



Information Security & Sharing

Dr. Elizabeth D. Liddy

Director

Center for Natural Language Processing
School of Information Studies
Syracuse University



A Concern? Your Concern?

- **The role of information specialists has broadened in recent years:**
 - Does it include responsibility for the security of the intellectual property of the organization?



Information Security

1. Protection from intruders

- Interruption of Service
- Protection of IP addresses
- Intrusion Detection

2. Protection from unwanted release to inappropriate recipients

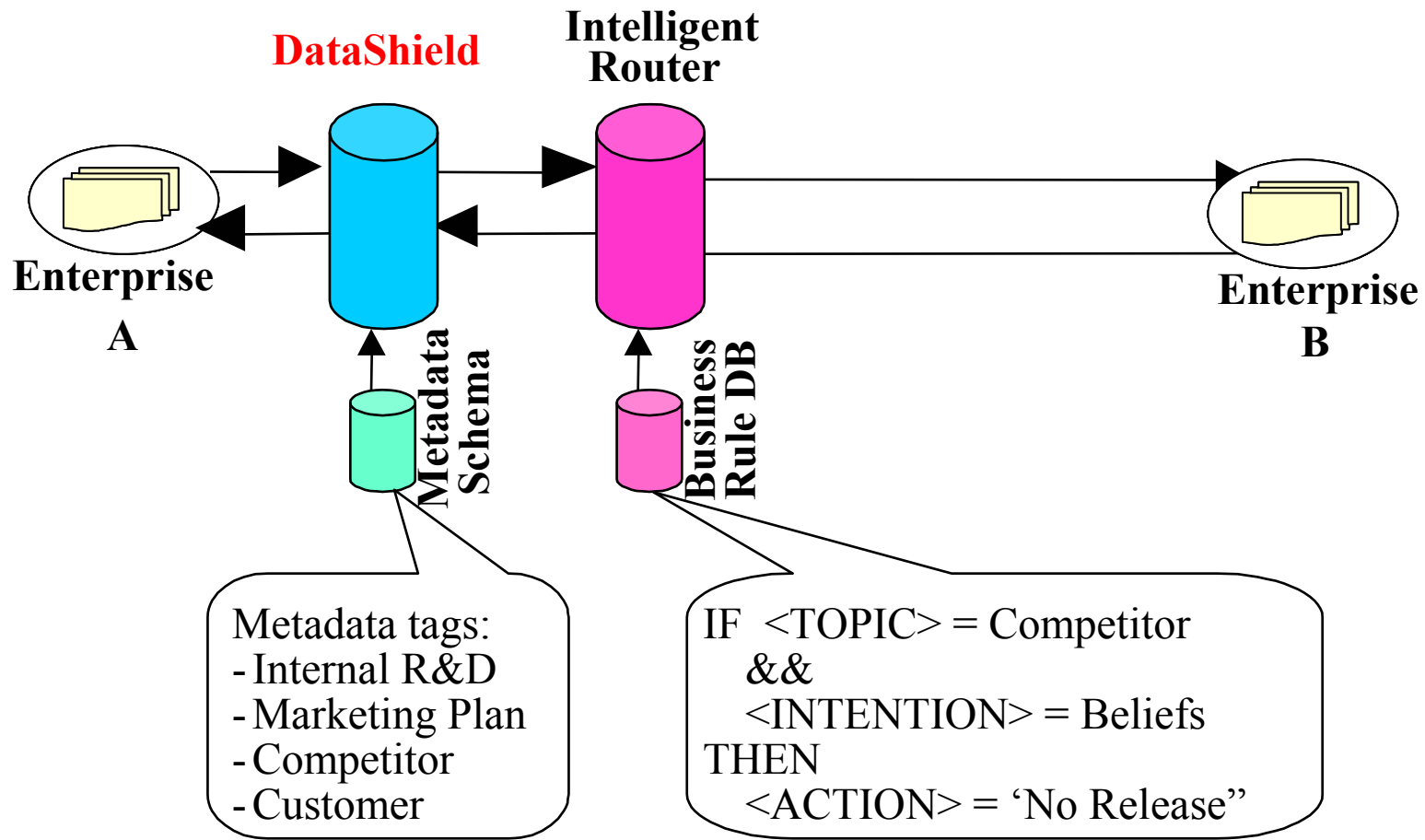
- Boundary Controllers



Boundary Controllers

Purpose is to prevent the unauthorized release of sensitive information:

- **Groups that must share some information use them to interconnect units that operate at different clearance or *need-to-know* levels**
- **Designed to enforce the Business Rules of an organization and to control the information flows between**
 - internal units
 - **the organization & the outside world**

