OIC - 2008

Automatic Ontology Creation from Text for National Intelligence Priorities Framework (NIPF)

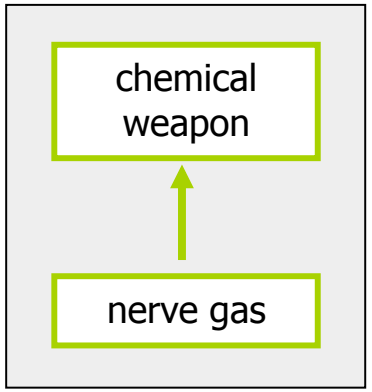
Mithun Balakrishna & Munirathnam Srikanth
Lymba Corporation
www.lymba.com
Richardson, Texas

Jaguar/KAT Overview

- Jaguar: Automatically builds Ontologies and Knowledge-Bases from the concepts and relationships between those concepts found in text.
- Constituents of a knowledge base
  - Concepts/Vocabulary ("weapon", "WMD", "launcher")
  - Relations ("anthrax" ISA "biological weapon", "anthrax" CAU "death")
  - Organization of Relations
    - Hierarchical
    - Contextual

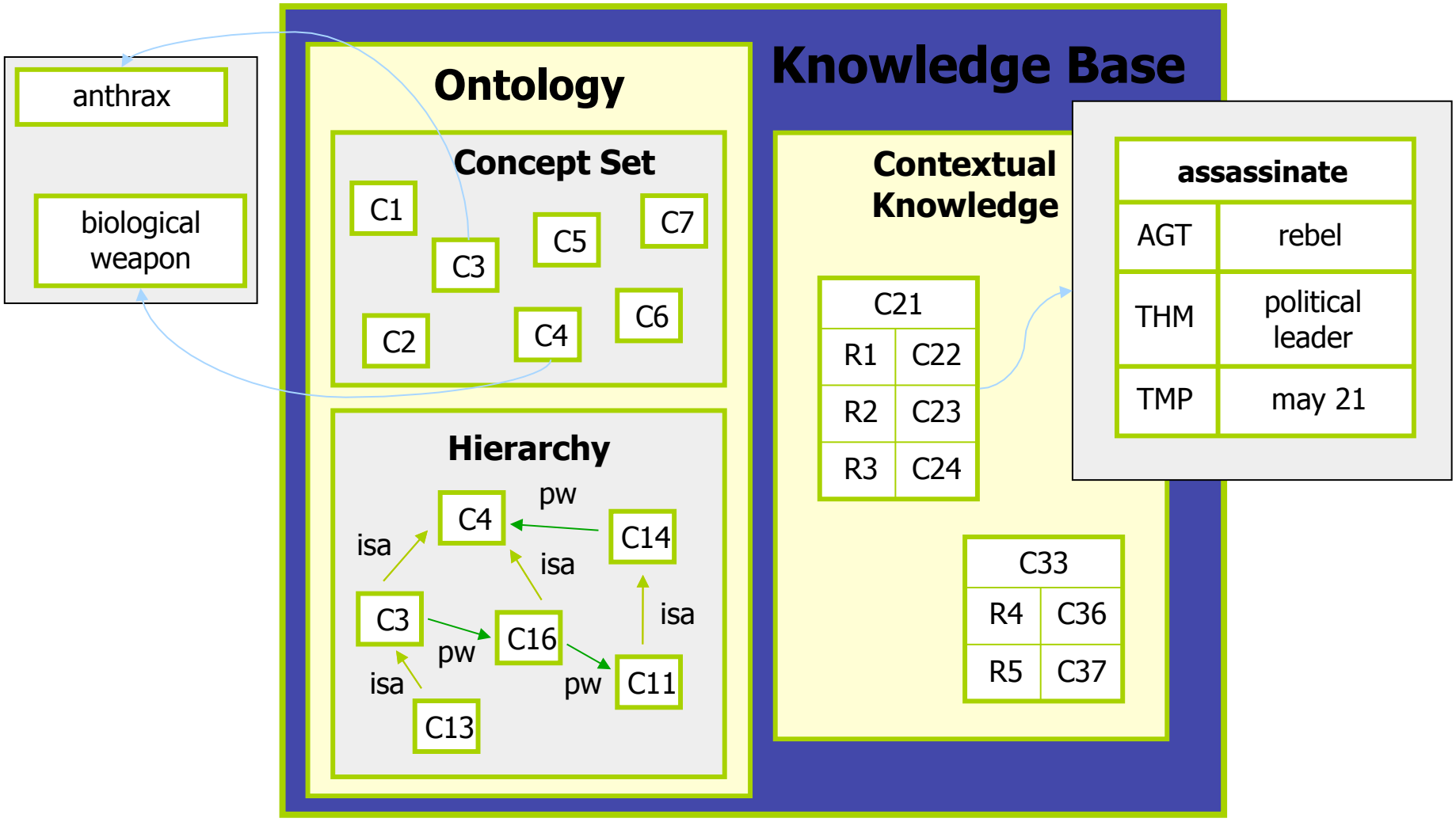
Types of Knowledge

- Universal (or ontological)
  - Represented in Hierarchies
  - Simple binary relations between concepts
  - *"Chemical weapons such as nerve gas, …"*
- Contextual
  - Represented in individual (semantic) contexts
  - Groups of relations centered on a common concept
  - *"The forces launched a full-scale attack on Monday"*

