International Technology Alliance
in
Network & Information Sciences

Information Extraction Using Controlled English to Support Knowledge-Sharing and Decision-Making

17th ICCRTS
June 19-21, 2012

Ping Xue, Stephen Poteet, Anne Kao
Boeing Research & Technology

David Mott, Dave Braines
IBM UK

Cheryl Giammanco, Tien Pham
Army Research Lab

MINISTRY OF DEFENCE

Critical Coalition Need – Sharing at the "edge"

- New military missions (e.g. reconstruction) call for cross-domain information sharing at the edge
- Sharing is more challenging and critical in a distributed environment
- Traditional model, where information flows up, gets processed centrally, and then flows down, does not meet this requirement
- Additional challenges
  - Information requirements for organizations are mission specific
  - Ability to analyze information depends on the operational tempo
  - Require analytics to provide Soldiers with context relevant information for their domain model and workflow
  - Team members from different domains often have different domain concepts due to different perspectives
  - Language variation -- differences in vocabulary, sentence structure, language usage and style
  - Metadata may have different meaning too

International Technology Alliance

- Network and Information Sciences International Technology Alliance (ITA) is a collaborative research alliance between the UK Ministry of Defence (UK MoD) and US Army Research Laboratory (US ARL), and a consortium of leading academic and industry partners.
- The ITA program started on May 12, 2006 with the strategic goal of producing fundamental advances in information and network sciences that will enhance decision making for coalition operations and enable rapid, secure formation of ad hoc teams in coalition environments and enhance US and UK capabilities to conduct coalition warfare. The first phase of the ITA program finished in 2011, and now the program is in its second phase (May 2011-May 2016)
- Part of the goal is to address shared understanding and information exploitation in a coalition environment
  - Work presented here is funded under this ITA effort

Information Extraction

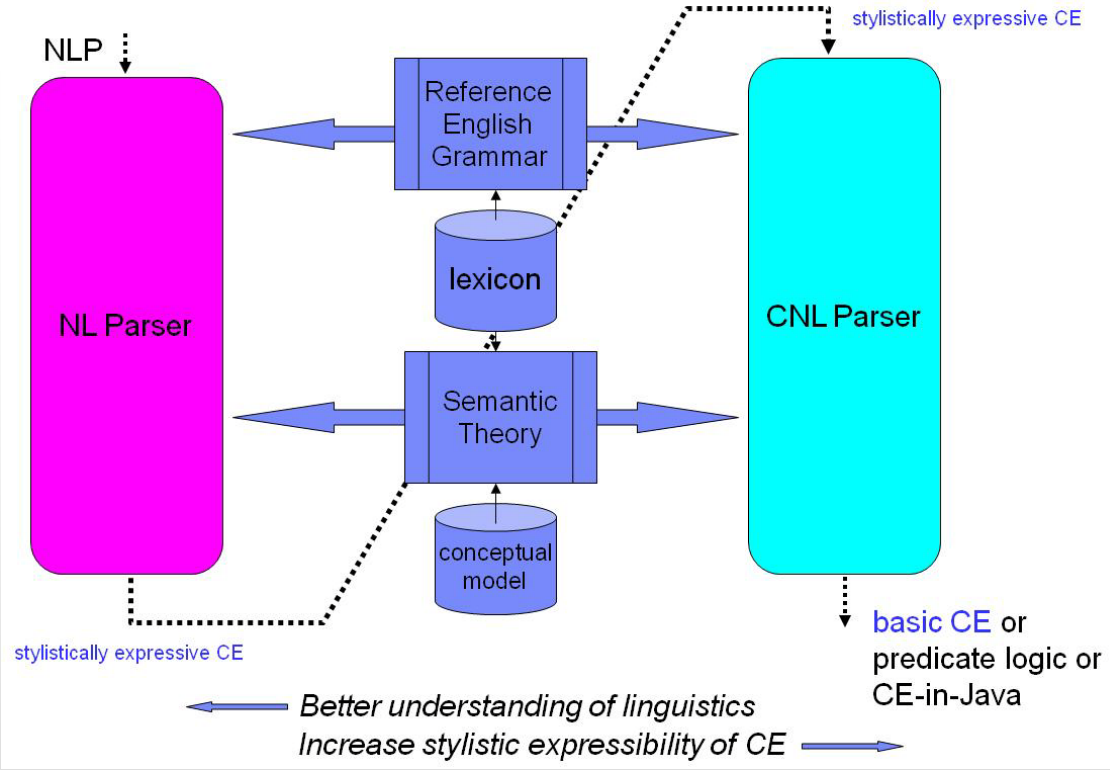
- Much of the info that needs to be shared is available in memos and reports
  - Not currently available for querying or inferencing
  - Needs to be extracted and standardized
- Information Extraction (IE) extracts the following from free text reports:
  - Entities (persons, places, organizations, equipment etc.)
  - Relations between them
  - Events and their participants, time, location and other properties
  - Processes
- Some can be done by general systems, but much requires domain-specific knowledge
- Controlled English provides a means for domain experts (e.g. military analysts) to enter their knowledge and fine-tune the IE system

