Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference

Interactive Information Extraction and Navigation to Enable Effective Link Analysis and Visualization of Unstructured Text

Emily Budlong, Carrie Pine, Mark Zappavigna
Air Force Research Laboratory,
Information Directorate,
Rome, NY 13440

James Homer
National Air and Space Intelligence Center,
Wright-Patterson AFB,
OH 45433

Charles Proefrock, John Gucwa
General Dynamics Information Technology,
Buffalo, NY 14068

Michael Crystal, Ralph Weischedel
Raytheon BBN Technologies,
Cambridge, MA 02138

Abstract

This paper describes the Advanced Text Exploitation Assistant (ATEA), a system developed to enable intelligence analysts to perform link analysis and visualization (A&V) from information in large volumes of unstructured text. One of the key design challenges that had to be addressed was that of imperfect Information Extraction (IE) technology. While IE seems like a promising candidate for exploiting information in unstructured text, it makes mistakes. As a result, analysts do not trust its results. In this paper, we discuss how ATEA overcomes the obstacle of imperfect IE by incorporating a human-in-the-loop for review and correction of extraction results. We also discuss how coupling consolidated extraction results (corpus-level information objects) with an intuitive user interface facilitates interactive navigation of the resulting information. With these key features, ATEA enables effective link analysis and visualization of information in unstructured text.

Introduction

Textual data overload is a pervasive problem impacting intelligence analysts. The amount of potentially relevant text available for analysis is far greater than what analysts can effectively review given limited time. As a result, much of the available data "falls on the floor," negatively impacting the quality of analyses and the quality of decisions based on those analyses. Intelligence analysts would like to use automated analysis and visualization (A&V) tools to exploit the information in unstructured (prose) text, such as intelligence reports, news articles, and documents. However, these tools require data to be in a structured form. As a result, analysts need some way to identify and structure relevant entity, event, and relationship information in text, in order to enable the use of automated A&V tools.

Most capabilities for structuring information found in unstructured text fall into two categories: fully automatic information extraction (IE) or manual markup. Fully automatic IE systems are able to process many documents in a short amount of time and store the entity, event, and relationship information in a structured form. However, results produced by these systems are imperfect and typically do not allow the user to make corrections to the extraction results. As a result, they lose credibility with users.

On the other hand, systems that enable users to manually markup information in text documents and automatically store the information in a structured form have more credibility with users, since in their experience, human markup is more accurate. These systems can speed up the manual markup process over unassisted copying and pasting of information found in text into another tool (e.g., Excel). However, the markup process is still slow and tedious. Therefore, using these systems on large document collections will not solve the textual data overload problem.

In order to address this design challenge, the Advanced Text Exploitation Assistant (ATEA) development team identified imperfect IE as a design constraint up-front. ATEA addresses this design constraint by leveraging the complementary strengths of the human and computer, and mitigating their respective weaknesses. The computer is much faster at processing documents than the human; however, the human is more accurate at identifying and tagging information relevant to them than the computer. ATEA leverages the speed of the computer to process large collections of documents in a short amount of time, and then enables the highly-accurate human to efficiently search, visualize, review, and correct the extraction results as needed. This provides a faster, more effective, and more reliable means to exploit information in unstructured text than either fully automatic or manual markup systems.

Advanced Text Exploitation Assistant

ATEA is a research and development prototype that enables automated extraction and visualization of entity networks based on information in large collections of unstructured text. Its goal is to support the information analyst's mission to develop and maintain situational awareness on a topic of interest. It does so by providing information extraction, entity consolidation, human-in-the-loop review and correction, semantic pattern search, and visualization of extracted information, in an intuitive user interface with multiple synchronized views. This supports automated link visualization, automated electronic dossier capabilities, information discovery, and persistence in a workflow that aligns with the analytic needs for exploiting information from unstructured text.

