A Case-based Approach to Content Analysis in Cross-Domain Information Sharing

Thomas Reichherzer, Badri Lokanathan, Ashwin Ram
Enkia Corp., Atlanta GA

Research support by Airforce Research Lab (AFRL), Rome, New York, under
SBIR Phase I (Contract number FA8750-08-C-0110) and
SBIR Phase II (Contract number FA8750-07-C-0117)

Presentation Outline

- Cross-Domain Information Sharing & Challenges.
- Automated Approaches for Reliable Human Review.
- Experimental Evaluation.
- Discussions and Conclusion.

Cross-Domain Information Sharing

- Government and industry alike benefit from information sharing.

  industry:
  - develop and expand new partnerships among business partners,
  - public relations

  government:
  - exchange of mission-critical information across different agencies
  - freedom-of-information act

- Sharing takes place across institutional boundaries or security domains.
  - information cannot be freely shared

Reliable Human Review

- Information must be reviewed to remove sensitive content.
  - released information must be in compliance with non-disclosure policies across security domains
  - policies guide release of information
- Information review completed by review officers (e.g. FDO).
  - reviewer identifies sensitive content in document to be removed priority release
  - review process is time intensive and requires significant human expertise
  - policies are complex and subject to changes

Reliable Human Review (RHR) presents a significant bottleneck to "just-in-time" information needs.

Just-Enough Information Sharing

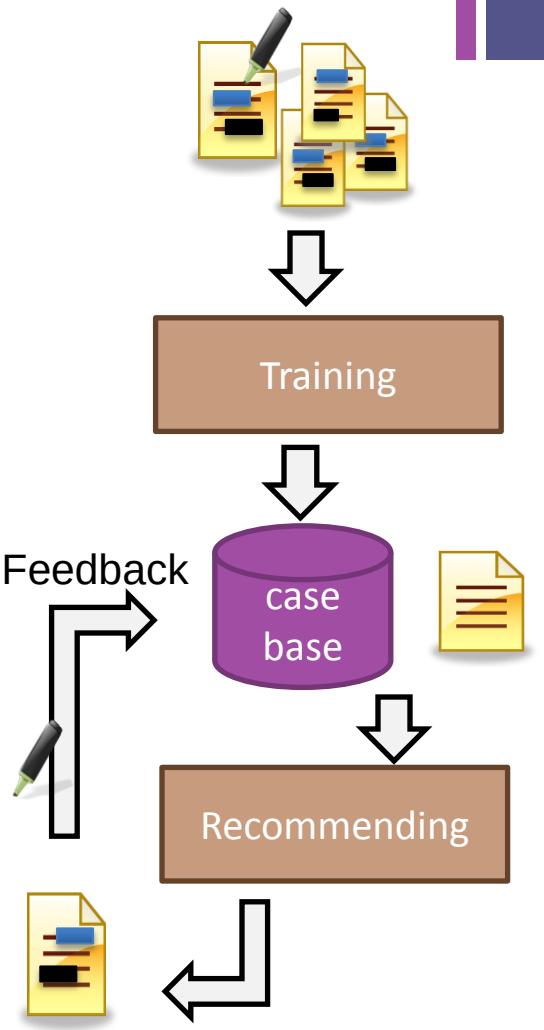
- **Problem**:
  - Identifying shareable information in documents is time consuming and laborious.
  - Security policies are high-level and difficult to capture by rules.
  - Timely dissemination of appropriate information is critical in crisis situations.

- **Need**:
  - Tools for assistance with the classification of information across multiple security domains.
  - Tools to develop and apply security policies.