Controlled English for Knowledge Sharing

- Controlled English
  - Controlled English (CE) is a type of controlled natural language
  - A controlled natural language is a subset of a natural language using a restricted set of grammar rules and a restricted vocabulary
  - Focus can be either for *human readability* or for *machine readability*
    - Our primary focus is the latter – for computer processing
    - However, readability of generated reports also important
- Controlled English can address two critical information needs
  - Need for normalization and organization of free-text description
  - Need for domain expertise for specifying and extending the domain model (ontology)
    - Including extending vocabulary and terminology, and relation to the domain concepts
- Challenge: how to balance naturalness and lack of ambiguity

ITA Controlled English

- ITA CE is consistent with First Order Predicate Logic
  - Based on Common Logic Controlled English (Sowa 2007)
- Syntax is compatible with existing ontology modeling languages such as OWL



Example of ITA CE

- English:
  - BCT patrol in South Baghdad discovers a bomb-making facility on Hilla Road
- ITA CE:
  - The patrol unit ‘|BCT patrol|’ finds the facility ‘|p6|’ and is located in the place ‘|South Baghdad|’ and is a NATO military unit.
  - The facility |p6| makes the device bomb and is located on the road ‘|Hilla Road|’

Unique Combination of Features ITA Controlled English Provides

- A user-friendly language in a form of English, in place of a standard formal *query language* (e.g., SPARQL or SQL), which enables the user to construct queries to information systems in a more intuitive way
- A precise language that enables clear, unambiguous *representation* of extracted information to serve as a semantic representation of the free text data that is amenable to rule-based inferencing
- A *common form of expression* used to build, extend and refine domain models by adding or modifying entity, relation, or event types, and specifying mapping relations between data models and terminology or language variants
- An intuitive means of *configuring system processing* (such as specifying entity types, rules, and lexical patterns)