Processing Pipeline

Before the user can search for information in ATEA, documents must first go through ATEA's document processing pipeline (Figure 1). The processing pipeline uses the Unstructured Information Management Architecture (UIMA) (Apache 2012), which allows for substitution and addition of processing components. This allows for customization of the processing pipeline to meet the needs of various users while minimizing the cost and time required to do so. The base configuration of the processing pipeline uses capabilities for document zoning, information extraction, within-document entity coreference, cross-document entity coreference, and database loading.

The first step in the processing pipeline is document zoning. This is the process of identifying the header, body (the main text section of the document), and footer sections in each document, as well as any metadata found within the header and footer sections. Document zoning helps reduce the amount of noise in the IE output by identifying sections of the document that should not be processed with the IE engine (e.g., header and footer sections). ATEA's document zoner uses a rule-based approach for accurate identification of document metadata (e.g., title, source, and document date) and for identifying the body (text).

Next, ATEA uses IE technology to automatically extract and structure *information-elements* (entities, events, values, relationships, and attributes) from what the document zoner identified as the body section of the document. The extractor identifies:

- Mentions of entities (e.g., people, organizations, geopolitical entities, vehicles), including proper names (e.g., "Fidel Castro"), pronouns (e.g., "he"), and descriptive phrases (e.g., "the leader of the Cuban Revolution");
- Simple verb-based events (e.g., "transports", "attacks");
- Value information (e.g., roles, dates);
- Entity-entity relationships (e.g., "subordinate");
- Entity-event relationships (e.g., "attacker"); and
- Attributes of entities and events (e.g., positions held by a person, the date of an event).

For each entity, event, and relationship identified, the extraction engine provides the *semantic type* of the information, such as ENTITY_PERSON_INDIVIDUAL or RELATIONSHIP_AFFILIATION_MEMBERSHIP. The IE engine used in ATEA's base configuration has been statistically trained to identify information of the entity, event, value, and relationship types and subtypes used in the National Institute of Standards and Technology's (NIST's) 2007 Automatic Content Extraction (ACE) Evaluation (NIST 2007). This extractor provides broad coverage and has been further customized to identify additional topic-specific information.

In addition to extracting entity, event, value, and relationship information, ATEA's IE engine performs within-document entity coreference. Within-document entity coreference is the process of determining whether multiple mentions of an entity within a single document refer to the same real-world entity. This enables consolidation of entity information within a document, particularly the relationships that have been extracted for an entity (Figure 2). For example, if a document reads "Fidel Castro led the Cuban Revolution. He overthrew Fulgencio Batista in 1959", the relationships between "Fidel Castro" and "led", and "he" and "overthrew" may be extracted. However, without within-document coreference to identify that "he" (in the 2nd sentence) is referring to "Fidel Castro" (in the 1st sentence), the connection between "Fidel Castro" and "overthrew Fulgencio Batista in 1959" would not automatically be made and displayed on the link analysis chart. The re-