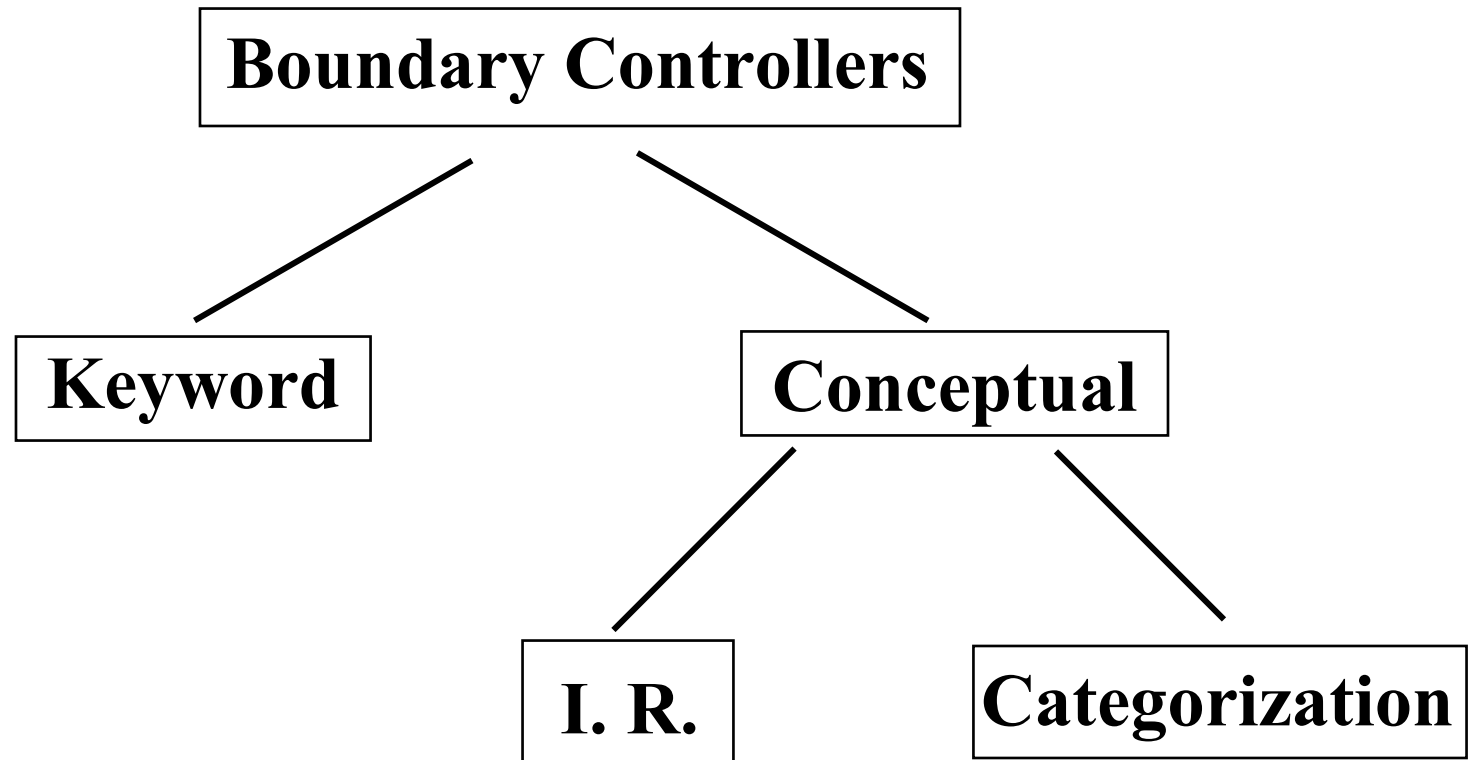




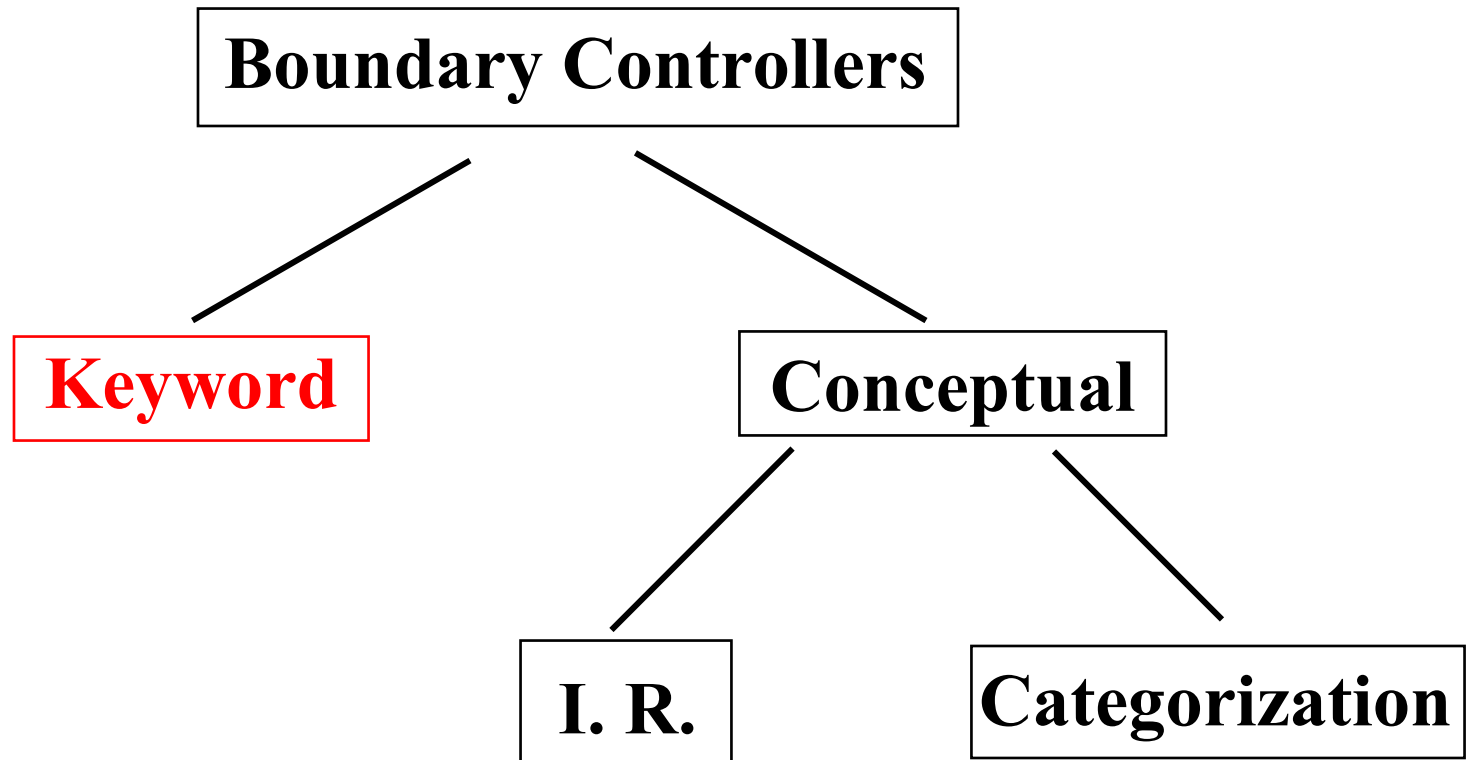
Imagine these situations:

- **Your company has decided to do R & D on a new approach to solving a known, expensive problem in their industry**
- **You're Microsoft and attempting to make life difficult for your competitors**
- **You're leading a military team within a multi-national force and need to share necessary info with other teams, but not all info**