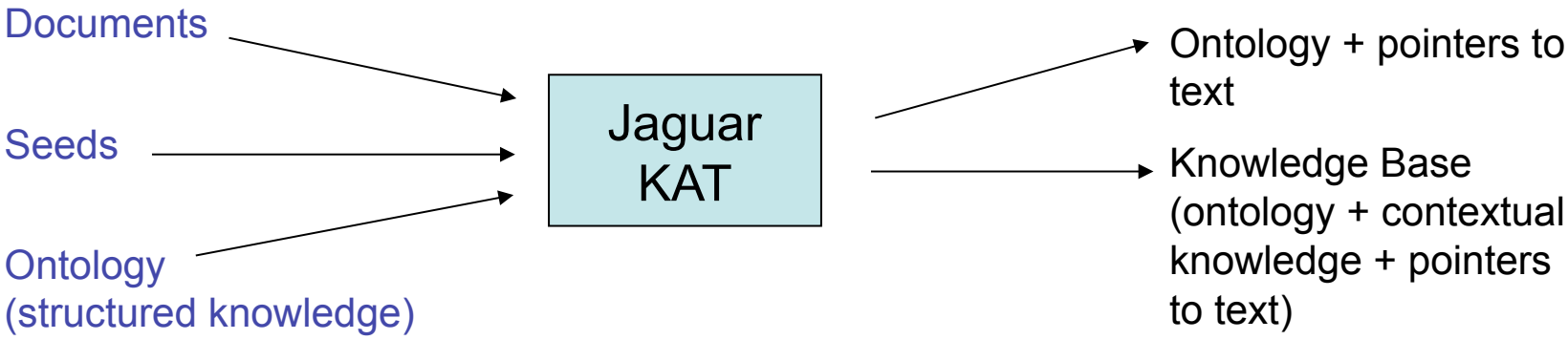


launch	
AGT	forces
THM	full-scale attack
TMP	monday

KB Constituents



Jaguar/KAT Overview



Functionality

1. Produce ontologies
2. Link concepts & relations to text
3. Visualize ontology
4. Edit ontology
5. Enhance an existing ontology
6. Merge two ontologies into a consistent ontology
7. Ontological search of documents (search documents using ontology)