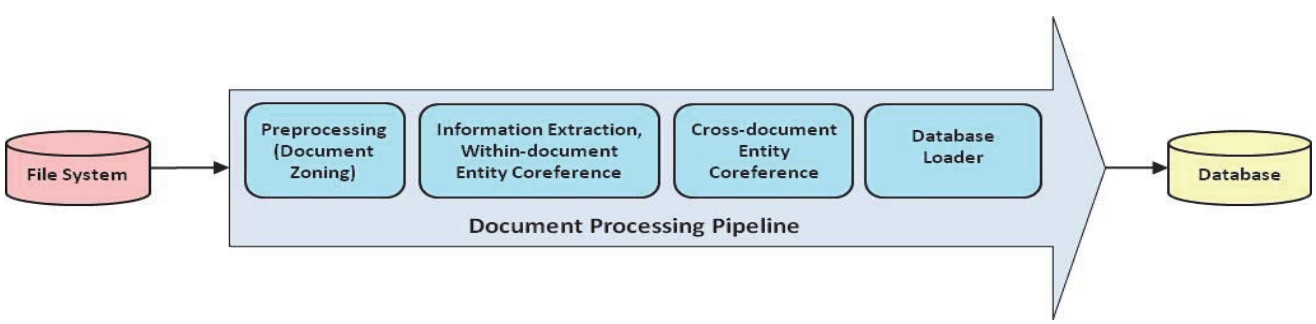

Figure 1: ATEA Document Processing Pipeline

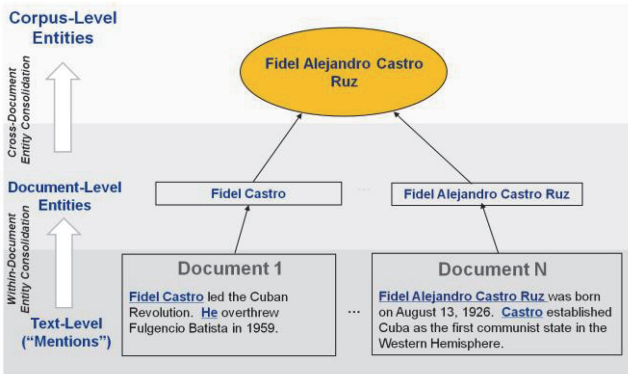

Figure 2: Within- and Cross-Document Entity Consolidation

sult of within-document coreference is a *document-level entity*.

Once within-document coreference is completed, ATEA's processing pipeline performs cross-document entity coreference. This process attempts to consolidate multiple document-level entities into a single *corpus-level entity* for each real-world entity across the entire document collection. This is important, because if a user has a collection of 500 documents about Fidel Castro, using only within-document coreference will result in 500 document-level entities for Fidel Castro. ATEA's cross-document coreference attempts to consolidate those 500 document-level entities into a single corpus-level entity (Figure 2). Creating corpus-level entities is a key aspect of ATEA, as that is what enables it to connect all information extracted on a real world entity to a single node in a link chart.

ATEA's cross-document coreference is currently conservative. This is based on feedback from analysts. They have more trust in a conservative system that under-consolidates entity information than one that is too aggressive and over-consolidates. Leveraging the strength of the human, the ATEA user interface allows the analyst to further consolidate document-level entities into a single corpus-level entity once they have confirmed the accuracy of doing so.

The final step in the processing pipeline is loading all of the extracted and consolidated information into a database. The database maintains the original extraction results with traceability to the source document. It also tracks corrections made by the users.

Search for Information

Once documents have been processed and the extracted entity, event, value, and relationship information is stored in the database, users can perform a *semantic pattern search* to start identifying relevant information. Semantic pattern search is useful because it raises search from the *keyword level* to the *information level.* It does so by leveraging the semantic types of extracted information. For example, using the keyword search "George Washington" to identify information about the person George Washington requires the user to sort through the documents returned that may contain information about President George Washington, George Washington University, and George Washington Bridge. ATEA allows the user to do the semantic pattern search "Who is George Washington?" This search will only return information about the *person* George Washington and any relationships that were extracted about him.

Semantic pattern search also allows the user to perform queries such as "Find all organizations with members," "Find all events with participants," and "Find all people linked to <Country>." This type of search, which leverages semantic categories of extracted information to filter results, helps users get to the *information* they are looking for faster. It is "less noisy" than keyword search alone, which returns *documents* containing various senses of a word, which the user then has to read, interpret and filter themselves. Semantic pattern search also supports information discovery, as the search results can include information the user did not know to look for, such as previously unknown relationships between people and organizations. ATEA provides a set of semantic pattern search templates, as well as interfaces to guide users through creating their own semantic pattern searches.

Navigate Returned Information

Once the user has run their semantic pattern search, ATEA presents the user with an *investigation folder*, which is the set of extracted information that was returned by the semantic search. Investigation folders allow different analysts to work on a subset of information relevant to their own assignments without interfering with each other. ATEA provides multiple synchronized views for the user to navigate the information-elements in their investigation folder (Figure 3, Figure 4). Investigation folders can be saved, allowing analysts to persist their work and return to it at a later time. They also provide a means of preserving corporate knowledge on an area of interest when an employee leaves.

ATEA provides a Document List (Figure 3, top pane) specifying those documents that contain information matching the semantic search. The Document View (Figure 3, left pane) allows the user to see the full text of a selected document[1].