Possible Approaches



Possible Approaches

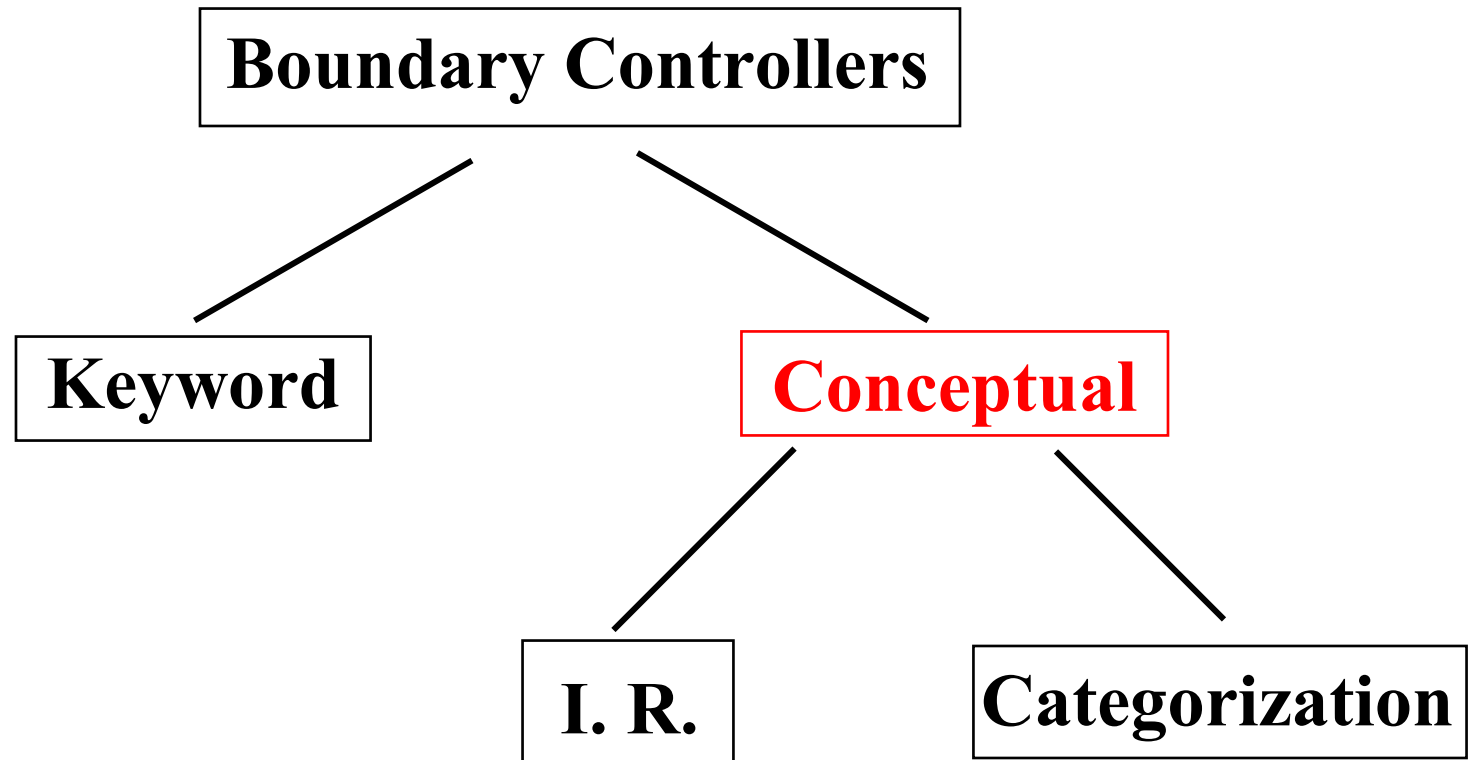




Keyword Approach

- **“Dirty Word List”**
 - Organization prepares a list of words, which if contained in a message / document, should prevent the sending of that document.
- **Process**
 - Outgoing messages / documents are tokenized into words & the words compared against the list. If there is any match, the document is not released.
- **Limitations**
 - A word can express many meanings (ambiguity)
 - Same meaning can be expressed by different words (synonymy)

Possible Approaches





Conceptual Approaches

- **Release/access decision is made at conceptual level**
 - what policies & documents *mean* and not simply the occurrence of so-called ‘dirty’ words
- **Will improve precision of the boundary controller in 2 ways:**
 - fewer documents that violate security policies / business rules will be released
 - more documents that do **not** violate policies / rules will be passed
- **More efficient than use of a human security monitor**
 - process documents and make decisions in far less time and for far less cost