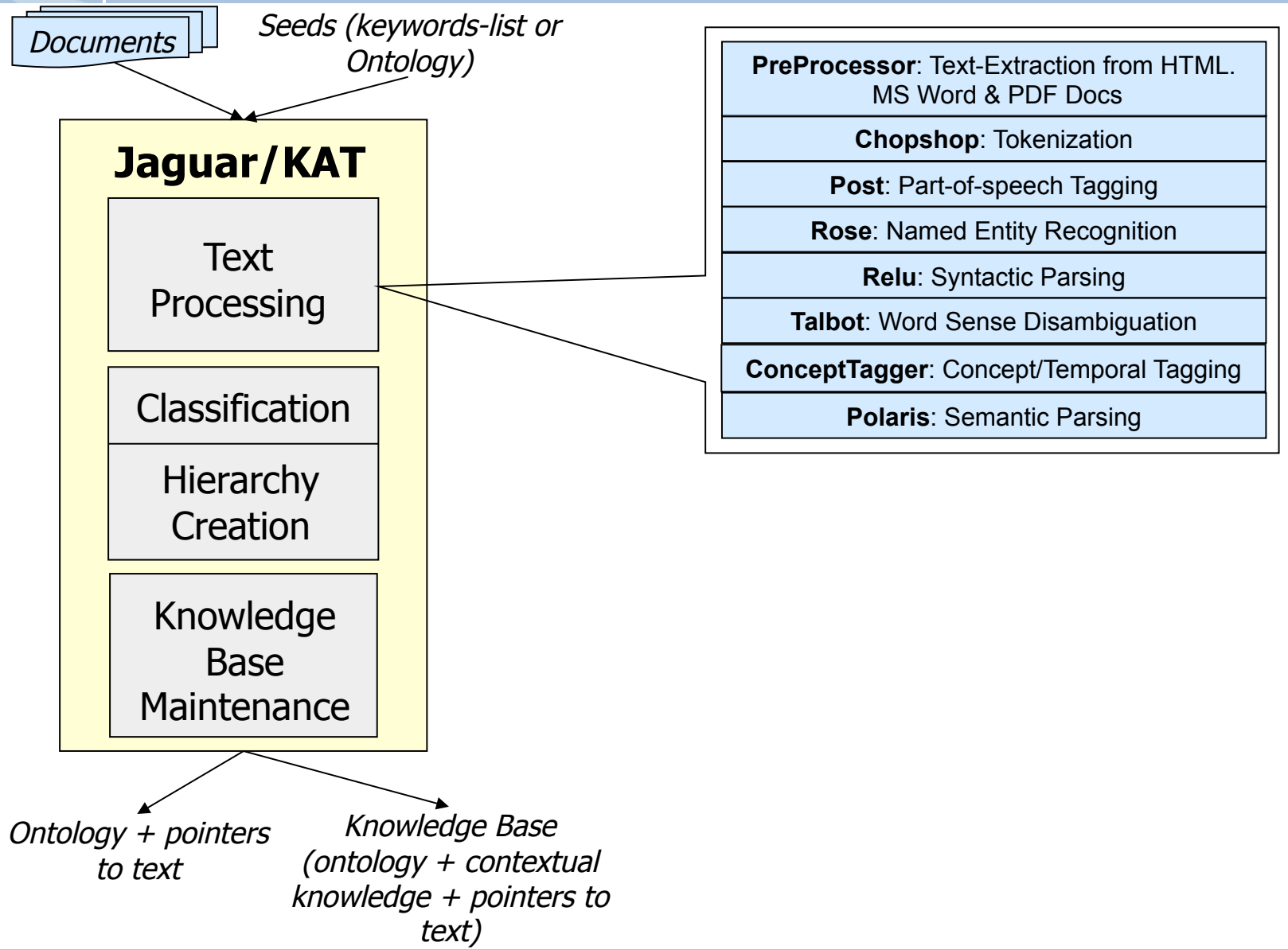
Knowledge Bases

- Ontology/KB creation overview
  - Knowledge *Extraction* from Text
    - Pattern recognition; Semantic Parsing
  - Knowledge *Representation* and *Storage*
    - Contextual vs. Universal
    - XML; Relational Database
  - Knowledge Base *Maintenance*
    - Conflict Resolution; Ontology Merging
    - User Interaction; Ontology Modification

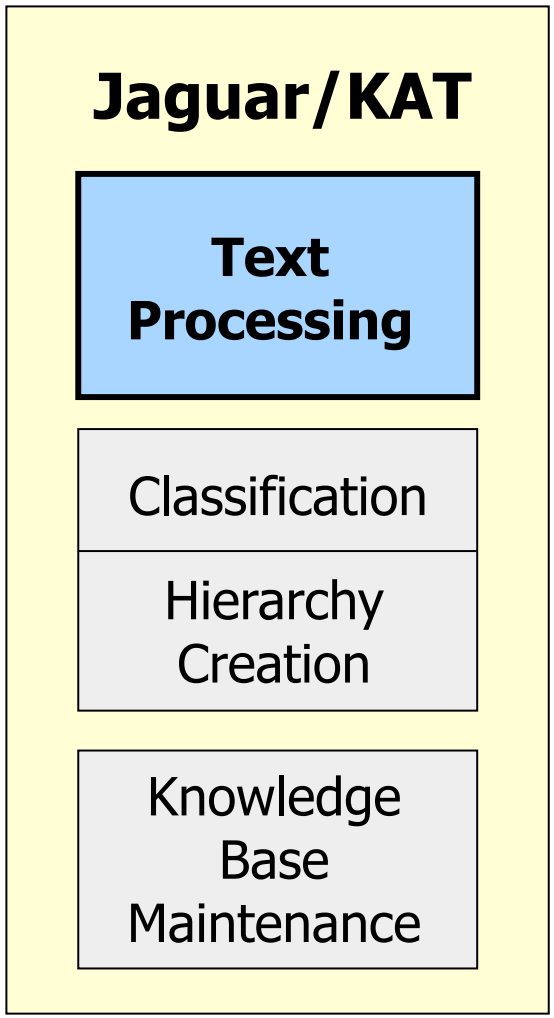
NIPF Ontology Creation

- NIPF is the *Director of National Intelligence's (DNI's) guidance to the Intelligence Community on the national intelligence priorities approved by the President of the United States of America*
- We use Jaguar to create an ontology library for the 33 topics defined in NIPF
  - For each NIPF topic, we collected 500 documents from the web (the Weapons topic was an exception and its collection had only 50 Wikipedia documents) and manually verified their relevance to the corresponding topic.
  - For each NIPF topic, Jaguar is provided with an initial seed set containing on average 51 concepts of interest

Jaguar – NIPF – Process & Modules



NIPF Ontology Creation



Text Processing

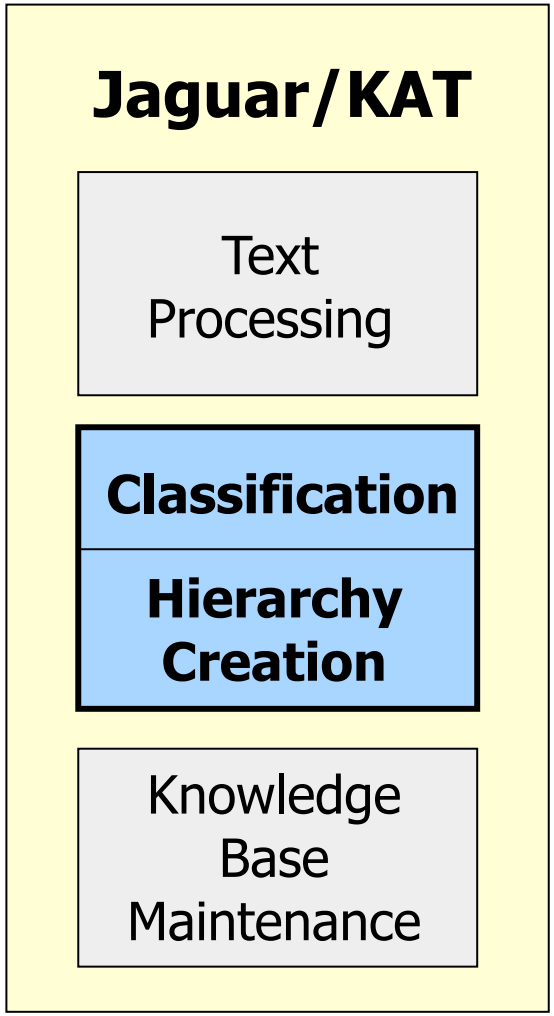
Input: Documents, Seeds

- Extract "concepts" of interest
- Extract binary relations (universal)
- Use Semantic Parser to obtain contextual knowledge

Output: Concepts, Contexts, Binary Relations

"The rebels had access to *chemical weapons*, such as *nerve gas* and other *poisonous gases*."