[1] Text in example document from (Wikipedia 2013)

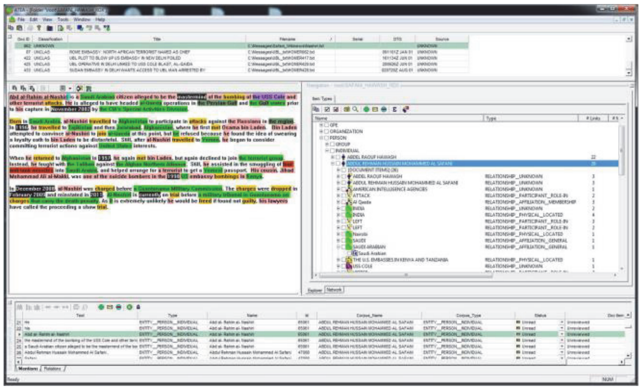

Figure 3: ATEA showing the Explorer Tree

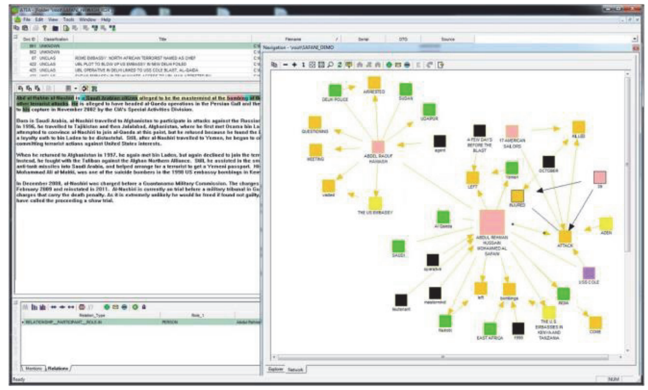

Figure 4: ATEA showing the Network View

ATEA also provides a Grid Pane (Figure 3, bottom pane) that shows the text mentions and relationships that were extracted in a structured form. Each time the user clicks on a mention or relationship in the Grid Pane, the views synchronize to show the user where in the original document the information was extracted from, allowing the user to check the correctness of the extracted information.

An Explorer Tree (Figure 3, right pane) organizes the returned information by semantic type in an expandable and collapsible navigation tree. This helps the user quickly navigate to what they are looking for and see what other information was returned by their search. The Explorer Tree shows the user information consolidated across the entire document collection for each returned entity, event, and value. It also lets them drill into the specific documents the information was extracted from.

Clicking on an information-element in the Explorer Tree will synchronize all of ATEA's views. The Grid Pane will show all mentions of the information-element, the Document List will show what document the currently selected mention was extracted from, and the Document View will show where in the original document the mention was found. This intuitive, analyst-centric navigation through information space has received a lot of positive feedback from our users.

ATEA also has a Network View (Figure 4) that visualizes the extracted information at the *corpus-level*. The Network View shows the user a link chart in which the nodes are corpus-level entities, events, and values, and the edges between the nodes are all the relationships and attributes extracted between them, from across the entire document collection.

Similar to the Explorer Tree, clicking on a node will synchronize the other views to show all of the extracted and cross-document consolidated mentions of that item in the Grid Pane, as well as where in the original document the currently selected mention was extracted from[2]. Clicking on an edge in the Network View will also synchronize the views, allowing the user to see where in the original document the relationship was extracted from, and to learn the context in which the relationship occurs.

ATEA provides an "Expand" function for entities, events, and values in the Explorer Tree and Network View. This allows the user to quickly bring additional mentions, relationships, and attributes, extracted from multiple documents, for the item of interest, into their investigation folder. ATEA's network chart in particular supports information discovery by giving the user the ability to follow the information wherever it leads them. It has the ability to "travel to" different nodes, which changes the anchor node of the network to show information about a different entity or event. Combined with the "Expand" function, this lets the user dig deeper into multiple entities and events without leaving the Network View. For example, if the initial search was "Who is Al Safani?" the user might start with their network chart anchored on Al Safani. In that chart, they notice a relationship between Al Safani and Hawash that they did not know about. They can "Expand" the node for Hawash to add additional information about him to the investigation folder, and then "travel to" Hawash in the network to view that new information, enabling them to quickly learn more about his relationships and attributes.