Our Conceptual Approach

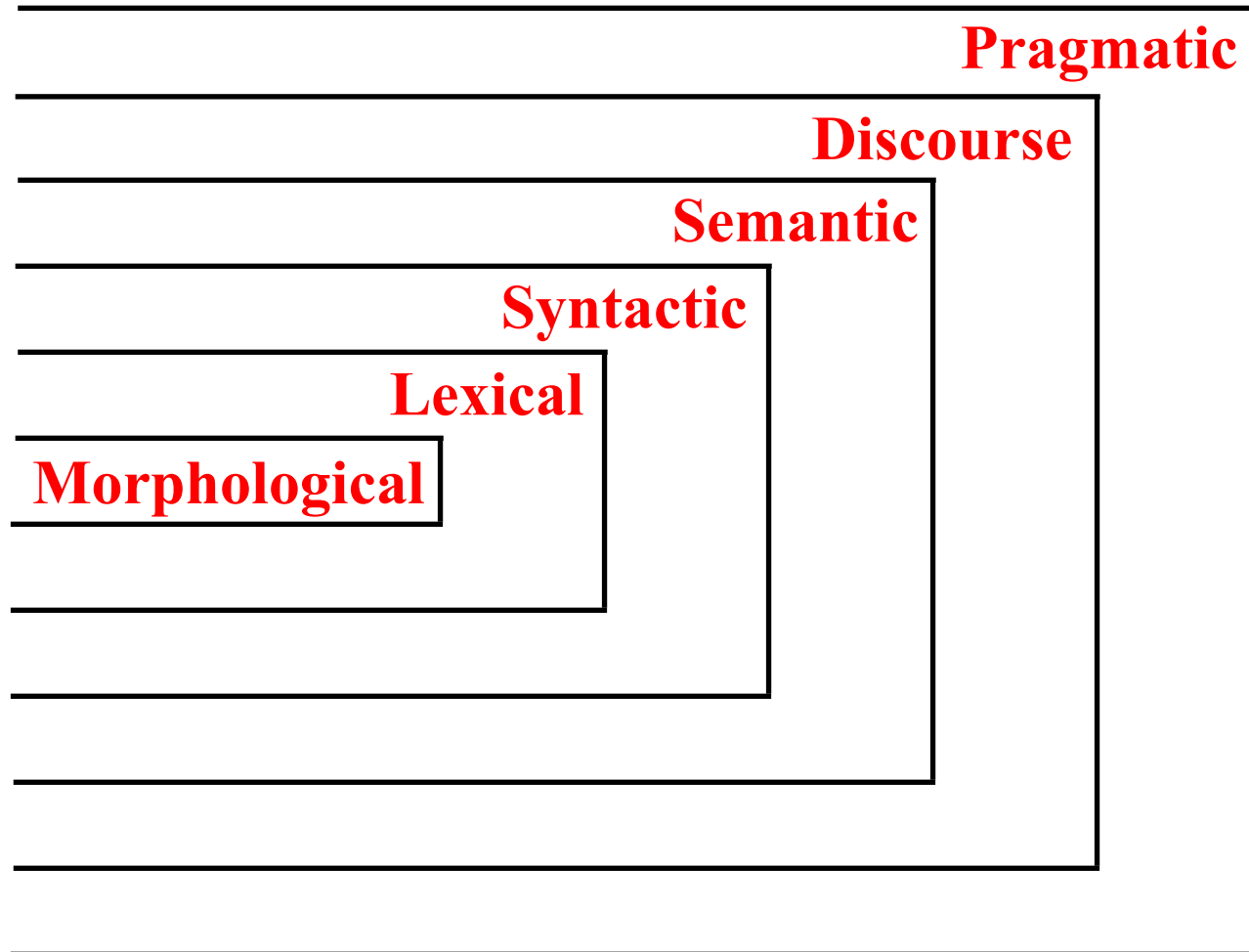
- **Based on a rich NLP interpreter**
 - Determine whether documents are ‘releasable’ or ‘unreleasable’ based on semantic comparison to business rules
 - Successfully prevent / permit passage of a document in accordance with the business rules
 - Multiple ways to do the comparison, each using rich features of text



Natural Language Processing

- **an approach which enables a system to accomplish human-like access to valuable information**
- **extracts both explicit and implicit meaning**
- **utilizes all levels of human language understanding when representing the contents of text**

Levels of Language Understanding





NLP Features for Boundary Controllers

- **Morphologically analyzed words**
- **Correctly POS-tagged words**
- **Phrasal concepts identified**
- **Proper Named entities categorized**
- **Semantic expansions incorporated**



Morphological Analysis

“Junior employees of the Acme Corporation must not describe specifications of company products in outgoing emails.”

Junior employee(s) of the Acme Corporation must not describe specification(s) of company product(s) in outgoing email(s).

Part-Of-Speech Tagging

“Junior employees of the Acme Corporation must not describe specifications of company products in outgoing emails.”

Junior|**ADJ** employee|**NN** of|**PREP** the|**ART**
Acme|**PN** Corporation|**PN** must|**MOD** not|**MOD**
describe|**VBG** specification|**NN** of|**PREP**
company|**ADJ** product|**NN** in|**PREP**
outgoing|**ADJ** email|**NN** .|.

Phrasal Concepts Identified

“Junior employees of the Acme Corporation must not describe specifications of company products in outgoing emails.”

<CN> Junior_employee </CN> of|PREP the|ART
Acme|PN Corporation|PN must|MOD not|MOD
describe|VBG specification|NN of|PREP <CN>
company_product </CN> in|PREP <CN>
outgoing_email </CN> .|.

Proper Named Entities Categorized

“Junior employees of the Acme Corporation must not describe specifications of company products in outgoing emails.”

<CN> Junior_employee </CN> of|PREP the|ART
<PN> Acme_Corporation; type=Company </PN>
must|MOD not|MOD describe|VBG
specification|NN of|PREP <CN> company_
product </CN> in|PREP <CN> outgoing_email
</CN> .|.

Semantic Expansions

“Junior employees of the Acme Corporation must not describe specifications of company products in outgoing emails.”

<CN> (Junior_employee; new_hire; level_1-to-6)
</CN> of|PREP the|ART <PN> Acme_
Corporation; type=Company </PN> must|MOD
not|MOD (describe; tell; explain) |VBG
(specification; size) |NN of|PREP <CN>
(company_product; Prod_Names) </CN>
in|PREP <CN> (outgoing_email; messages;
postings) </CN> .|.



