

# **Combining Intelligent Agents with an NLP-based Search Engine**

**Elizabeth D. Liddy, Ph.D.**

**Center for Natural Language Processing  
School of Information Studies  
Syracuse University**

**January 12, 2001**



# Mission Statement

---

To advance the development of *human-like language understanding software capabilities* for government, commercial, and consumer applications.



# Mission Statement

---

To advance the development of *human-like language understanding software capabilities* for government, commercial, and consumer applications.

- **Basic and applied research**



# Mission Statement

---

To advance the development of *human-like language understanding software capabilities* for government, commercial, and consumer applications.

- Basic and applied research
- **Building on our recognized capabilities in Natural Language Processing**



# Topics:

---

- 1. EVA – an intelligent information agent system**



# Topics:

---

1. **EVA – an intelligent information agent system**
2. **eQuery – an NLP-based 2-stage information retrieval system**



# Topics:

---

1. **EVA – an intelligent information agent system**
2. **eQuery – an NLP-based 2-stage information retrieval system**
3. **Future – EVA + eQuery**



# Intelligent Agents

---

- **Autonomous** - can act/react on behalf of the user to achieve a goal.
- **Adaptive** - can learn from examples, experiences, other agents, etc, and evolve and co-evolve with the changing environment.
- **Cooperative** - can communicate and collaborate with other agents to achieve a common goal.



# Goals of EVA Project:

---

- **To develop a prototype system for NIMA and its national security customers which would:**
  - provide timely, accurate, and complete information in response to analysts' needs



## Goals of EVA Project (cont'd):

---

- **To develop a prototype system for NIMA and its national security customers which would:**
  - provide timely, accurate, and complete information in response to analysts' needs
  - continuously bring new, relevant information to analysts' attention without requiring them to re-search



## Goals of EVA Project (cont'd):

---

- **To develop a prototype system for NIMA and its national security customers which would:**
  - provide timely, accurate, and complete information in response to analysts' needs
  - continuously bring new, relevant information to analysts' attention without requiring them to re-search
  - adapting to individual users' styles, interests, techniques, and preferences



# EVA Capabilities

---

- **Searches** for geospatial information sources on the **Web**



# EVA Capabilities

---

- **Searches** for geospatial information sources on the Web
- **Retrieves** documents & images using captions and the associated textual information



# EVA Capabilities

---

- **Searches** for geospatial information sources on the Web
- **Retrieves** documents & images using captions and the associated textual information
- **Filters** information based on dynamic user profiles



# EVA Capabilities

---

- **Searches** for geospatial information sources on the Web
- **Retrieves** documents & images using captions and the associated textual information
- **Filters** information based on dynamic user profiles
- **Learns** and **evolves** by applying a Neural Genetic Algorithm (NGA) to Web agents



# EVA Capabilities

---

- **Searches** for geospatial information sources on the Web
- **Retrieves** documents & images using captions and the associated textual information
- **Filters** information based on dynamic user profiles
- **Learns** and **evolves** by applying a Neural Genetic Algorithm (NGA) to Web agents
- **Combines** both web-crawling and meta-searching agents



# EVA Capabilities

---

- **Searches** for geospatial information sources on the Web
- **Retrieves** documents & images using captions and the associated textual information
- **Filters** information based on dynamic user profiles
- **Learns** and **evolves** by applying a Neural Genetic Algorithm (NGA) to Web agents
- **Combines** both web-crawling and meta-searching agents
- **Coordinates** multiple agents working on various aspects of the task in parallel