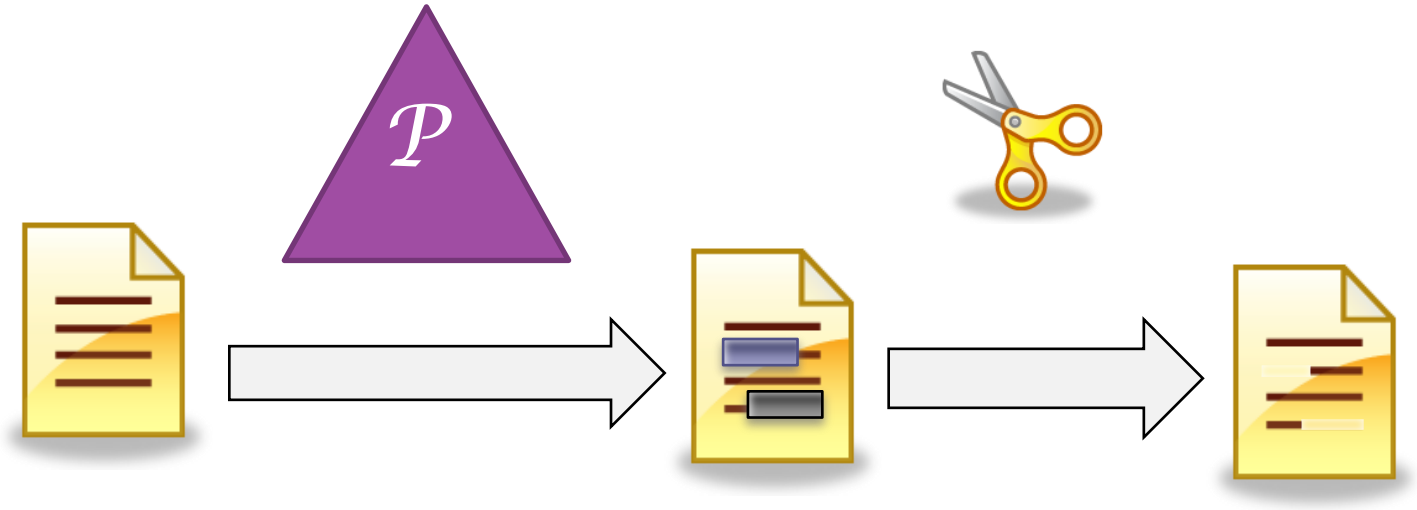
- **Proposed Approach**:
  - Assist RHR by automatic text classification of unstructured text.
  - A combined approach of Natural Language Processing (NLP) and Case Based-Reasoning (CBR) to automate process of selecting sensitive content.

Assisted RHR: Automation Steps

- Select documents that have been marked up by RHR.
  - mark-up indicates sensitive content with respect to release domain
  - mark-up captures non-disclosure policies
- Feed marked up documents into a text classifier.
  - "learn" security policies from mark-up information
  - mark-up is labeled into categories
  - train different classifiers for different release domains
- Apply classifier to unmarked documents.
  - use feedback from RHR to adjust classifier



Sanitizing Unstructured Text Workflow



Analyst opens/authors document in MS Word or Web Editor.

Analyst selects policy to be applied.

Analyst reviews automatically generated markup.

Analyst edits text to enforce selected policy.

Policy Creation Workflow

Act of Terrorism

Defense

Defense

Acts of Terrorism

Secret

Analyst reviews existing markup.

Analyst identifies markup categories.