NIPF Ontology Creation



Classification/Hierarchy Creation

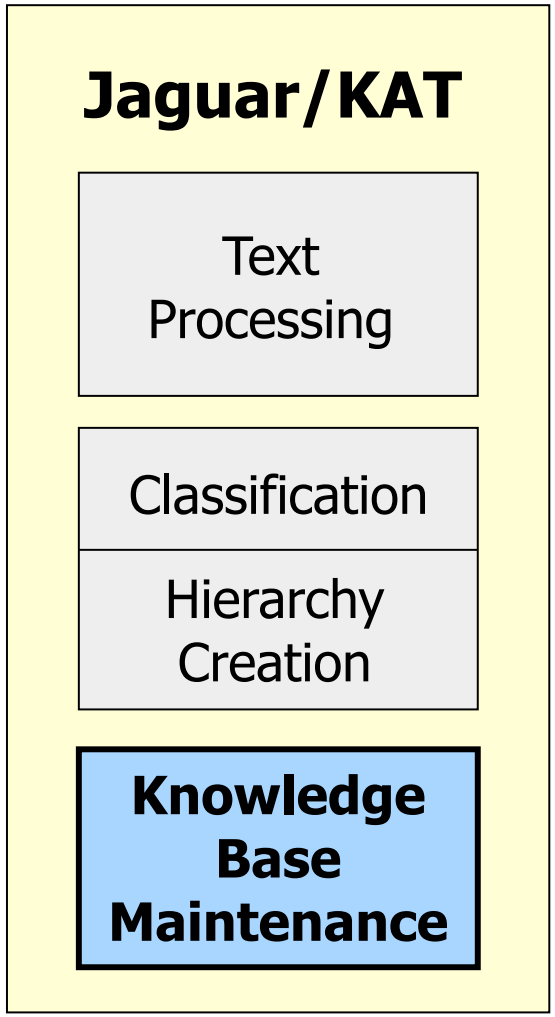
Input: Concepts, Binary Relations

- Classify each concept against every other using defined procedures, obtaining set of ISA relations
- Add all ISA and other binary relations to the hierarchy using *conflict resolution*

Output: Hierarchy of relations

"Scud missile" *ISA* "missile"
"Iraqi standing_army" *ISA* "Asian army"
"weapons inspection team" *ISA* "inspection team"

NIPF Ontology Creation



Knowledge Base Maintenance

- Knowledge Base Merging
- Visualization
- Knowledge Base Editing
  - User Interaction
  - Modifications

NIPF Ontology/KB Creation - Example

Iteration 0

Sentences

1. The rebels had access to chemical weapons, such as nerve gas and other poisonous gases
2. The nerve gas was created using numerous fluorine-based compounds