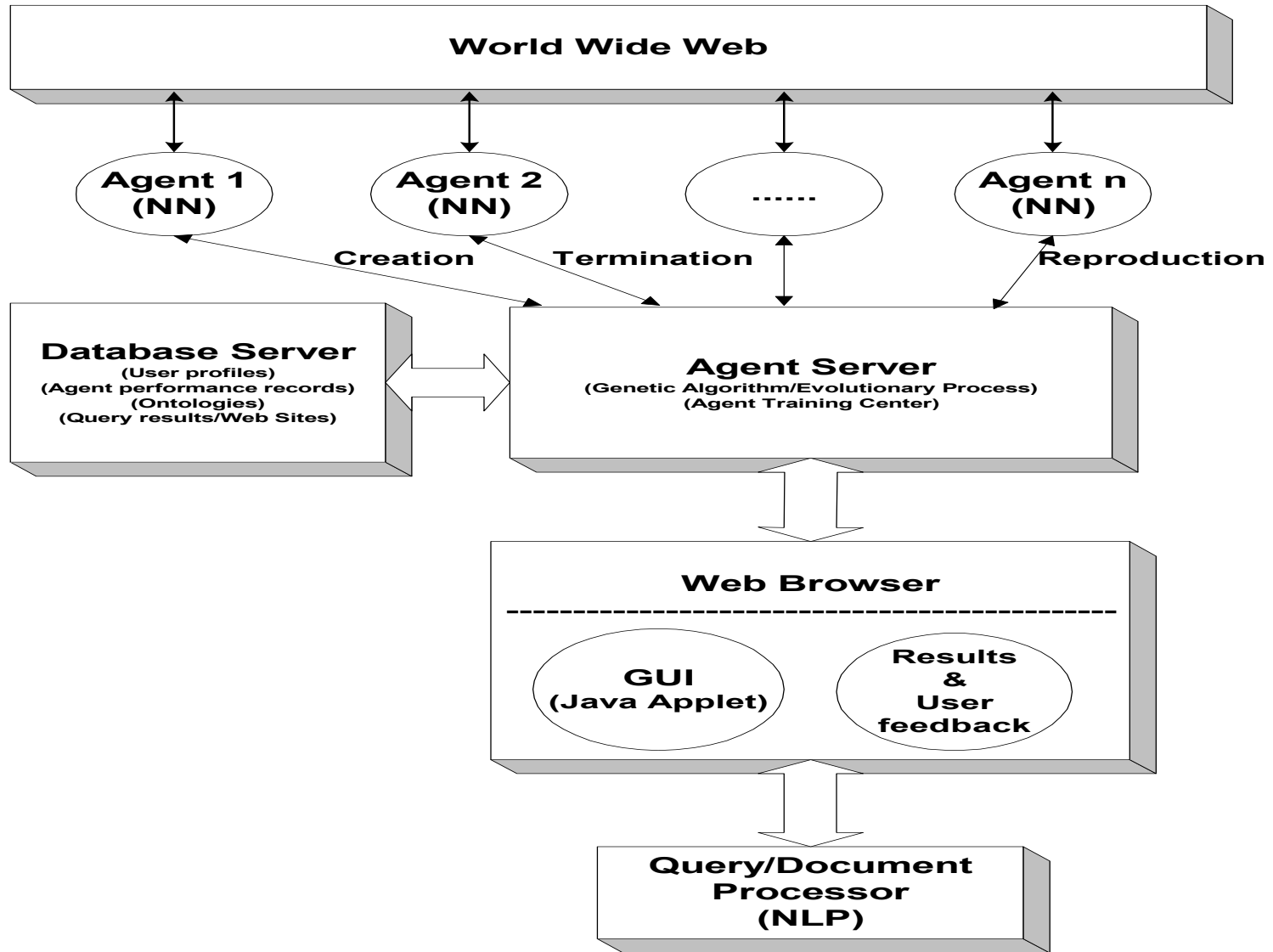


# EVA Capabilities

---

- **Searches** for geospatial information sources on the Web
- **Retrieves** documents & images using captions and the associated textual information
- **Filters** information based on dynamic user profiles
- **Learns** and **evolves** by applying a Neural Genetic Algorithm (NGA) to Web agents
- **Combines** both web-crawling and meta-searching agents
- **Coordinates** multiple agents working on various aspects of the task in parallel
- **Performs** automatic relevance feedback to improve retrieval results

**Figure 1. EVA System Architecture**





# Image Retrieval Using Text

**The <title> field**

**The <meta> field**

**The <img> field**

**The image caption - text following the image's URL until end of the paragraph, or until a link to another image is encountered**

**Relative importance of the terms in these fields is determined by the Neuro-Genetic Algorithm**



# Sample Queries

- Accelerometers
- Active tracking systems
- Altazimuth instrument
- Achromatic lens
- Cadastral maps
- Contact printing frame
- Continuous strip cameras
- Densitometers
- Fiber optics
- Gravity anomaly map
- Gyrocompass

...



