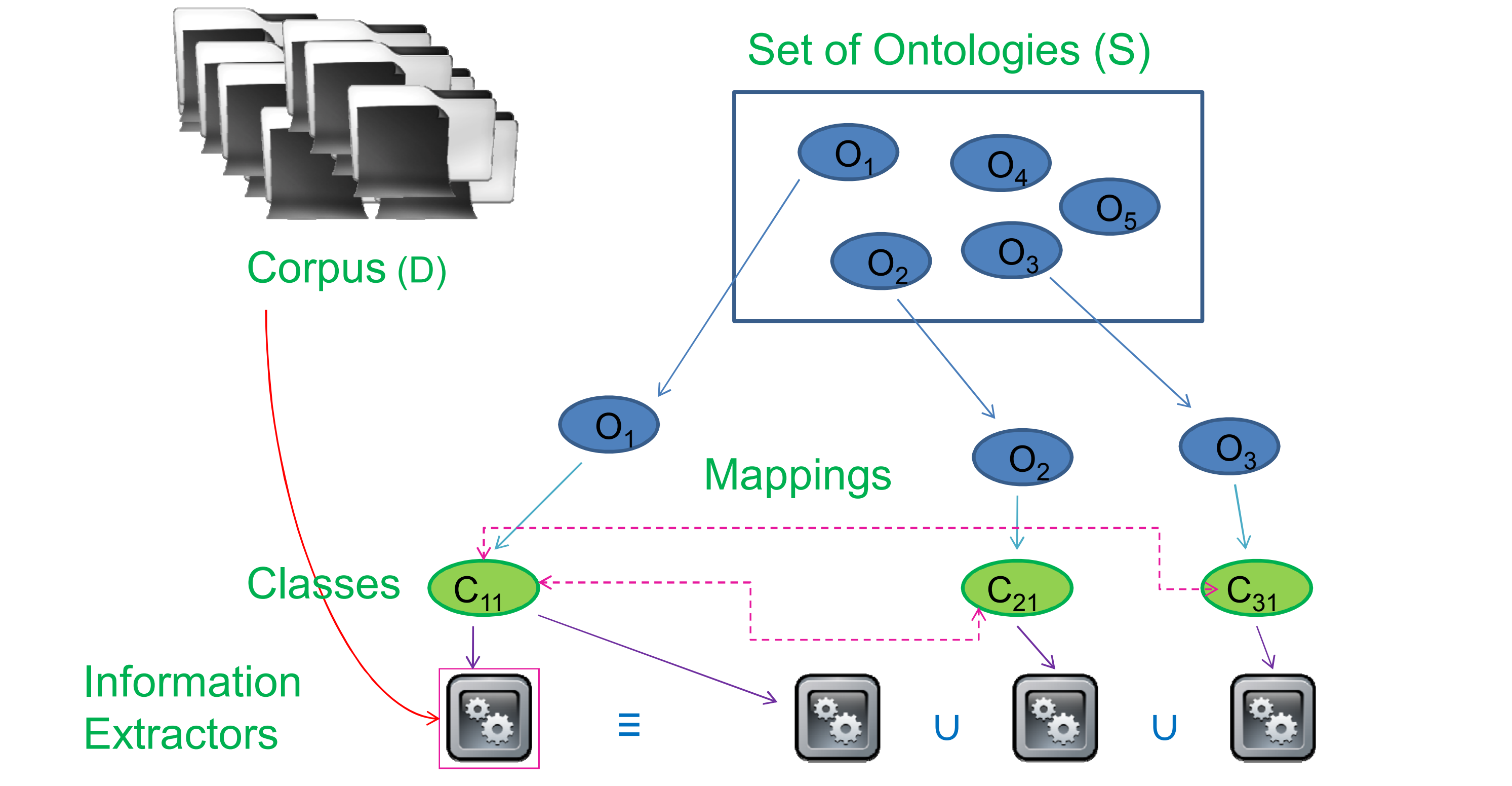
Scenarios for having multiple ontologies for a domain:

1. *Specializing in sub-domains: e.g.,* North American universities and British universities for the domain of universities.
2. *Providing different perspectives: e.g.,* classes for *"Husband*" and *"Wife"* vs. a property "isSpouseOf" for the domain of marriages.

Using multiple ontologies specializing in sub-domains in OBIE*: Assign* an ontology for each document of the text corpus.



Using multiple ontologies providing different perspectives in OBIE: *Reuse* information extractors based on *mappings* between the concepts of the ontologies and take the *union* of the results.



Experiments

Case Study 1: University Ontologies

Ontologies:

1. An ontology for North American universities
2. An ontology for universities of other regions
3. A common ontology for the entire domain

Text Corpus:
Selected web pages from 100 university web sites

Information Extraction Technique:
Linguistic extraction rules

Results:

System	Domain	Precision (%)	Recall (%)	F1 (%)
Single Ontology	North America	52.86	37.00	43.53
	Other Regions	47.83	52.38	50.00
	All Universities	50.86	41.55	45.74
Multiple Ontology	North America	54.65	44.34	48.96
	Other Regions	52.17	57.14	54.54
	All Universities	53.79	47.97	50.71

Case Study 2: Terrorism Ontologies

Ontologies:

1. An ontology derived from the structure of the key files of the 4th Message Understanding Conference (4th MUC)
2. An ontology from the Mindswap group of University of Maryland

Text Corpus:
200 files from the text corpus of the 4th MUC

Information Extraction Technique:
Classification

Results: (for Mindswap ontology)

System	Class (Scope)	Precision (%)	Recall (%)	F1 (%)
Single Ontology	Agent	31.73	41.75	36.06
	Organization	59.39	28.32	38.35
	All Classes	41.92	33.47	37.22
Multiple Ontology	Agent	32.72	56.49	41.44
	Organization	51.94	40.74	45.06
	All Classes	40.85	46.77	43.61

Conclusion

Experimental results support our hypothesis that using multiple ontologies in OBIE results in a higher recall.

Future Work

- Conducting more case studies on different ontologies and different text corpora to verify that the results are generic.
- Designing a *component-based approach for OBIE* based on the reuse of *information extractors*.

Reference: Daya C. Wimalasuriya and Dejing Dou, Using multiple ontologies in information extraction, In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), pp. 235-244.