Seed Concepts

1. Weapon

Iteration 1

Sentences being used

1. The rebels had access to chemical *weapons*, such as nerve gas and other poisonous gases

Concepts/SRs Extracted

Chemical Weapon —*ISA*→ Weapon

Seed Concepts Used/Added

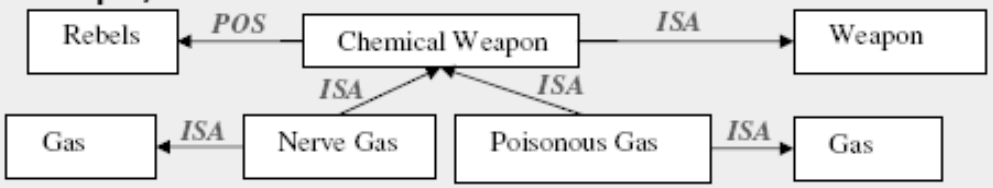
1. *Weapon*
2. Chemical Weapon

Iteration 2

Sentences being used

1. The rebels had access to *chemical weapons*, such as nerve gas and other poisonous gases

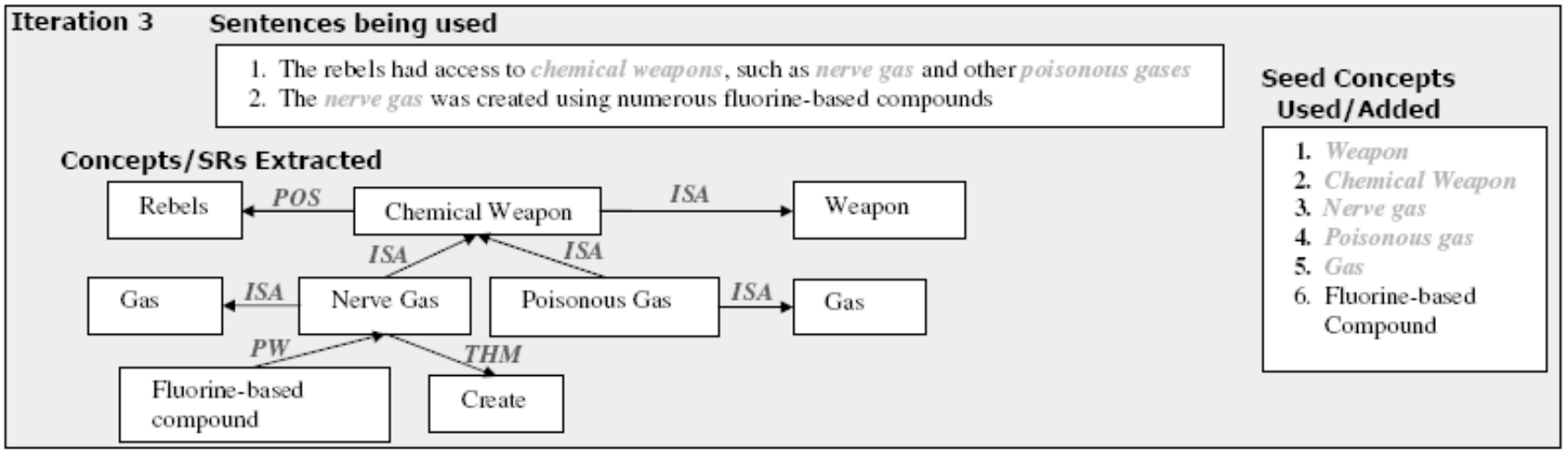
Concepts/SRs Extracted



Seed Concepts Used/Added

1. *Weapon*
2. *Chemical Weapon*
3. Nerve gas
4. Poisonous gas
5. Gas

NIPF Ontology/KB Creation - Example

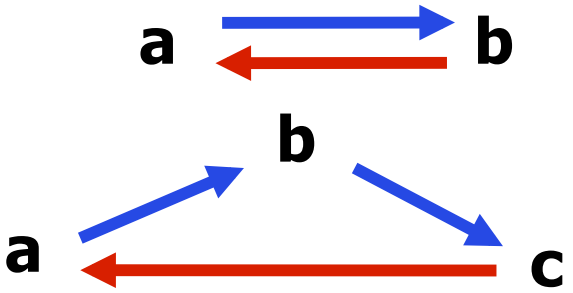


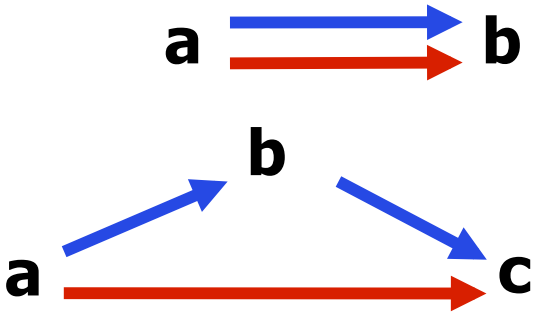
Conflict Resolution

- Properties of Hierarchical Relations
  - Transitive
    - ‘a ® b’ AND ‘b ® c’ implies ‘a ® c’
  - Strictly non-symmetric
    - ‘a ® b’ implies NOT ‘b ® a’
- Example: IS-A relation
  - ‘cat ISA mammal’ AND ‘mammal ISA animal’ implies ‘cat ISA animal’
  - ‘cat ISA mammal’ implies NOT ‘mammal ISA cat’

Types of Conflicts

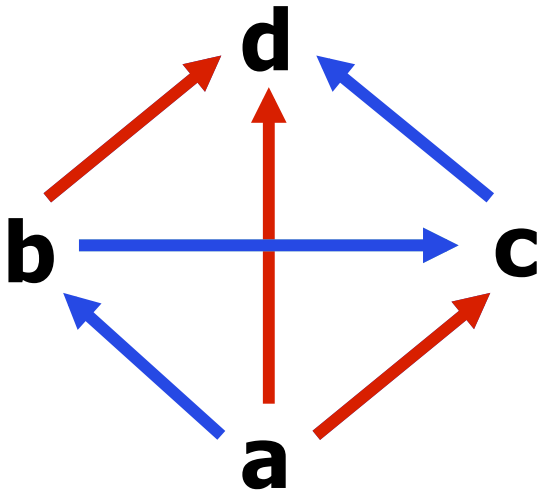
- Anomalies
  - Simple loops
  - Cycles
- Redundancies
  - Duplicate Relations
  - Jump Links





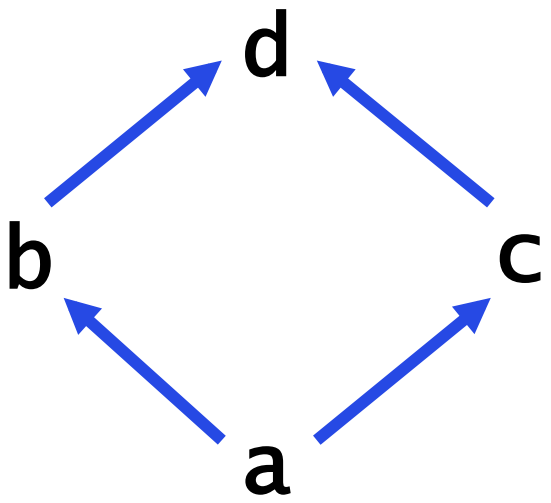
More on Jump Links

- Multiple paths from one node to another are acceptable
  - As long as no ‘single link’ duplicates a path



Does fewer links mean less knowledge?