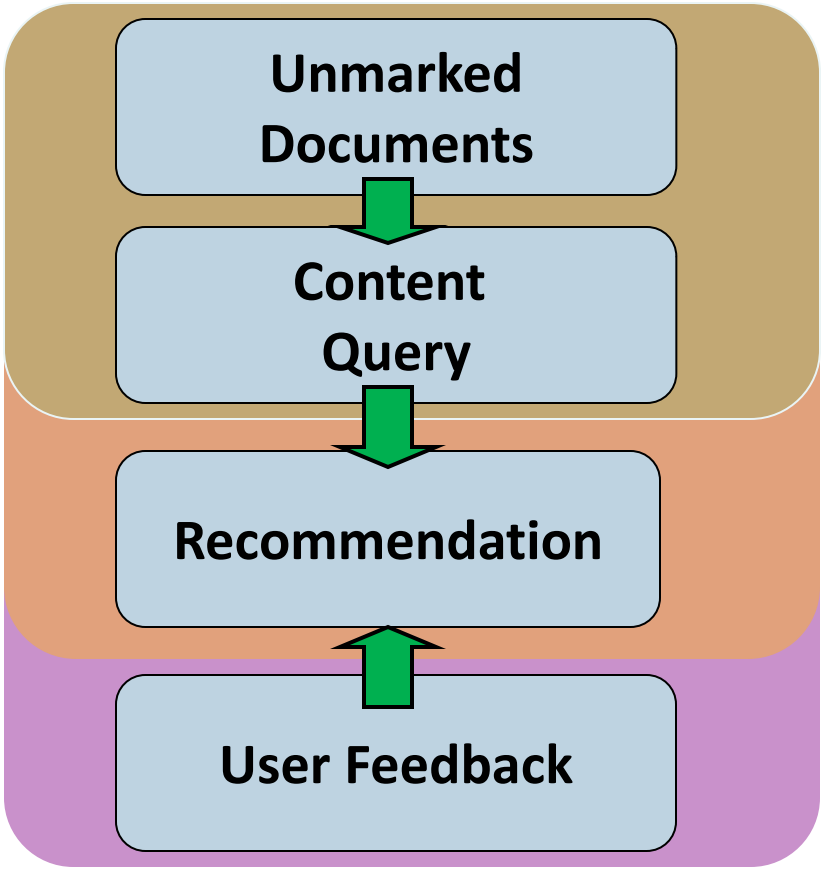
Analyst creates release policies.

Problem-Solving Steps

Text Analysis

Decision-Making

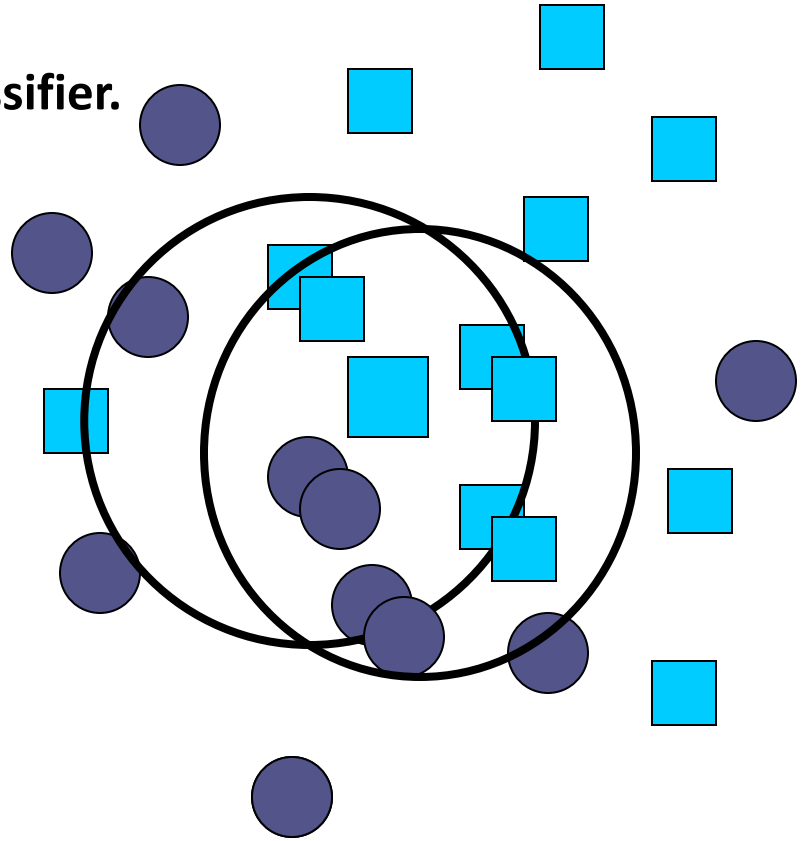
Learning



Mark-up recommendations are generated at sentence level!