ATEA has been designed to work on dual monitors, which allows the Explorer Tree and Network View to use an entire monitor each for easier viewing. This is particularly useful when viewing large link charts in the Network View, as they can easily take up an entire monitor.

Human-in-the-Loop Review and Correction

One of ATEA's key innovations is that it leverages the analytical strength of the human to review and correct extracted information, manually mark up additional information of interest, and indicate what information is relevant to their task. This improves acceptability of automatically extracted information and increases user confidence by enabling them to examine, validate, and correct the ex-

[2] Text in example document from (Wikipedia 2013)

tracted information. Users can make corrections to any of the extracted information, including correcting within and cross document coreference, modifying the span of a text mention, updating document- and corpus-level entity display names, correcting the semantic type of the information, and correcting relationship information.

All of the changes made during review and correction are stored in the database, so that others can benefit from the corrections and additions made. Additionally, users can see if information has already been marked as "Reviewed," meaning that a human has verified that the information was extracted correctly. This enables groups of users to work collaboratively and save time by leveraging the work of their colleagues.

Enabling Other Analysis and Visualization Tools

ATEA has been designed to support the use of other analysis tools, in particular, to enable the use of automated link analysis and visualization tools. ATEA allows the user to export investigation folder results in XML or CSV format, which then can be transformed and imported into other automated analysis tools. It also has database views and web services for obtaining extraction results, including any additional markup and corrections made through ATEA's review and correction workflow. ATEA also includes web services for submitting documents to the processing pipeline from other tools, such as a front-end document retrieval tool. These capabilities allow ATEA to be used in a workflow that aligns with the analytic needs for exploiting information from unstructured text, which includes leveraging document retrieval, information extraction, and automated link analysis and visualization tools.

Systems Engineering Approach and its Contribution to Achieving an Intelligent User Interface

ATEA was developed using an evolutionary spiral prototyping process. This was a gradual process that typically involved some "hits" and/or "misses" at each step. At the end of each spiral, the development team would get analyst feedback on the current prototype; update and identify requirements from user, developer, and program management perspectives; prioritize and scope those requirements based on available resources; and then plan for the next spiral. This process was repeated over five years, resulting in gradual improvements that achieved a baseline capability that shows promise in meeting users' requirements.

Several key innovations in ATEA's user interface are the result of analyst feedback. For example, an early version of ATEA did not include the Network View, as the development team had been told that other link analysis tools existed. However, analyst feedback was that they needed a better way to visualize the information across the document collection prior to exporting information to the large scale link analysis tools. The analysts were also adamant about being able to see the extracted information in context in the original document, as they would not trust a "black box" that simply output structured information. This led to development of multiple synchronized views that always shows the extracted information in context in the original document. The Network View ability to "travel to" new nodes, and expand nodes to bring in additional information and relationships about that entity/event, was also a result of analyst feedback. They needed a straightforward way to "chase the rabbit", to follow the information wherever it took them.

A Closer Look at ATEA's IE Component

The previous sections have presented ATEA at a fairly high, system-level. This was a conscious decision, in order to emphasize that it is the *system-level* design innovations that set ATEA apart, enabling successful transition of imperfect IE technology from the lab and into a useful text exploitation capability. But the Artificial Intelligence technology that makes ATEA's functionality even possible is IE technology. This section will take a closer look at ATEA's IE Component.

Vendor-Independence

Vendor-independence - not being "locked-in" to using any one vendor's or researcher's solution - is a key design criterion for ATEA's IE component. This is important because IE technology is continuously evolving, so it is impossible to know whose technology will best meet users' needs in the future. Additionally, different IE capabilities may be customized to perform better for different users' domains (areas of interest). As a result, it is in the users' best interest for the ATEA design to be flexible enough to use whichever IE engine best meets their specific needs at a given point in time. The UIMA architecture, with its plug-in, plug-out components, was selected for ATEA because it supports this design criterion, and because it is an emerging community standard. To date, three different IE engines have been integrated into ATEA. Each has its own strengths and capabilities that make it better suited for a particular research goal.