Links = 4

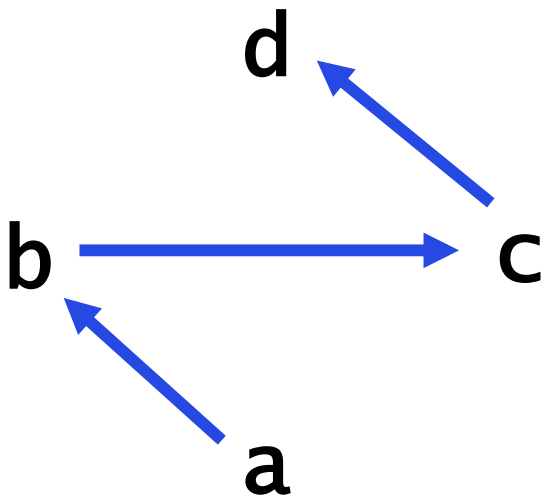


Assertions

1. a → b
2. a → c
3. b → d
4. c → d
5. a → d

Does fewer links mean less knowledge?

Links = 3



Assertions

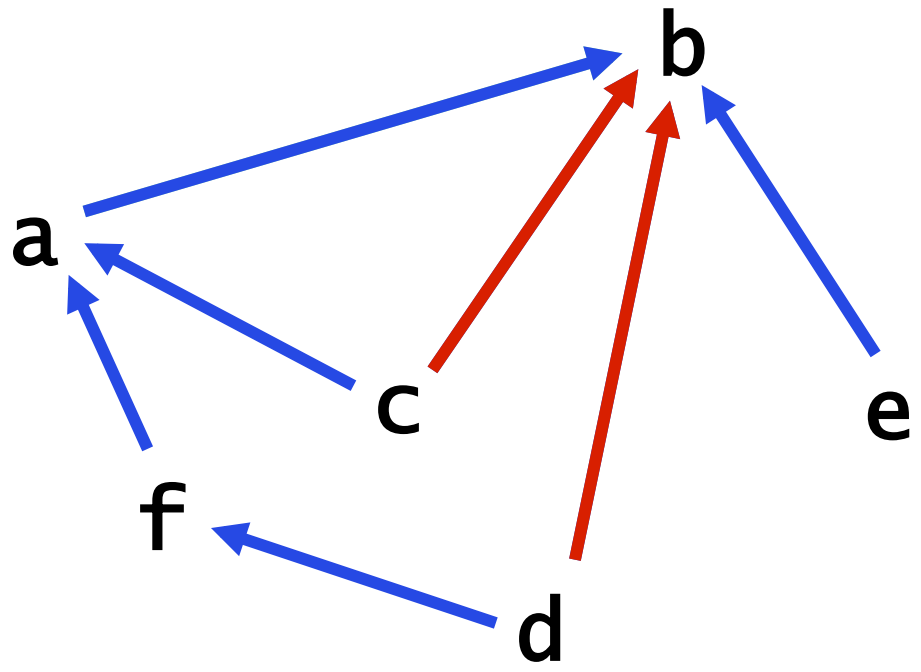
1. a → b
2. a → c
3. b → d
4. c → d
5. a → d
6. b → c

Conflict Resolution Algorithm

- Approach Used: *Prevention*
  - Start from an *empty hierarchy* and an input relation set
  - Add a relation from the input set to the hierarchy, if:
    - It does *not form a cycle*
    - It is *not redundant* (does not duplicate a path)
  - After the addition of any relation, algorithms *(jump link removal)* are run to ensure that all jump links are removed

Jump Link Removal

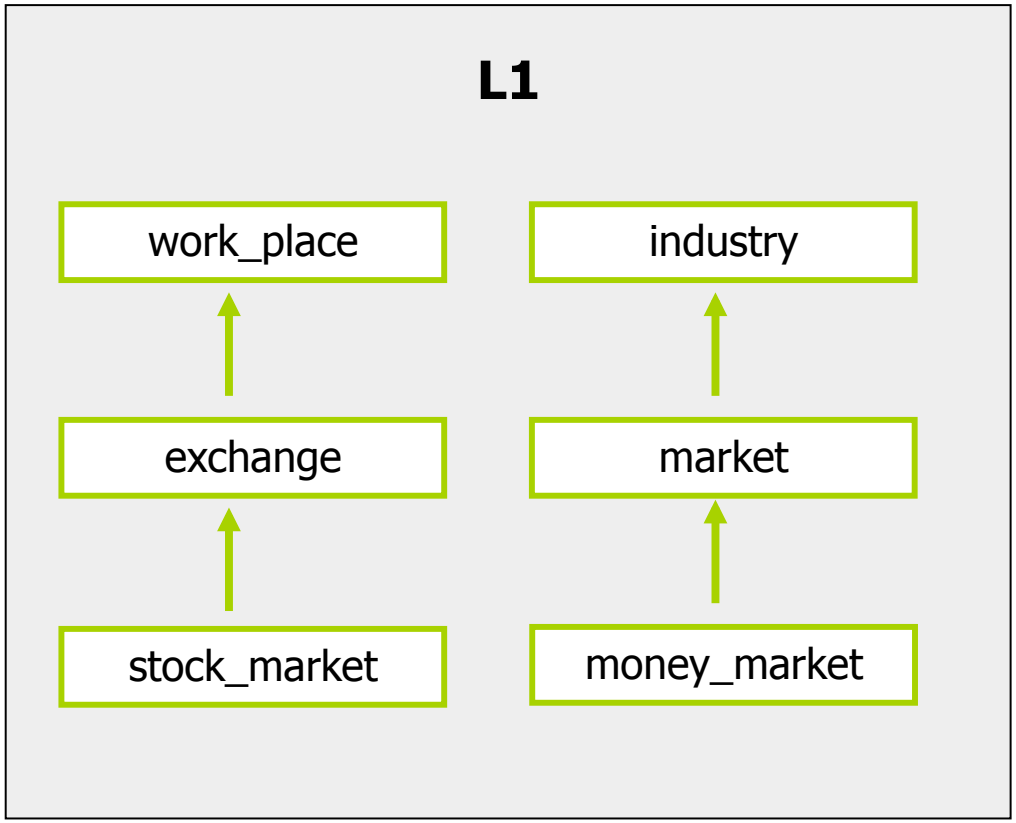
- When it is safe to add ‘a ® b’, remove links from direct descendents of ‘b’ to ‘b’, if they have a path to ‘a’