# Query Expansion

We differentiate 2 types of terms:

- **Domain-specific terms (from MC&G Handbook; Getty Thesaurus of Geographic Names, etc)**
- **Non-domain specific terms (from generic sources such as WordNet)**
- **Both used for initial expansion:**
  - cadastral map → property map
  - accelerometer → direction, speed, altitude



# Neural network

- **Query terms are used to initialize the 3-layer feed-forward neural network**
- **User relevance feedback provides additional training patterns**
- **Automatic relevance feedback can provide training patterns in the absence of the user**



# Automatic Relevance Feedback

- **Top 10 retrieved pages are assumed relevant**
- **From these,  $X$  terms are selected from the window surrounding matched words, and added to the query for the next round of search**
- **Repeat  $N$  times. Our research shows that retrieval results stabilize after 5 or 6 rounds**
- **Use the final top 10 retrieved pages that occur at least  $N/2$  times in all  $N$  runs to expand the original query**



# The Neuro-Genetic Algorithm

- **Mimics the natural selection process to control and improve the overall performance or ‘fitness’ of the agent population**
  - **Facilitates agent evolution**
  - **Learns from examples**
  - **Evolves the best/most parsimonious feature set**
- **Operates at an inter-agent perspective**
  - **Extends individual NN agents’ local perspective**



# The Neuro-Genetic Algorithm

**NGA has the following steps:**

- 1. Define parameters**
- 2. Define fitness**
- 3. Create population**
- 4. Evaluate fitness**
- 5. Select mate**
- 6. Reproduce**
- 7. Mutate**
- 8. Test convergence. (*If no, go back to 4. If yes, stop.*)**

# Evaluation

	Precision at 10	Precision at 20
<b>NN-Agent Retrieval</b>	<b>.753</b>	<b>.656</b>

# Evaluation

	Precision at 10	Precision at 20
<b>NN-Agent Retrieval</b>	<b>.753</b>	<b>.656</b>
<b>NGA-Agent Retrieval</b>	<b>.830</b>	<b>.748</b>
<b>Improvement</b>	<b>9.7 %</b>	<b>8.7%</b>



# Web Page Classification

- **Personalized - based on users' preferences**



# Web Page Classification

- **Personalized - based on users' preferences**
- **Types of text categorization techniques:**



# Web Page Classification

- **Personalized - based on users' preferences**
- **Types of text categorization techniques:**
  - Naïve word matching – shared words between text and category names/descriptions



# Web Page Classification

- **Personalized - based on users' preferences**
- **Types of text categorization techniques:**
  - Naïve word matching – shared words between text and category names/descriptions
  - **Thesaurus-based matching – relies on ontologies for expanding set of terms**



# Web Page Classification

- **Personalized - based on users' preferences**
- **Types of text categorization techniques:**
  - Naïve word matching – shared words between text and category names/descriptions
  - Thesaurus-based matching – relies on ontologies for expanding set of terms
  - Empirical learning of term-category associations – utilizes relevance feedback to learn contextualized semantic associations



# Neuro-Genetic Classification Results

- **Incorporates ensemble averaging**
  - Linear combination of the outputs from 10 neural and neuro-genetic classifiers
- **Tested on 10 categories of interest (Remote Sensing, Botany, Land Surveying, Cartography, GIS, etc)**
- **83% correct assignment of pages**



# EVA Summary

- **Efficiently scours diverse online information sources**
- **Autonomously accesses, evaluates, retrieves, & fuses information**
- **Learns & evolves into a better information gathering tool**
- **Adapts to each user's styles, techniques, preferences, & interests**
- **Provides individualized results**
- **Implemented in JAVA**



# Topics:

---

1. **EVA – an intelligent information agent system**
2. **eQuery – an NLP-based 2-stage information retrieval system**



# eQuery

## **An NLP-based, 2-stage Information Retrieval System**

- Extracts important concepts, relations and events from text (*both documents & queries*)
- Provides document retrieval, question-answering, and input to visual summarization

# Core Technology

## Natural Language Processing



# Core Technology

## Natural Language Processing

- A technology which enables a system to accomplish human-like understanding of text



# Core Technology

## Natural Language Processing

- A technology which enables a system to accomplish human-like understanding of text
- Extracts both explicit and implicit meaning