- Note: users of ITA CE would need brief training to learn ITA CE

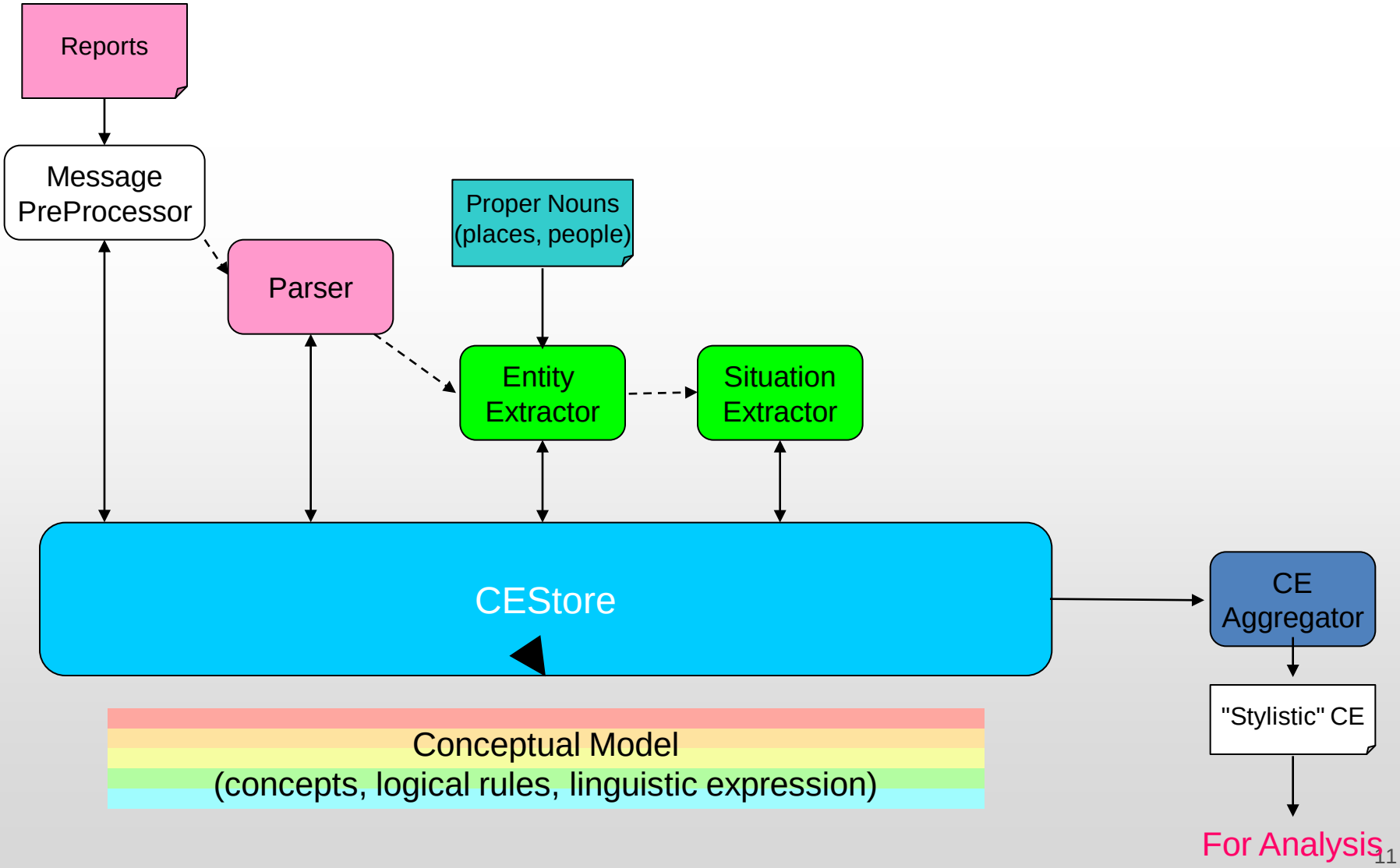
Benefits of ITA Controlled English

- Provides an intuitive, natural language based capability for end user to directly query information as needed from the edge of the network
- Provides user with a better understanding of what the back-end system does (e.g. in inferencing)
- Provides multiple systems a way to exchange information
- Provides end user a way to augment the system's knowledge base
  - E.g. add new vocabulary, relationship
- Provides end user a way to configure the back end system
  - E.g. apply different rules etc.
- Supports users' use of their own domain model and vocabulary to interact with the system