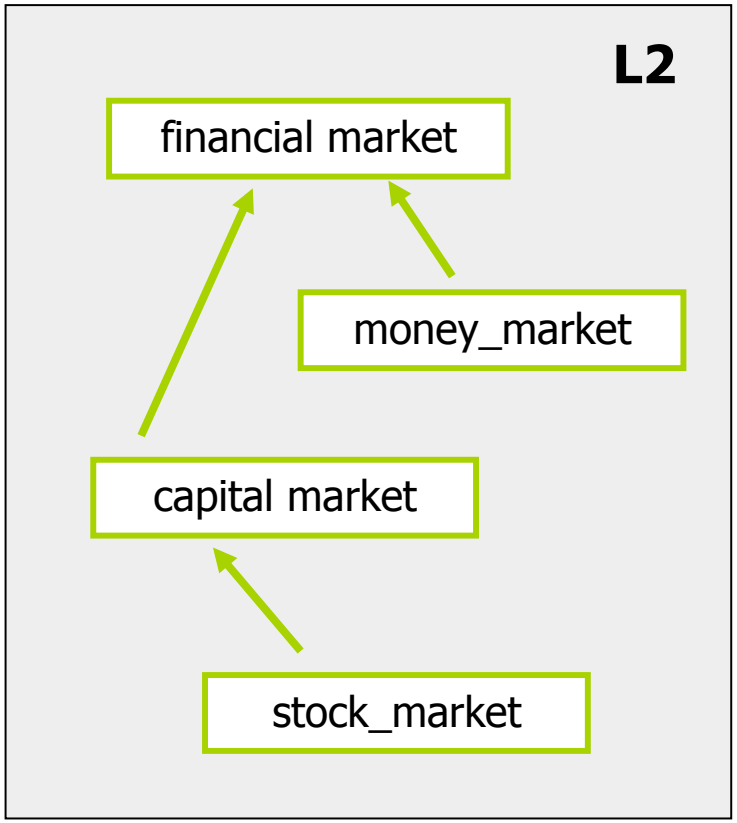


Knowledge Base Merging

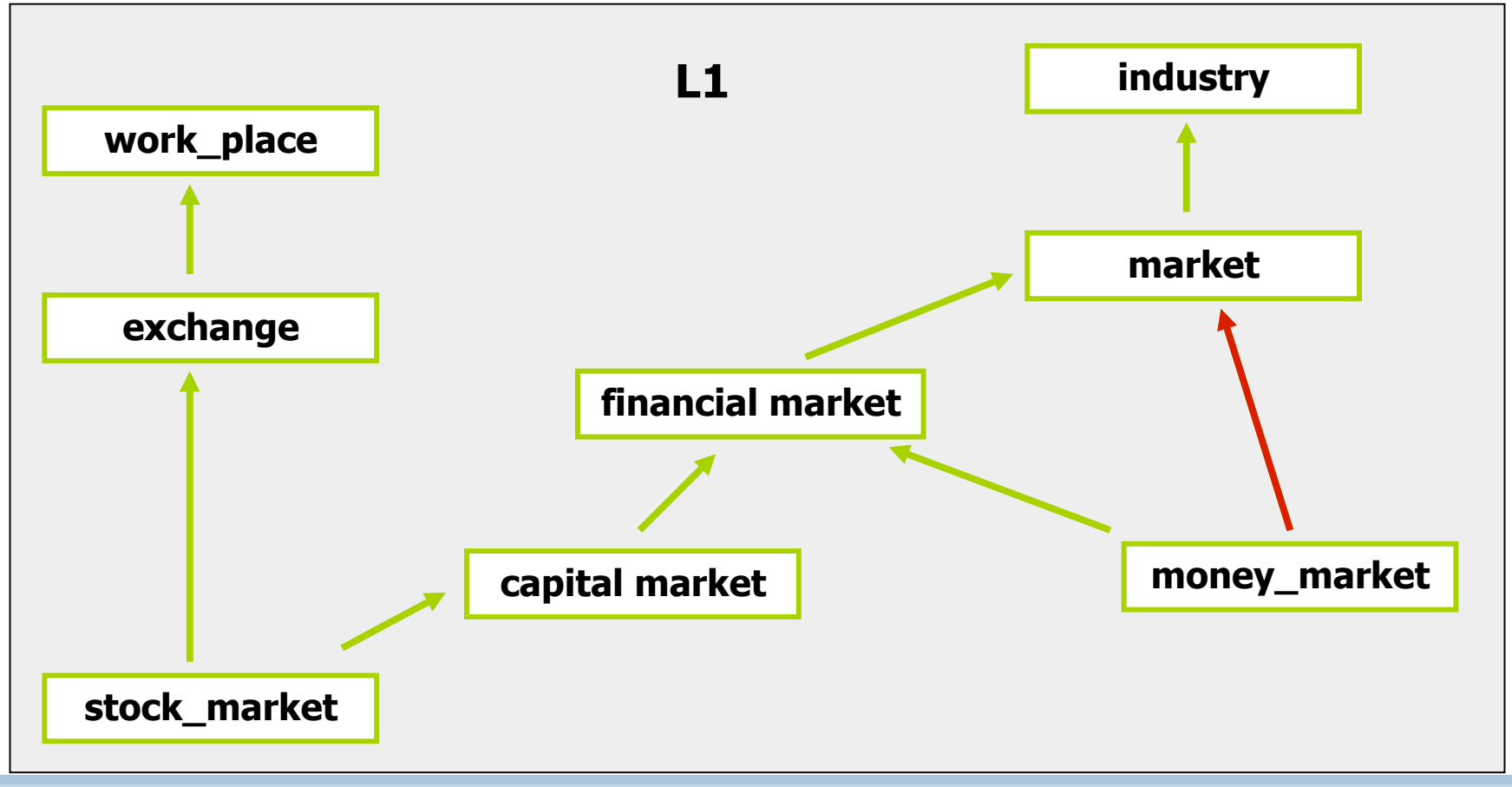
- Current Approach
  - Label the bigger ontology L1, and the other L2
  - Merge *concepts* (from those in L2 into those of L1)
  - Copy all *contexts* (from L2 to L1)
  - Add all *relations* (from the hierarchy of L2 to the hierarchy of L1) using the conflict resolution algorithm
  - Additionally, *classify* all concepts in L1's hierarchy against concepts in L2's hierarchy (form relation set R)
  - *Add relations* from R into L1's hierarchy (conflict resolution)