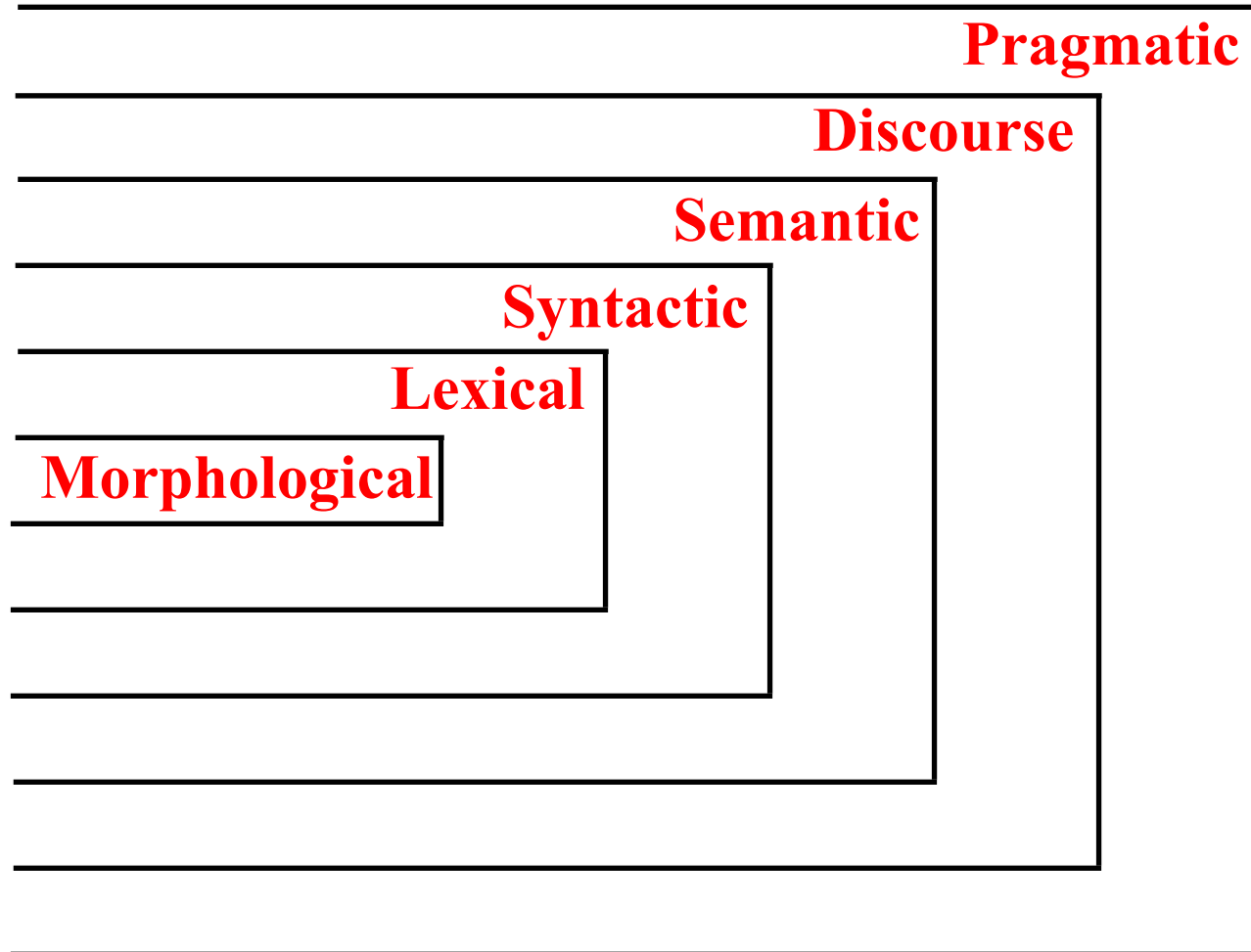


# Core Technology

## Natural Language Processing

- A technology which enables a system to accomplish human-like understanding of text
- Extracts both explicit and implicit meaning
- Utilizes all levels of human language understanding when representing the contents of text

# Levels of Language Understanding





# NLP Processing

---

*03/14/1999 (AFP)...* the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...



# NLP Processing

---

*03/14/1999 (AFP)*... the extremist Harkatul Jihad group,  
reportedly backed by Saudi dissident Osama bin Laden ...

... the|**DT** extremist|**JJ** Harkatul|**NP** Jihad|**NP** group|**NN** ,|,  
reportedly|**RB** backed|**VBD** by|**IN** Saudi|**NP** dissident|**NN**  
Osama|**NP** bin|**VB** Laden|**NP** ...