Baseline Instantiation: SERIF

The IE engine used in ATEA's baseline configuration is BBN's Statistical Entity & Relation Information Finding (SERIF) system (Boschee, Weischedel and Zamanian 2005). SERIF is a general language understanding component that has been used in applications such as distillation, question answering, and extraction in English, Chinese, and Arabic. To extract information from text, SERIF determines a logical form for each sentence, resolves co-

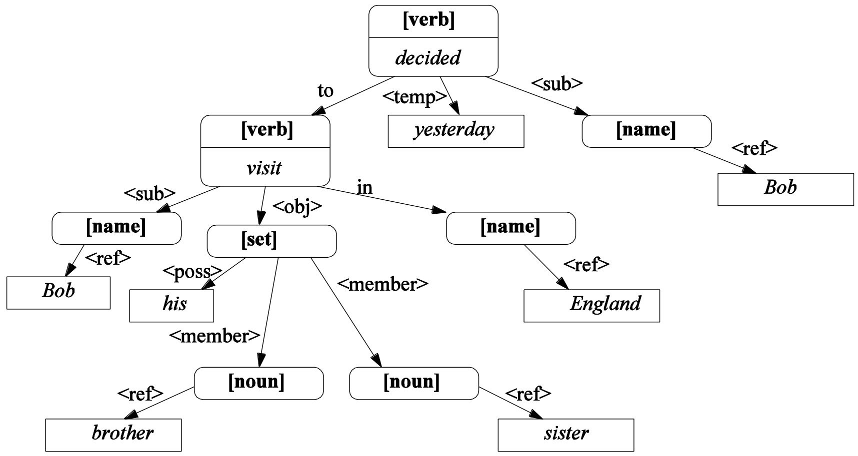

Figure 5: Example Proposition Tree

reference (e.g., pronouns) within and across sentences, and determines the meaning of ambiguous words in context.

There are four dimensions to interpreting text in SERIF:

(1) **Parsing** text using a lexicalized probabilistic context-free grammar. The F-measure on parsing benchmarks is ~89%.

(2) A small number of rules that map parse trees into a simple (propositional) **logical form**. Propositions, the logical form for a sentence, are represented in one coherent tree structure giving the semantic interpretation of the sentence, abstracted from its syntactic realization. A sentence may consist of one or several proposition trees, created by replacing mention arguments of propositions with propositions referencing those mentions. Figure 5 illustrates the proposition tree for the sentence: *"Yesterday, Bob decided to visit his brother and sister in England."* Intermediate nodes in the tree represent propositions. Leaves (e.g., *sister*) represent arguments of propositions that cannot be expanded. Both intermediate nodes and leaves may have text associated with them. The labels on the arcs between nodes are the argument labels.

(3) **Resolving coreference** (e.g., pronouns or definite descriptions) using a hybrid of statistical predictions and handcrafted rules. Proposition trees are further enhanced with coreference, e.g., if *his brother* is mentioned by name earlier in the document, the name is associated. The algorithm performs a left-to-right scan through the text, one sentence at a time, classifying whether a group of words – usually a noun phrase, a pronoun, or a verb – in the sequence represents the introduction of a new entity, or a reference to a previously identified entity. It does this using five models, one each for proper names, nominal mentions of known named entity types, pronouns, other nominal mentions, and verbs. The features extracted by these models primarily are based on the words themselves, and other syntactic and semantic information tagged on top of the text as a pre-processing step using other statistical models such as a syntactic parses, named entity tagging, proposition structure, etc.

(4) Mappings from propositional form to the **customer-specified ontology**. In this case, the ACE entity, event, value, and relationship structures.