How Recommendations are generated

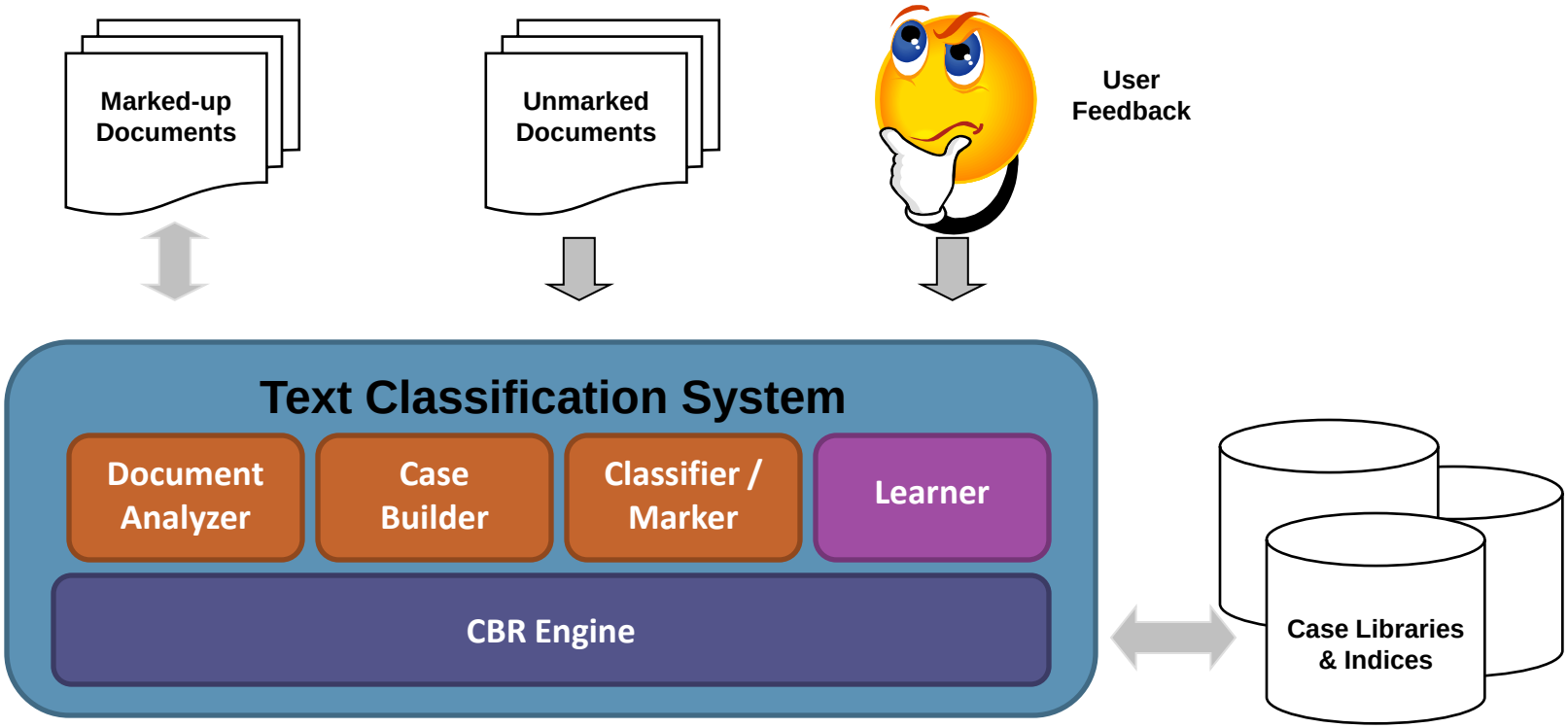
1. **User gives classifier some examples.**
2. **User presents a new problem to classifier.**
3. **Classifier retrieves similar cases.**
4. **Classifier decides on classification.**
5. **Classifier gives a recommendation.**