# NLP Processing

---

03/14/1999 (AFP)... the extremist Harkatul Jihad group,  
reportedly backed by Saudi dissident Osama bin Laden ...

... the|DT extremist|JJ Harkatul|NP Jihad|NP group|NN ,|,  
reportedly|RB backed|VBD by|IN Saudi|NP dissident|NN  
Osama|NP bin|VB Laden|NP ...

... the|DT extremist|JJ <entity> Harkatul\_Jihad\_group </entity> ,|,  
reportedly|RB backed|VBD by|IN <entity> Saudi\_dissident  
</entity> <entity> Osama\_bin\_Laden </entity> ...

# NLP Processing

---

03/14/1999 (AFP)... the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

... the|DT extremist|JJ Harkatul|NP Jihad|NP group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NP dissident|NN Osama|NP bin|VB Laden|NP ...

... the|DT extremist|JJ <entity> ref=1; type=*terrorist group*; Harkatul\_Jihad\_group </entity> ,|, reportedly|RB backed|VBD by|IN <entity> ref=2; type=*nationality* Saudi\_dissident </entity> <entity> ref=3; type=*person*; Osama\_bin\_Laden </entity> ...

# NLP Processing

---

03/14/1999 (AFP)... the extremist Harkatul Jihad group, reportedly backed by Saudi dissident Osama bin Laden ...

... the|DT extremist|JJ Harkatul|NP Jihad|NP group|NN ,|, reportedly|RB backed|VBD by|IN Saudi|NP dissident|NN Osama|NP bin|VB Laden|NP ...

... the|DT extremist|JJ <entity> ref=1; type=*terrorist group*; Harkatul\_Jihad\_group </entity> ,|, reportedly|RB backed|VBD by|IN <entity> ref=2; type=*nationality* Saudi\_dissident </entity> <entity> ref=3; type=*person*; Osama\_bin\_Laden </entity> ...

# NLP Processing

---

03/14/1999 (AFP)... the extremist Harkatul Jihad group,  
reportedly backed by Saudi dissident Osama bin Laden ...

... the|DT extremist|JJ Harkatul|NP Jihad|NP group|NN ,|,  
reportedly|RB backed|VBD by|IN Saudi|NP dissident|NN  
Osama|NP bin|VB Laden|NP ...

... the|DT extremist|JJ <entity> ref=1; type=*terrorist group*;  
Harkatul\_Jihad\_group </entity> ,|, reportedly|RB backed|VBD  
by|IN <entity> ref=2; type=*nationality* Saudi\_dissident </entity>  
<entity> ref=3; type=*person*; Osama\_bin\_Laden </entity> ...