One of the main reasons SERIF was selected as ATEA's baseline IE component was that it already met many of ATEA's IE requirements. It provided a general-purpose extraction engine that could be further customized to meet additional, user-specific IE requirements. In addition, it had been a top performer at NIST's ACE competitions (NIST 2007). These and other factors made SERIF an attractive choice as the IE component in ATEA's baseline configuration.

ATEA System Evaluation

The benefits of using ATEA have been demonstrated in functional experiments. In one experiment, the task was to identify the entities in a social network from a collection of 1,250 documents. Manually, it took one analyst 6 months to complete the task, and when finished, she swore she would never do it again. During the task, the analyst was only able to identify the entity information in 10-20 documents a day. Using ATEA to complete the same task, it took another analyst just 2 weeks to identify the same social network. Here, it took roughly 3 hours to process the documents with ATEA, and the analyst spent the remainder of the time using ATEA to identify the relevant entity information to form the social network. With ATEA, an analyst identified the same social network 13 times faster.

Another experiment looked at the amount of time it took to answer 15 requests for information (RFIs) using a collection of 88 documents. Using a collaborative search, discovery, and knowledge management tool that allowed the users to manually markup information, it took 8 analysts 24 hours collectively (3 hours per analyst) to fully answer 4 RFIs and partially answer 8 RFIs, leaving 3 RFIs unanswered. Another analyst processed the same document collection with ATEA, and using its user interface was able to answer all 15 RFIs, in some cases with more complete answers, in less than 3 hours. In this case, the task was accomplished 8 times faster using ATEA and more complete answers were achieved.

Related Work

There has been little work that combines use of broad coverage, corpus-level information extraction; a human-in-the-loop, both for review and correction of imperfect extraction results and to navigate the information-space; and the use of multiple synchronized views to intuitively navigate results. (Culotta, Kristjansson, McCallum and Viola 2006) describes a system that automatically extracts contact information from unstructured online text sources (e.g., e-mail messages, web pages). The system supports the user in form filling by automatically populating the fields with extracted information, allowing the user to view the document the information was extracted from, and to correct the extracted information where necessary. The system then feeds the corrections back to the extractor to improve performance and reduce future corrections needed. While this system does support correction of extraction results, its extraction scope is much narrower than ATEA's broad coverage extraction, and its user interface is simpler than the interface ATEA provides to support information navigation at the document and corpus levels.

(Li, Srihari, Niu and Li 2003) introduces the concept of corpus-level Entity Profiles (EPs), which consolidate all information extracted on an entity from a collection of documents. ATEA leveraged this concept.

(Crystal and Pine 2005) discusses an R&D prototype that applied IE with a human-in-the-loop for automated org-chart generation (AOCG). The AOCG was narrower in scope than ATEA in terms of its task, the type of input used (message traffic) and the types of relationships extracted (subordination relations). The AOCG supported multiple views for more intuitive analysis of information from text, but was document- vs. information-centric.

(Minton et al. 2012) describe the Entity Intelligence Portal (ENTEL) that helps intelligence analysts track and analyze information on entities from many Web sources. ATEA and ENTEL are similar in a number of ways: both help intelligence analysts exploit information from large volumes of text, both use information extraction technology, both are heavily dependent on entity resolution technology, and both systems' architectures are viewed as information pipelines. The main differences between ATEA and ENTEL result from the different user applications they were designed to address. Where ATEA is primarily focused on enabling users to perform link analysis and visualization from information in unstructured text (typically documents, reports, and articles), ENTEL is primarily focused on monitoring and tracking entities of interest from open sources on the Web (typically accessing semi-structured data sources such as public databases). Because ENTEL exploits Web sources, it uses an agent platform to harvest web data on an ongoing basis. ATEA does not; it loads documents from a directory for processing. Additionally, because ATEA is focused on processing unstructured text documents, it emphasizes review and correction of imperfect IE results, in order for analysts to trust what they are using. ATEA also places a larger emphasis on a user interface that supports intuitive navigation, helping analysts discover additional entities associated with an entity-relation network of interest.