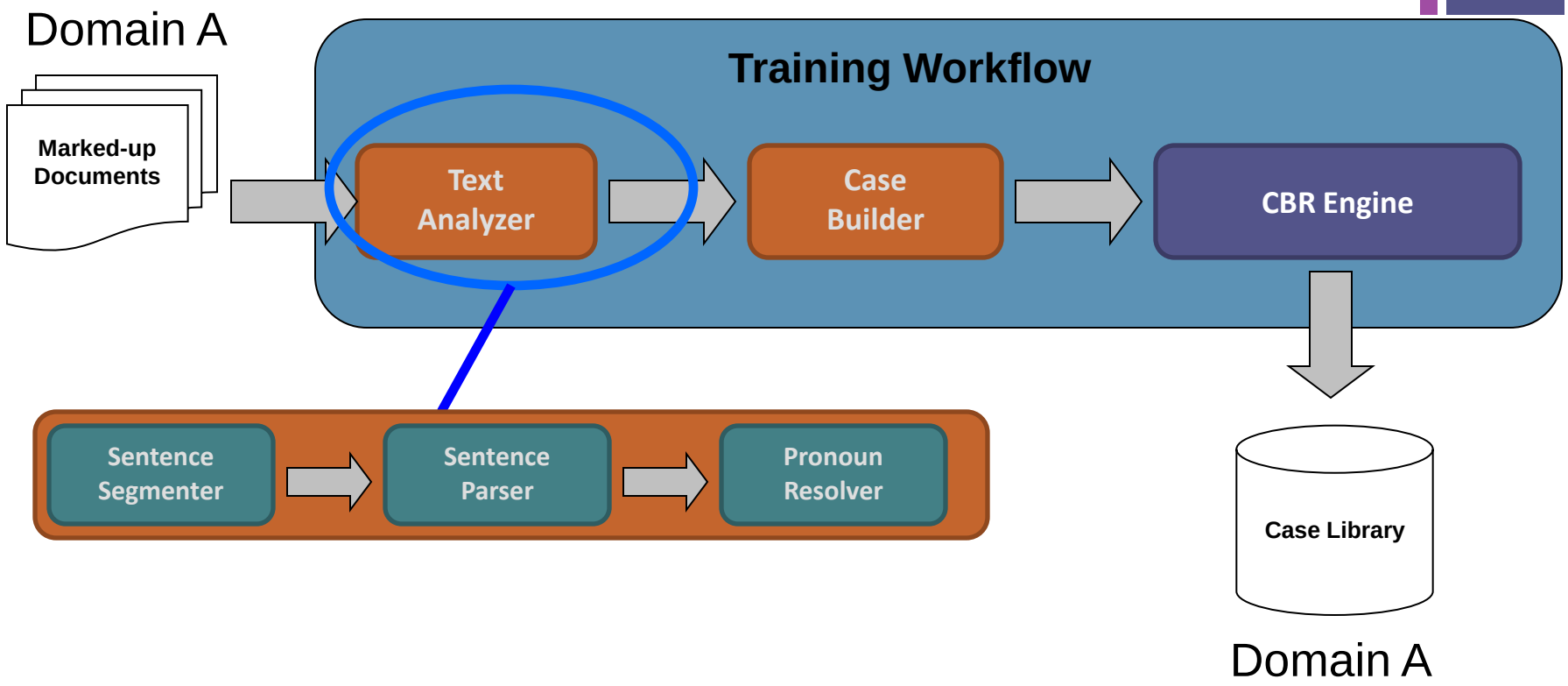
Sensitive

Non-sensitive.