# NLP Processing

---

## EVENT: SUPPORT

relation: agent

entity: osama\_bin\_laden|3 *person*

relation: recipient

entity: harkatul\_jihad|1 *terrorist group*

relation: characteristic

entity: harkatul\_jihad|1 *terrorist group*

state: extremist

relation: manner

state: reportedly

relation: isa

entity: osama\_bin\_laden|3 *person*

entity: dissident

relation: characteristic

entity: dissident

entity: saudi|2 *nationality*



# What NLP can extract:

---

## 1. Domain-independent Entities

- *Person, Country, Organization, Company*



# What NLP can extract:

---

## 1. Domain-independent Entities

- *Person, Country, Organization, Company*

## 2. Domain-independent Relations

- *Agent, recipient, location, point-in-time*



# What NLP can extract:

---

## 1. Domain-independent Entities

– *Person, Country, Organization, Company*

## 2. Domain-independent Relations

– *Agent, recipient, location, point-in-time*

## 3. Domain-dependent Entities

– *Terrorist, subsidiary, perpetrator*



# What NLP can extract:

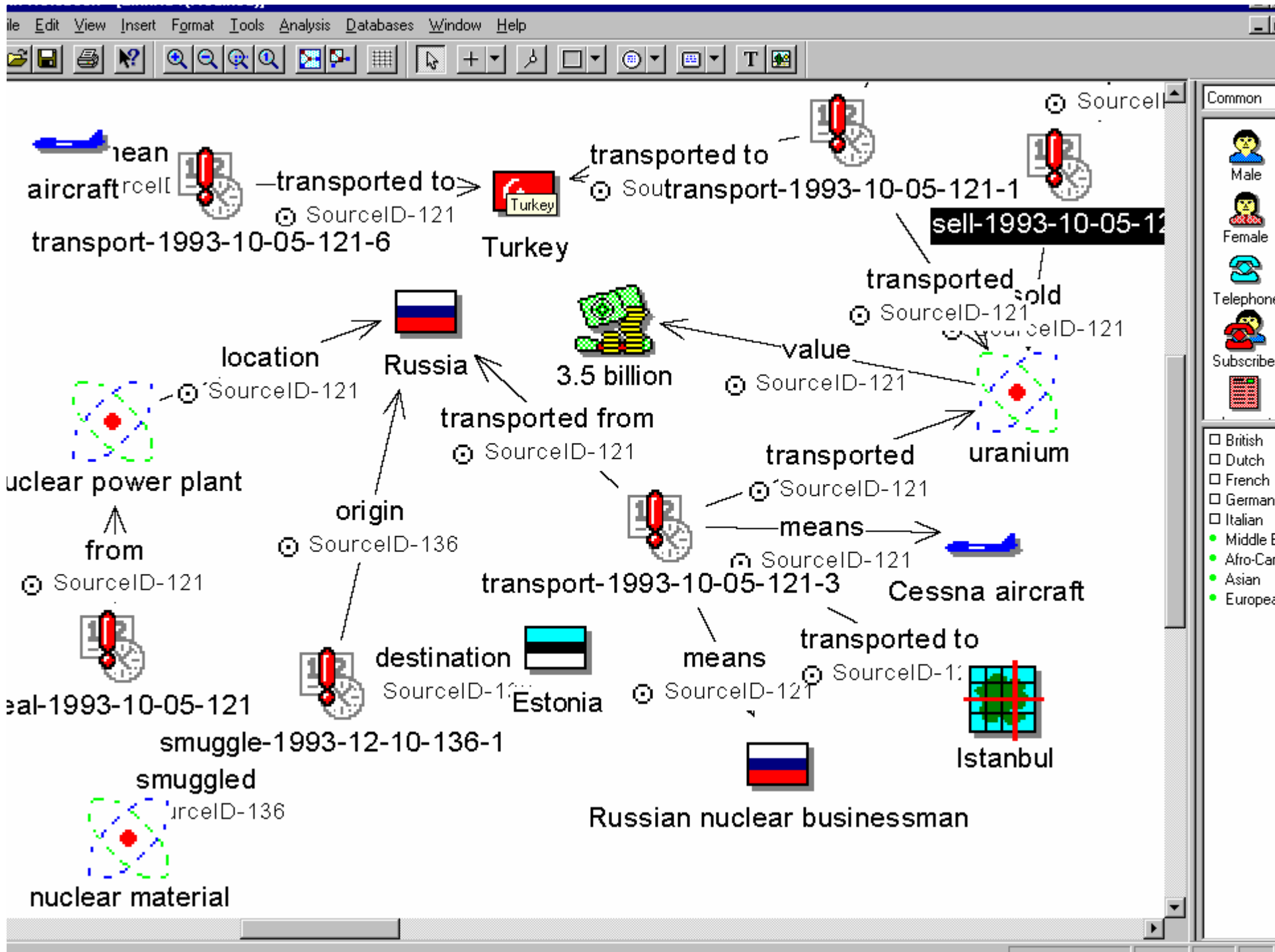
---

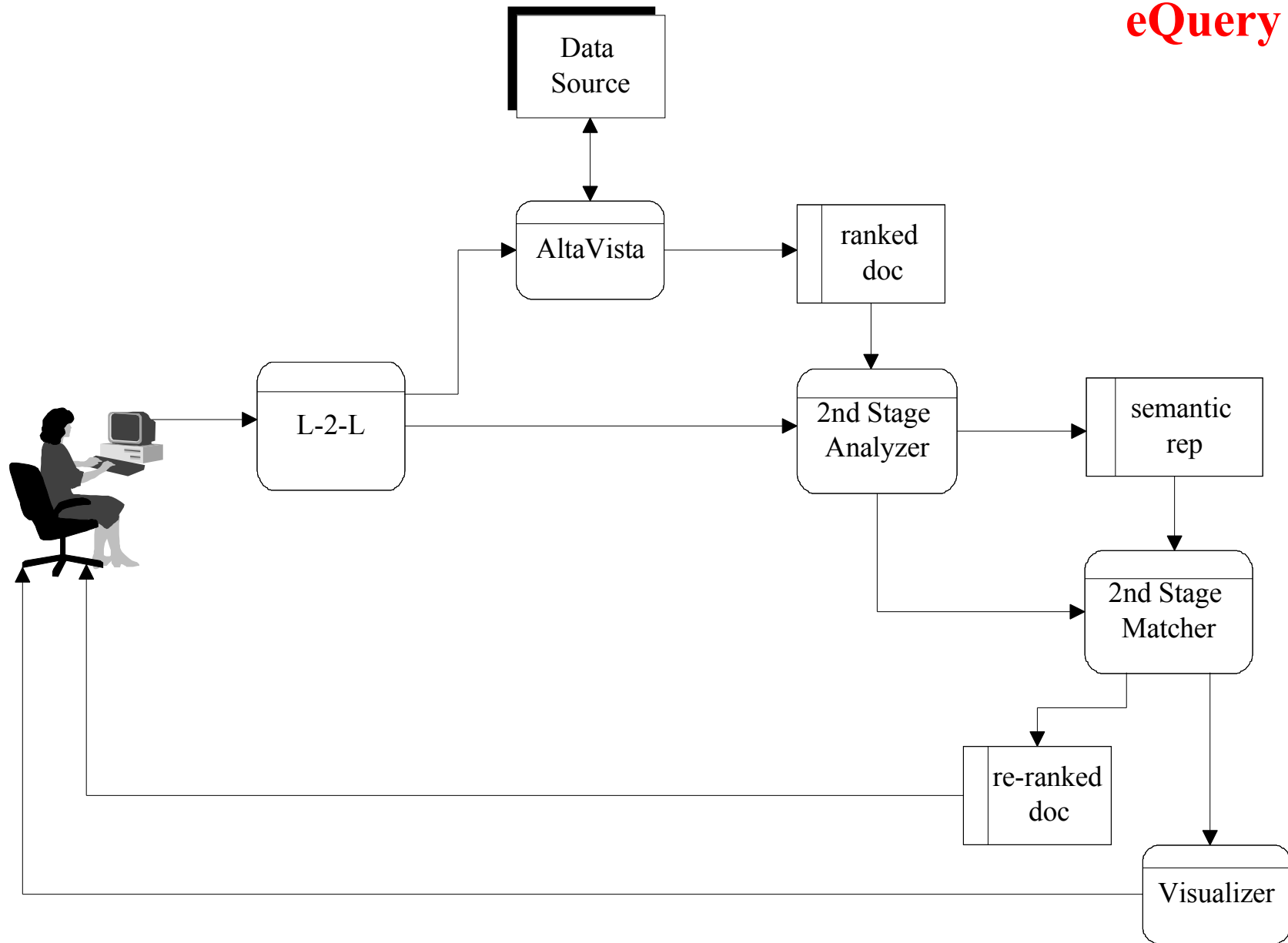
- 1. Domain-independent Entities**
  - *Person, Country, Organization, Company*
- 2. Domain-independent Relations**
  - *Agent, recipient, location, point-in-time*
- 3. Domain-dependent Entities**
  - *Terrorist, subsidiary, perpetrator*
- 4. Domain-dependent Relations**
  - *Takeover, complain*

# What NLP can extract:

---

- 1. Domain-independent Entities**
  - *Person, Country, Organization, Company*
- 2. Domain-independent Relations**
  - *Agent, recipient, location, point-in-time*
- 3. Domain-dependent Entities**
  - *Terrorist, subsidiary, perpetrator*
- 4. Domain-dependent Relations**
  - *Takeover, complain*
- 5. Model-specific Events**
  - *Acquisitions & Mergers*
  - *Terrorism*
  - *Nuclear Smuggling*







## L-2-L: Language-to-Logic

- **Converts a NL query to rich, conceptual, logical search requirements for submittal to COTS search engine(s)**



## L-2-L: Language-to-Logic

- **Converts a NL query to rich, conceptual, logical search requirements for submittal to COTS search engine(s)**
- **Provides correct identification of:**
  - *required* concept
  - implicit logical requirements
  - semantic relations amongst concepts
  - multi-word concepts
  - appropriate stemming of terms
  - expansion of terms with synonymous words/phrases