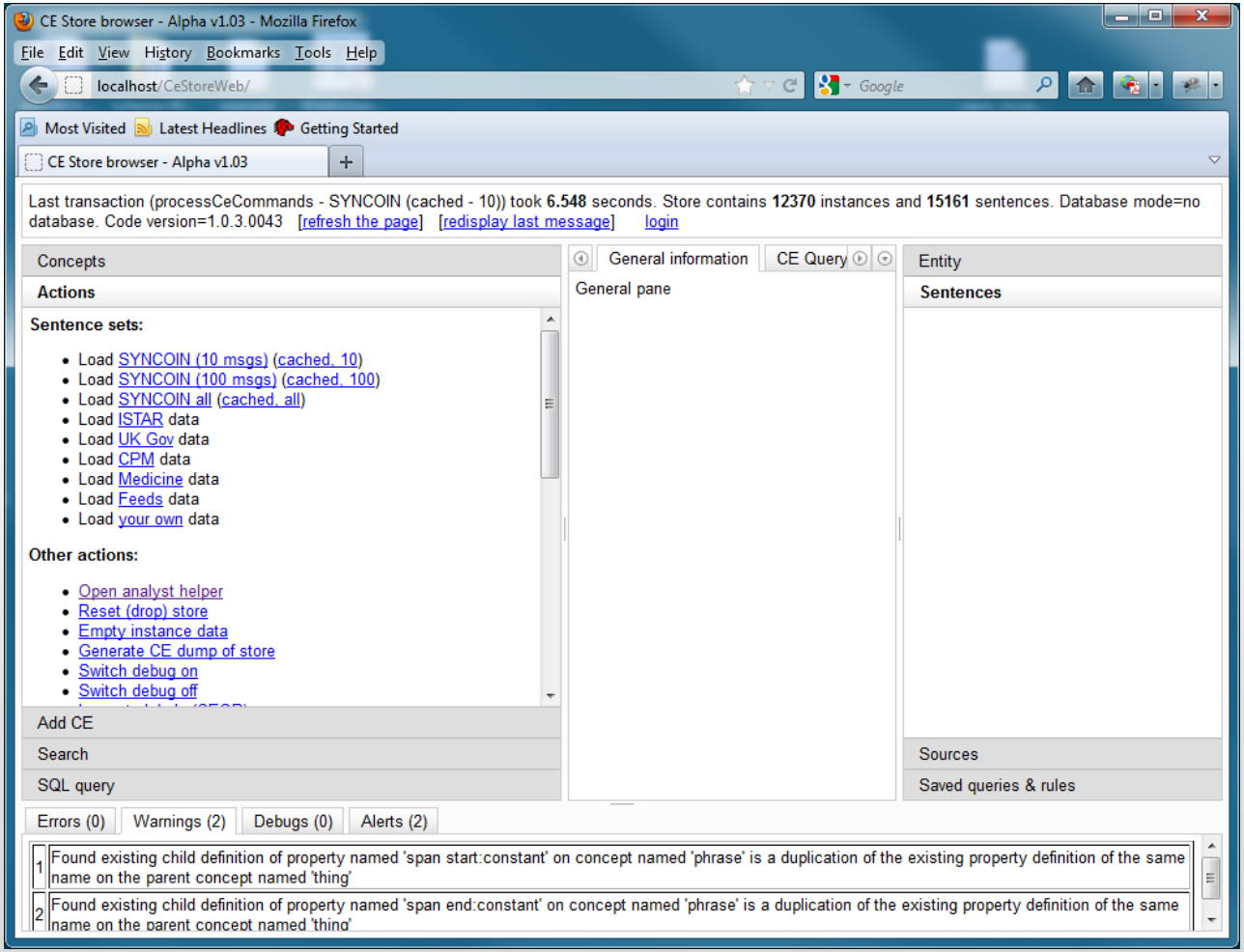
Controlled English for Information Extraction

- Two major purposes of CE for IE
  - CE can represent the end result of linguistic processing - as the unambiguous semantic representation language
    - A *domain model* is needed, but can be constructed and maintained by domain experts using CE
      - E.g. IED, village
  - Provides a means to analyze and process natural language
    - A *linguistic model* is needed, but can be constructed and maintained using CE
      - E.g. lexicon, noun phrase
    - Domain experts can leverage this to specify the language used to talk about things in the domain model