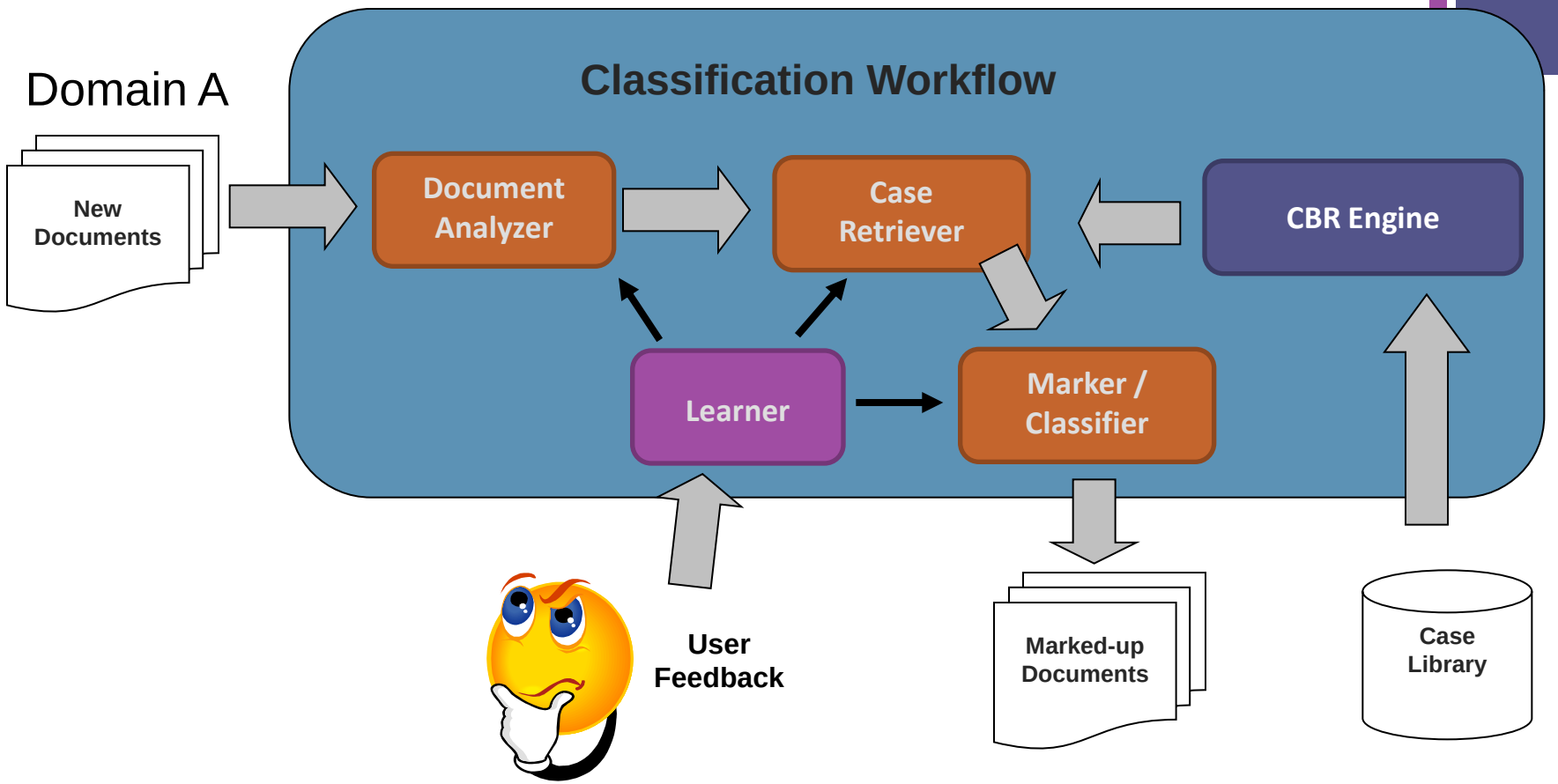
Architecture Overview



Building a Case Base



Generating Recommendations



Input Sentence - Example

BEARCLAW aircraft operating inside Friendlandia along the Narcotica border have intercepted communications indicating the site of a large heroin processing facility at PK848972 approx 4km north of the village of Lago Springo.

information not to be disclosed

Case Construction - Parsing

[[NP BEARCLAW aircraft] [VP operating inside Friendlandia along the Narcotica border]] [VP [VP have intercepted] [NP communications] [S [VP indicating] [NP the site of a large heroin processing facility] [PP at PK848972 approx 4km north of the village of Lago Springo.]

Case Construction - Mapping

[[PROBLEM
[SUBJECT BEARCLAW aircraft]
[SUBJECT_SUB1 operating inside Friendlandia along the Narcotica border]
[PREDICATE have intercepted]
[OBJECT communications]
[OBJECT_SUB1 indicating the site of a large heroin processing facility at PK848972 approx 4km north of the village of Lago Springo]]
[SOLUTION
[MARKUP operating Friendlandia Narcotica border]
[CLASSIFICATION NDP-1 Category 1?]]