Logical - Semantic Representation

“Junior employees of the Acme Corporation must not describe specifications of company products in outgoing emails.”

If ISA (?X, junior_employee) and ISA (?Y, Acme_product) and ISA (?Z, email) and RCPT (?Z, ?P) and LOC (?P, outside_network) and CONT (?Z, ‘ASSOC (?Y, ?A) & MEAS (?A, ?B)’), then CHRC (?Z, nonreleasable).



Logical – Semantic Representation (cont'd)

“Junior employees of the Acme Corporation must not describe specifications of company products in outgoing emails.”

If **ISA** (?X, junior_employee) and **ISA** (?Y, Acme_product) and **ISA** (?Z, email) and **RCPT** (?Z, ?P) and **LOC** (?P, outside_network) and **CONT** (?Z, ‘ASSOC (?Y, ?A) & MEAS (?A, ?B)’), then **CHRC** (?Z, nonreleasable).



Sample Message

Sally Tobias, a recent hire at Acme, sends a message to Jerry Haviland, who is not affiliated with Acme:

“I’ve really enjoyed working on the Levantar 3000 Project, but it’s been full of some real surprises.

When we first spec’d it out, we anticipated that the final product would be less than a foot in diameter. However, the Levantar 3000 has turned out to be a full 36.4 inches around.”



Semantic Representation of Message

ISA (Sally_Tobias, *junior_employee*)

LOC (Jerry_Haviland, *outside_network*)

ISA (Levantar_3000, *Acme_product*)

ASSOC (Levantar_3000, *diameter*)

MEAS (diameter, *36.4_inches*)



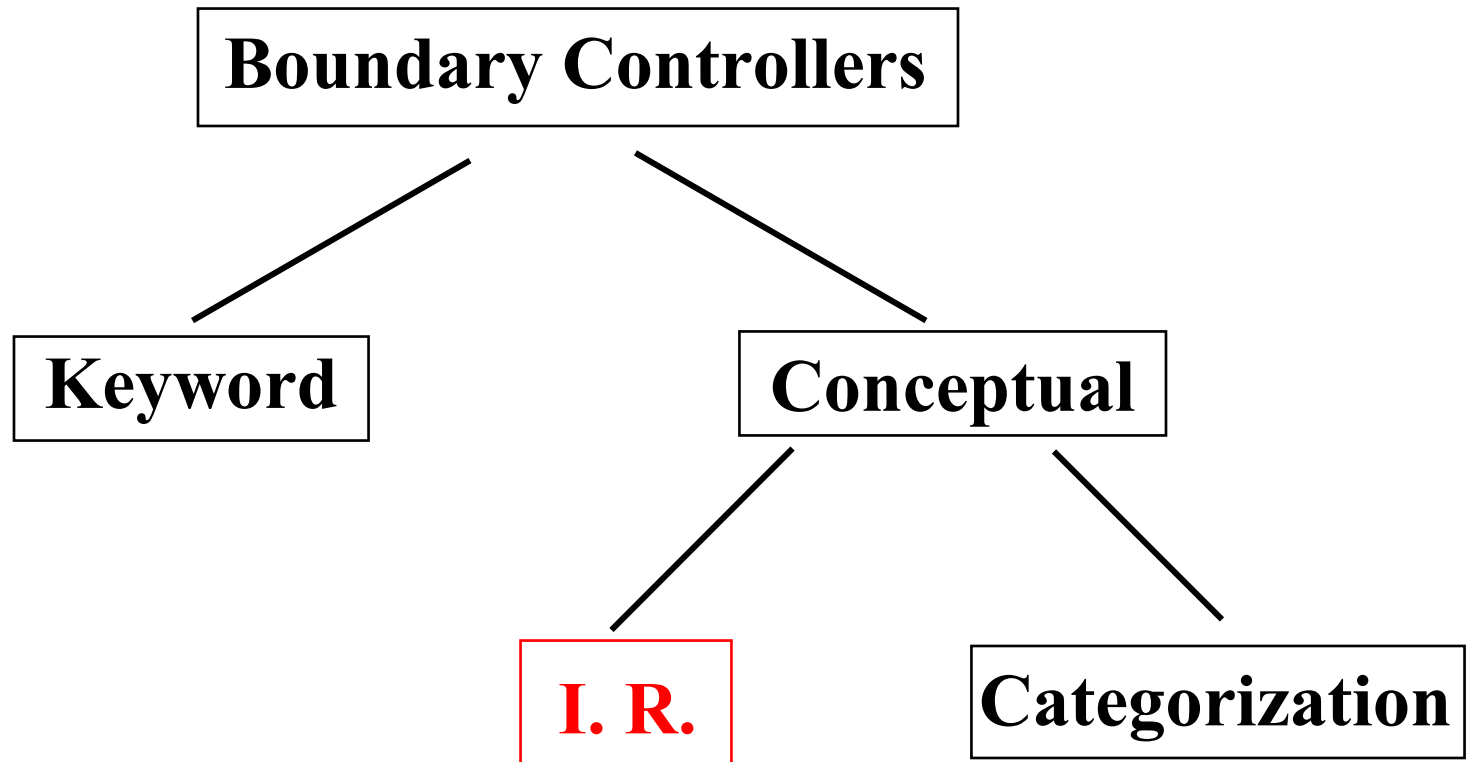
Outcome of Comparison

The semantic model of the message satisfies the conditions of a Business Rule.