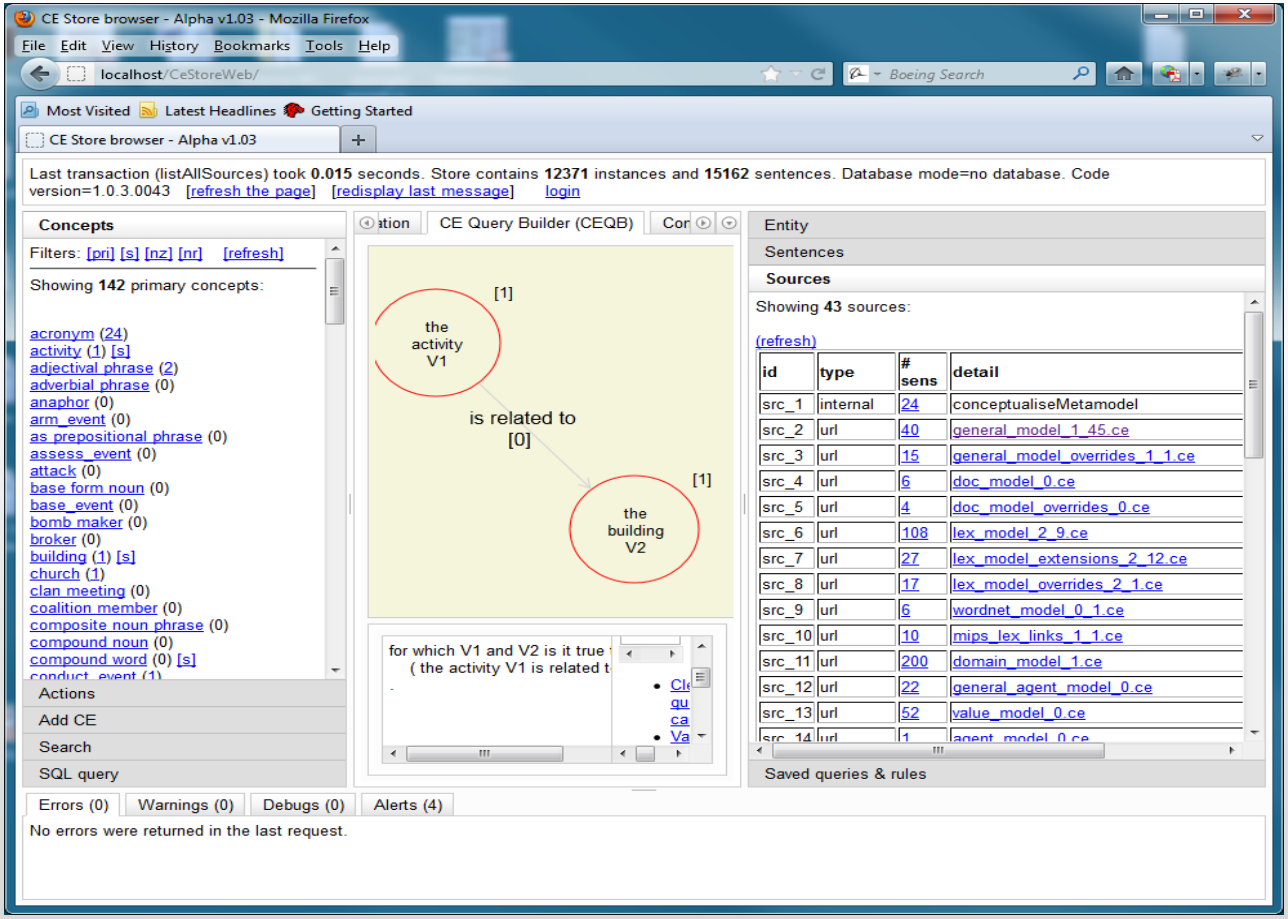
ITA CE Architecture



CEStore: Concepts, Actions



CEStore: CE Query Building



Conclusion

- ITA CE provides
  - A user-friendly language for querying
  - A user-friendly system-to-user report representation
  - A user-friendly way to add to and refine the information extraction system
  - A common form of representation that maps between domain models
  - A common form of expression for terminology or language variants

- Future work
  - Build on the currently completed three phases of transition (at UK) and extend syntax, semantics and domain models
  - Refine ITA CE to make it more user friendly
  - Extend CEStore functionalities and make it more user friendly