Merging Hierarchies





Merging Hierarchies



NIPF Ontology/KB Evaluation

- We evaluated the quality of 4 Jaguar NIPF ontologies by comparing them against manual gold annotations
- Our evaluations are focused on the
  - *Lexical Level*
  - *Vocabulary, or Data Layer Level*
  - *Other Semantic Relations Level*
- Viewing an ontology as a set of semantic relations between two concepts, the human annotators:
  - Labeled an entry *correct* if the concepts and the semantic relation are correctly detected by the system else marked the entry as *Incorrect*
  - Labeled a *correct* entry as *irrelevant* if any of the concepts or the semantic relation are irrelevant to the domain
  - From the sentences *added new entries* if the concepts and the semantic relation were omitted by Jaguar

NIPF Ontology/KB Evaluation - Metrics

$$Pr(Correctness) = \frac{N_j(correct) + N_j(irrelevant)}{N_j(correct) + N_j(incorrect) + N_j(irrelevant)}$$

$$Pr\begin{pmatrix} Correctness \\ + \\ Relevance \end{pmatrix} = \frac{N_j(correct)}{N_j(correct) + N_j(incorrect) + N_j(irrelevant)}$$

$$Cvg(Correctness) = \frac{N_j(correct) + N_j(irrelevant)}{N_g(correct) + N_g(irrelevant) + N_g(added)}$$

$$Cvg\begin{pmatrix} Correctness \\ + \\ Relevance \end{pmatrix} = \frac{N_j(correct)}{N_g(correct) + N_g(added)}$$

$N_j(.)$ gives the counts from Jaguar's output

$N_g(.)$ correspond to counts in the user annotations

NIPF Ontology/KB Evaluation - Results

Semantic Relation	Definition	Example	Code
ISA	X is a (kind of) Y	[XY] [John] is a [person]	ISA
Part-Whole/Meronymy	X is a part of Y	[XY] [The engine] is the most important part of [the car] [XY] [steel][cage] [YX] [faculty] [professor] [XY] [door] of the [car]	PW
Cause	X causes Y	[XY] [Drinking] causes [accidents]	CAU

NIPF Topic	Unique Semantic Relations					Unique Concepts		
	ISA	PW	CAU	Others	**Total**	In ISA/PW/CAU	Others	**Total**
Weapons	1683	766	113	946	3508	2620	1012	3473
Missiles	2939	2296	646	2692	8573	5982	3539	7873
Illicit Drugs	2356	2040	817	5464	10677	5107	4982	7935
Terrorism	2590	4219	1497	5405	13711	7929	6247	11638

NIPF Ontology/KB Evaluation - Results

Number of Annotators	NIPF Topic	Precision		Coverage		F-Measure	
		Correctness	Correctness+ Relevance	Correctness	Correctness+ Relevance	Correctness	Correctness+ Relevance
3	Weapons	**0.610090**	**0.501499**	0.702424	0.657122	**0.653009**	0.568859
1	Missiles	0.533867	0.485364	0.793775	**0.777747**	0.63838	**0.597715**
2	Illicit Drugs	0.471938	0.274506	0.801422	0.701122	0.594053	0.39454
1	Terrorism	0.388788	0.291019	**0.822285**	0.776206	0.527953	0.423323

Conclusions

- We presented a semi-automatic development of an ontology library for the NIPF topics
- We used Jaguar-KAT, a state-of-the-art tool for knowledge acquisition and domain understanding, with minimized manual intervention to create NIPF ontologies loaded with rich semantic content
- We also defined evaluation metrics to assess the quality of the NIPF ontologies created using our methodology
- The results look very promising and show that a decent amount of knowledge was automatically and accurately extracted by Jaguar from the input document collection while keeping the manual intervention in the process to a minimum