## L - 2 - L Query Representation

**“I would like information about indictments against  
Bosnian war criminals.”**

*indictment\* +Bosnian “war criminal\*”*



## L - 2 - L Query Representation

**“I would like information about indictments against  
Bosnian war criminals.”**

*indictment\* +Bosnian “war criminal”*

**“I want to know about efforts to bring suspects of the  
Lockerbie bombing to trial.”**

*effort\* bring\* suspect\* +Lockerbie bomb\* trial\**



## Comparison of AltaVista to eQuery

**HPKB Queries**

**AltaVista**

**eQuery**

**Ave. Precision @ 5**

**.40**

**.80**



# Topics:

---

1. **EVA – an intelligent information agent system**
2. **eQuery – an NLP-based 2-stage retrieval system**
3. **Future Research – EVA + eQuery**



# Future Research: EVA + eQuery

---

- **Allow analysts to search for relevant information on more complex topics**



# Future Research: EVA + eQuery

---

- **Allow analysts to search for relevant information on more complex topics**
- **Increase impact of EVA by utilizing richer features in queries, pages, documents**



## Future Research: EVA + eQuery

---

- **Allow analysts to search for relevant information on more complex topics**
- **Increase impact of EVA by utilizing richer features in queries, pages, documents**
- **Provide even better recall & precision for needs of intelligence analysts**



## Future Research: EVA + eQuery

---

- **Allow analysts to search for relevant information on more complex topics**
- **Increase impact of EVA by utilizing richer features in queries, pages, documents**
- **Provide even better recall & precision for needs of intelligence analysts**
- **Empower eQuery to search continuously**



## Future Research: EVA + eQuery

---

- **Allow analysts to search for relevant information on more complex topics**
- **Increase impact of EVA by utilizing richer features in queries, pages, documents**
- **Provide even better recall & precision for needs of intelligence analysts**
- **Empower eQuery to search continuously**
- **Enable continuous updates of visualizations**



# Summary Findings

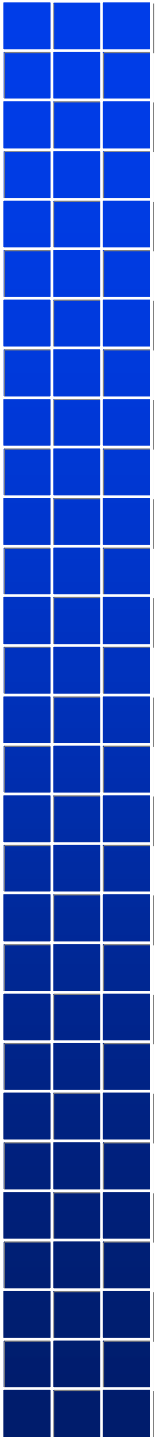
---

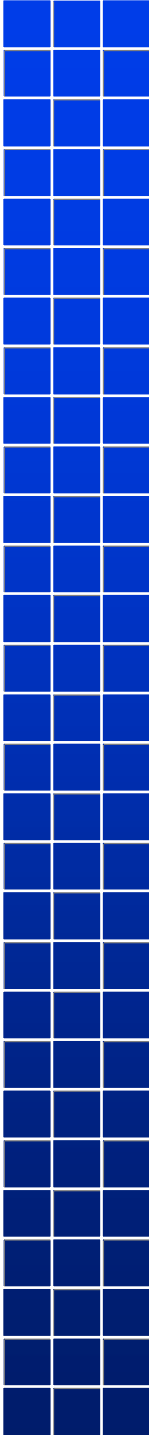


# Summary Findings

---

1. **Ongoing access to geo-spatial information can be effectively achieved with NGA-based, autonomous, adaptive agents - EVA**
2. **NLP-based IR can outperform standard search engines – eQuery**
3. **The 2-staged model can make sophisticated searching efficient - eQuery.**
4. **It is hypothesized that results can be further improved by integrating **EVA** with the 2-stage NLP-based **eQuery** System.**





# **Combining Intelligent Agents with an NLP-based Search Engine**

**Elizabeth D. Liddy, Ph.D.**

**Center for Natural Language Processing  
School of Information Studies  
Syracuse University**

**January 12, 2001**