The message is barred from release.

Possible Approaches

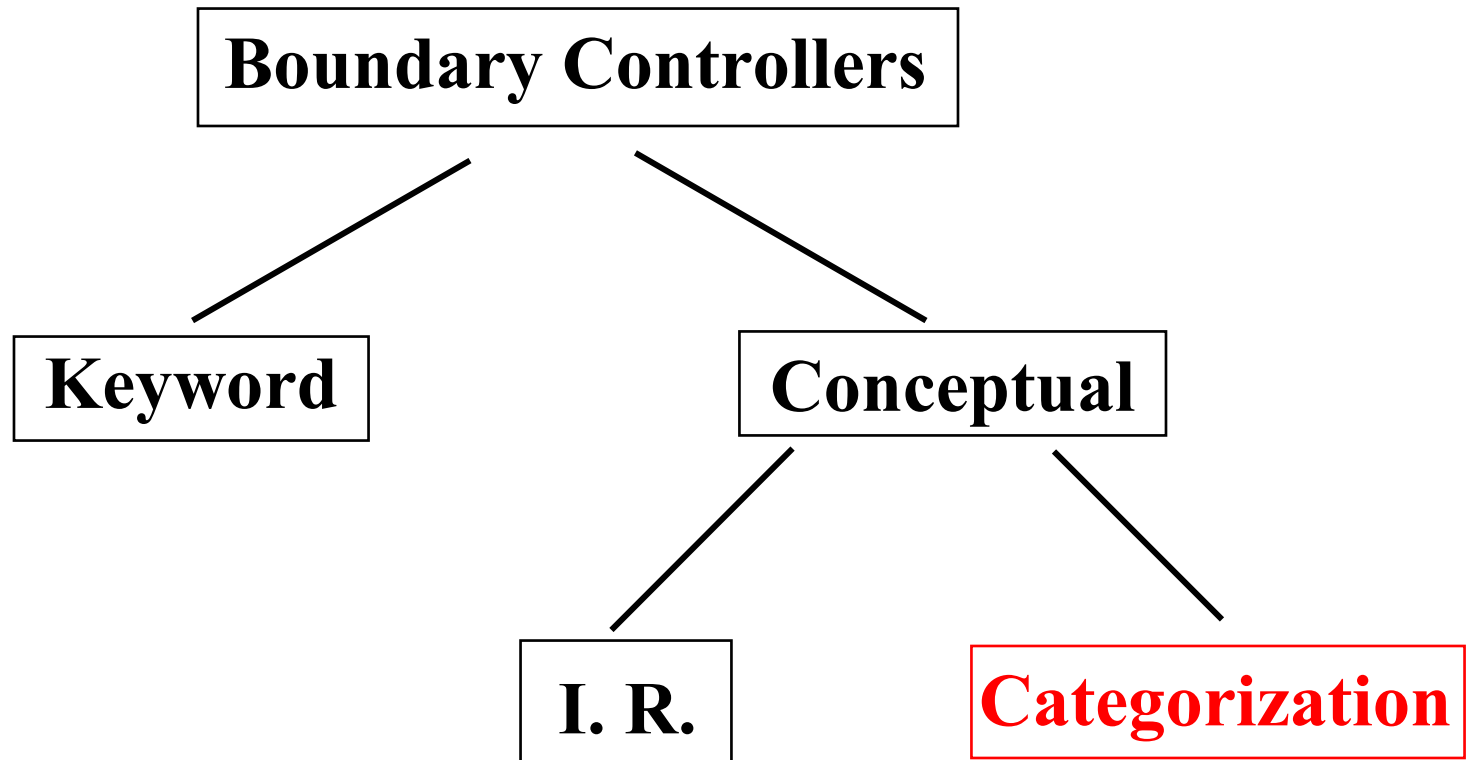




Information Retrieval Approach

- **Takes an ordinary textual description of an organization's Business Rules**
- **System constructs a semantic releasability model of each rule or set of rules**
- **Compares the meaning content of outgoing messages against these models**
 - If they are sufficiently similar & if the releasability level of the matching model exceeds that of the sender or the recipient, the document is barred from release