Feature Vector of Case

[[PROBLEM
[SUBJECT w_1 w_2 … w_n]
[SUBJECT_SUB1 w_1 w_2 … w_n]
[PREDICATE w_1 w_2 … w_n]
[OBJECT]
[OBJECT_SUB1 w_1 w_2 … w_n]]
[SOLUTION
[MARKUP w_1 w_2 … w_n]
[CLASSIFICATION c]]

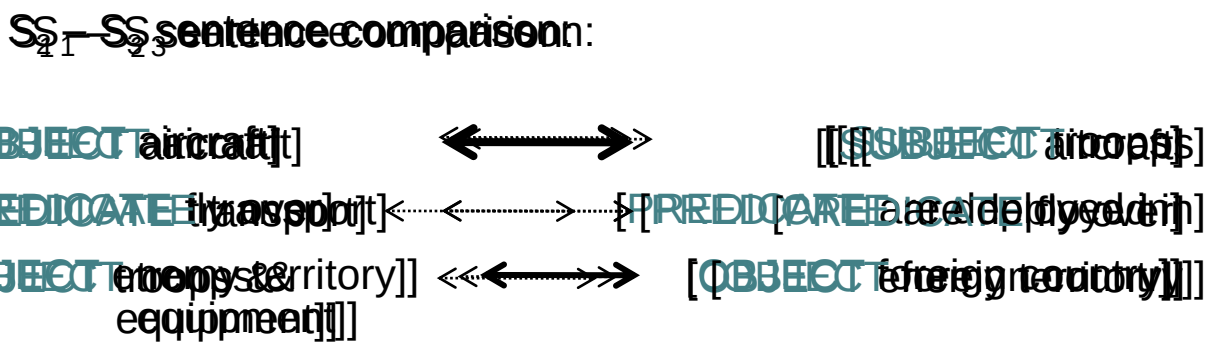
Features: SUBJECT, SUBJECT_SUB1, PREDICATE, OBJECT, OBJECT_SUB1, …

Distance between Sentences

S_1: aircraft transport troops & equipment.

S_2: aircraft fly over enemy territory.

S_3: troops are deployed in foreign country.



Distance Metric (Sentences)

Let $S_1 = \langle f_{11}, f_{12}, f_{13}, f_{14}, f_{15} \rangle$

$S_2 = \langle f_{21}, f_{22}, f_{23}, f_{24}, f_{25} \rangle$

features

$$D(S_1, S_2) = k_1 sim(f_{11}, f_{21}) + k_2 sim(f_{12}, f_{22}) + k_3 sim(f_{13}, f_{23}) + k_4 sim(f_{14}, f_{24}) + k_5 sim(f_{15}, f_{25})$$

Parameter: k_1, k_2, k_3, k_4, k_5

Distance Metric (Words)

If $f_{11} = [w_{111}, w_{112}, \ldots, w_{11n}]$ and

$f_{21} = [w_{211}, w_{212}, \ldots, w_{21n}]$

where

$sim(f_{11}, f_{21}) = sim(w_{111}, w_{211}) +$

$sim(w_{111}, w_{212}) + \ldots + sim(w_{11n}, w_{21n1})$

Lin Thesaurus

Performance Evaluation – Data Set

- IMdb database.
  - movie descriptions on selected movies rated PG, PG-13, and R
  - description was manually classified using 6 different categories:

Category	Number of cases
general violence	65
graphic	730
nudity	498
drug use	349
dark topic	298
sexual content	6

Performance Evaluation – Training and Testing

- Train classifier using marked-up information from movie descriptions.
  - Condition A: use all mark-up data or 1946 cases.
  - Condition B: use 90% of the mark-up data (randomly selected) or 1752 cases.
- Error rate:

	Condition A	Condition B
miss-labeled	30.5%	33.1%
not-labeled	1.6%	1.6%

- Under condition B, some cases generalize.

Conclusions

- Information sharing is critical; automated methods are needed.
  - methods need to go beyond keyword checking
- Proposed approach captures human expertise in classifying information.
  - policies are indirectly captured as cases in case base
- Markup and classification generated at sentence level, not document level.
- Direct feedback from reviewer refines and revises system's classification knowledge.
- Scalability affected by larger training sets.
  - more examples improve accuracy
  - more examples slow down classification