There has been more work in the area of interactively training an extraction engine. (Jakel et al. 2012) allows a user to interactively train an extraction engine by providing feedback to the system to learn from in subsequent training iterations. The system described in (Das Sarma, Jain and Srivastava 2010) helps users and developers understand why certain information was extracted, diagnose problems with the extraction results, and repair the extraction system to improve performance on future iterations. Although these systems support user interaction with an extraction system, it is for the purpose of training and improving the performance of the extractor itself, rather than enabling the user to leverage imperfect IE results for analytic tasks as ATEA does. While not feasible for the initial ATEA baseline, incorporating an interactive training capability is a future goal for the system.

Conclusion

The Advanced Text Exploitation Assistant has shown that by leveraging the complementary strengths of the human and computer, it is possible to enable effective link analysis and visualization from large volumes of unstructured text. By leveraging information extraction technology that incorporates a human-in-the-loop for review and correction of corpus-level extraction results, along with an intuitive user interface to navigate the results, ATEA achieved a text exploitation capability that is both useful and useable.

Acknowledgements

The authors would like to acknowledge Dr. Dan Howland, Mr. Bill Taylor, Dr. Jeannette Neal, Mr. Dave Sanders, Dr. Janet Miller, Dr. Rohini Srihari, and Mr. Richard Smith for their key contributions to this research over the years, as well as all the other people who have supported us.

References

Apache. 2011. *Apache UIMA*. [Online]. Available from: http://uima.apache.org/. [Accessed: 17 September 2012].

Boschee, E.; Weischedel, R.; and Zamanian, A. 2005. Automatic Information Extraction. In *Proceedings of the First International Conference on Intelligence Analysis*.

Crystal, M. R and Pine, C. 2005. Automated org-chart generation from intelligence message traffic. In *Proceedings of the First International Conference on Intelligence Analysis*.

Culotta, A.; Kristjansson, T.; McCallum, A. and Viola, P. 2006. Corrective Feedback and Persistent Learning for Information Extraction. *Artificial Intelligence* 170(14): 1101-1122.

Das Sarma, A.; Jain, A. and Srivastava, D. 2010. I4E: Interactive Investigation of Iterative Information Extraction. In Proceedings of the 2010 ACM SIGMOD, 795-806. New York, N.Y.: ACM.

Jakel, R.; Metzger, S.; Daivandy, J. M.; Hose, K.; Hunich, D.; Schenkel, R. and Schuller, B. 2012. Interactive Information Extraction based on Distributed Data Management for D-Grid Projects. In *Proceedings of EGI Community Forum 2012 / EMI Second Technical Conference*, PoS(EGICF12-EMITC2)031.

Li, W.; Srihari, R.; Niu, C. and Li, X. 2003. Entity Profile Extraction from Large Corpora. In *Proceedings of the Pacific Association for Computational Linguistics 2003.*

Minton, S.; Macskassy, S.; LaMonica, P.; See, K., Knoblock, C.; Barish, G., Michelson, M. and Liuzzi, R. 2011. Monitoring Entities in an Uncertain World; Entity Resolution and Referential Integrity. In *Proceedings of the Twenty-Third Innovative Applications of Artificial Intelligence Conference*, 1681-1688. Palo Alto, CA:AAAI.

NIST. 2007. *The ACE 2007 (ACE07) Evaluation Plan*, 22 February 2007. [Online]. Available from: http://www.itl.nist.gov/iad/mig/tests/ace/2007/doc/ace07-evalplan.v1.3a.pdf. [Accessed: 18 September 2012].

Ramshaw. L.; Boschee, E.; Freedman, M.; MacBride, J.; Weischedel, R. and Zamanian, A. 2011. Serif Language Processing – Effective Trainable Language Understanding. In Olive, J.; Christianson, J.; McCary, J., editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation.* Springer, pp. 636-644, 2011.

Wikipedia. 2013. *Abd al-Rahim al-Nashiri*. [Online]. Available from: http://en.wikipedia.org/wiki/Abd_al-Rahim_al-Nashiri. [Accessed: 4 April 2013]. Used under Creative Commons Attribution-ShareAlike license: http://creativecommons.org/licenses/by-sa/3.0/.