Possible Approaches





Categorization Approach

- **Manually categorize a training set of outgoing messages as to which Business Rule applies, and whether they breach the rule**
- **Run training documents through NLP modules for rich feature marking**
- **Machine Learning Classifier learns a vector of distinguishing terms, phrases, concepts, entities, and relationships**
- **Vector is used to categorize new messages**



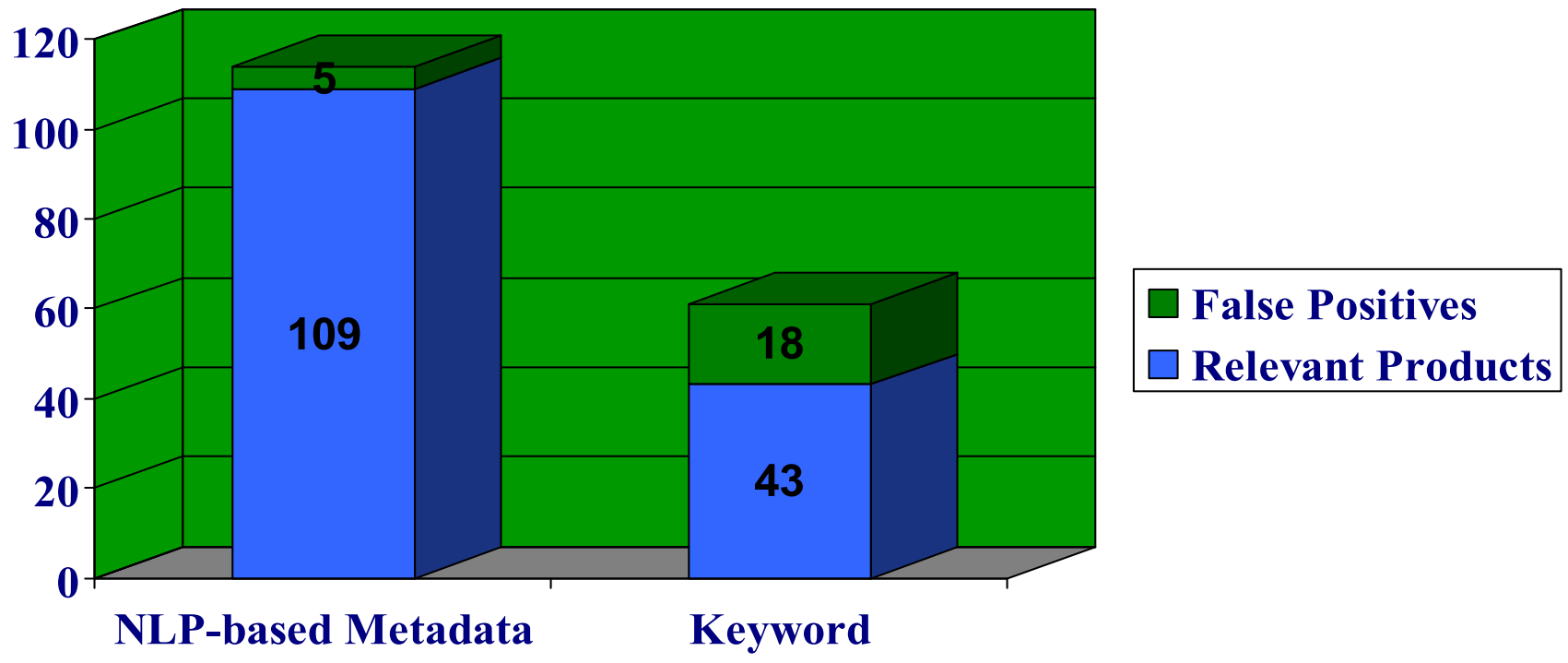
Government Test

- **DARPA-funded project**
- **Goal - to empirically compare standard Boundary Control approach to NLP-based conceptual approaches**
- **Used solution-united.com's DataShield**
 - Utilizing NLP feature extraction
 - Probabilistic text classifier & uncertainty sampling algorithm (Lewis & Gale, '94)
- **Ran against a commercial keyword system**



Government Test Results

	<u>Recall</u>	<u>Precision</u>
'Keyword'	38%	70%
DataShield	99%	96%





Currently

- **Many make a trade-off between:**
 - effective, inefficient boundary control by hiring and training a security monitor
 - OR -
 - efficient, but ineffective boundary control by relying on a ‘keyword’ system



Future

- **Conceptual, NLP-based technologies hold promise for effective & efficient Boundary Control**
 - Internally
 - Supports corporate “need to know” policies among departments & individuals
 - Externally
 - Prevents advertent & inadvertent security slip-ups