# Researching Persons & Organizations

## AWAKE: From Text to an Entity-Centric Knowledge Base

Elizabeth Boschee, Marjorie Freedman, Saurabh Khanwalkar, Anoop Kumar, Amit Srivastava, Ralph Weischedel
Raytheon BBN Technologies Corp.
10 Moulton St.
Cambridge, MA, USA

***Abstract*—We describe a pilot experiment building a capability to automatically read documents, develop a knowledge base, support analytics, and visualize the information found. The capability allows someone researching a topic of interest of focus on analysis and synthesis rather than on reading. We show how information from multiple modalities (speech, text, structured databases) and multiple approaches (ontology driven and open information extraction) can be fused to create a resource about both previously known and novel entities. We describe an extensible framework for language understanding tools that allows for scalability, plug-and-play of alternative components, and incorporation of additional input streams, including video, images, and foreign language text.**

***Keywords—entity disambiguation, entity discovery, automatic knowledge base construction. information extraction***

## I. INTRODUCTION

Fusing information in a knowledge base from a variety of modalities (e.g. written sources, audio sources, and structured sources) requires a range of technologies, e.g. speech recognition, information extraction, and natural language processing(NLP). Here we describe a framework and approach to fusing information across several modalities and technologies, focusing on a specific application that enables faster, more effective research about people and organizations. To perform such research today, one reads (or listens to) a wide range of sources (interviews, news articles, existing databases) gathered for example via web query. If one then wants to use automatic analysis tools, one must manually enter key information into a knowledge base (KB) about the persons/organizations of interest. Our goal is to automate the process of understanding text and mapping it into a knowledge base, to support both automatic analytics on the discovered facts and manual browsing of the knowledge base. Access to the KB allows a researcher to spend less time reading (and doing data entry) and more time on interpretation of the data. We focus on three sources of data: newswire, broadcast news, and the crowd-sourced, structured knowledge represented in DBpedia.

## II. CHALLENGES TO THE STATE OF THE ART

A full survey of the challenges of information extraction (IE) from text is beyond the scope of this paper. Rather, we summarize here several key challenges faced in extracting a coherent set of facts about persons and organizations from a large body of multi-media sources.

*Deduplication*. Many genres have documents that are substantially repeated, such as later versions of the same newswire story, re-tweets, and quotations in discussion forums. Similar challenges arise with the audio track of YouTube posts. Unless the system recognizes such duplication, frequency statistics in big data, which can be informative measures of confidence and importance, will be skewed.

*Name Variation & Entity Disambiguation*. Name variation arises from using only part of the full name, abbreviations, acronyms, transliteration, and spelling/typographical errors. For example, a few of the full forms we have observed for Hezbollah are Hizbollah, Hezbola, Hizbullah, Hezbullah, and Hizbulla. Developing an accurate knowledge base requires that a system when given a mention of a person or organization in a document is able to match that entity to its appropriate referent regardless of how that mention is spelled. In addition to recognizing variation in an entity's proper name, the system must also distinguish between multiple individuals sharing the same name (e.g. *The Times* is the formal name of newspapers in both Louisiana and New Jersey and the informal referent of many others). The challenge of disambiguation grows as the volume of data to be processed increases.

*Entity Discovery*. New individuals and organizations arise continually in a growing collection. Even given a starting entity database containing more than one million entries, only about 25% of named persons and organizations in our corpus were judged by the AWAKE system to map to a known entity with reasonable confidence. To support an accurate analytics, the system must detect such new real-world persons and organizations and add them to the KB. A particular challenge is detecting new individuals that share name strings with known individuals.

*Mapping text to knowledge*. The same relation may be stated (or implied) in quite diverse ways in the original documents. For example, "*Edison invented the light bulb*," "*Edison built the first light bulb*," and "*Edison was granted a U.S. patent for the light bulb*," all imply an inventor/invention relationship between Edison and the light bulb. Detecting a relationship, classifying it into a particular ontology, and extracting its arguments correctly are crucial to building a

knowledge base. We align fact extraction output with facts as represented in a crowd-sourced knowledge base (DBpedia). Here (and also in the case of aligning distinct fact extractors) the challenge is not the numerous ways in which language can represent a fact, but rather the challenge of ontology alignment between similar but not identical representations of knowledge.

*Erroneous Extractions*. Information extraction technology will generate incorrect information. In NIST's TAC 2013 evaluation of fact extraction, the top performing system's F-Score was 37.3 (out of 100)[1]; the highest level of precision in the top 5 systems was 61.4%. A means of coping with that error rate is essential. Furthermore, we target the synthesis of several different research technologies (speech recognition, fact extraction, and entity linking). None of these technologies are perfect and errors in one can impact another. For example, an error in speech recognition can lead to the erroneous extraction of a fact.

*Fusion with External Sources*. The challenge of name variation and entity disambiguation arises from different external databases, since there is no guarantee even in manually curated data that the same canonical name will be used in the different databases.

## III. Related Work

Several researchers have explored elements of the language processing tasks used to develop AWAKE. NIST's TAC Knowledge Base (KB) Population Evaluation includes three relevant tasks: Slot Filling, Entity Linking, and ColdStart. The data processed for the TAC tasks is text[4], and does not incorporate speech input as in our work. A second point of departure is the relation set. While some of the relations in our ontology (*title*, *affiliation*) are similar to their TAC counterpart (*title, member_or_employee);* we incorporate several relations which change more frequently (*statement, visited)*. The latter type of relation, our incorporation of OpenIE, and our *Updates* view provide access to novel, interesting information about entities even when some information is already known.

The Slot Filling (SF) evaluation focuses exclusively on structured relation extraction and largely ignores the challenge of cross-corpus entity linking. SF is structured such that rather than extracting all relations and building a KB, participants typically perform an information-retrieval style query and extract facts over only those documents retrieved [5], [6], [7]. A retrieval based approach does not support the graph-based view which we described earlier. Describing the details of our document-relation extraction system is not the focus of this paper, but as with many other systems we combine linguistic processing, distant supervision, and some number of hand-crafted rules [8], [5].

Entity Linking (EL) focuses entirely on the process of resolving name-strings to KB entities. As with the Slot Filling task, systems do not build a KB, but rather assign an ID to name strings, each of which has provenance in some document. As such, systems process thousands, not hundreds of thousands of documents (as described in this work). Furthermore, systems are only responsible for linking selected entities (and not every named entity in the corpus). Our framework for entity-linking and discovery follows from this prior work in that most EL systems factor the task into resolution to a pre-existing KB and detecting unknown entities [9], [10]. The KB used for TAC has been manually curated and thus has fewer errors than the DBPedia crowd-sourced KB. The TAC KB is also smaller than our KB (~1.1M nodes contrasted with 818K nodes in TAC KB) [4]. We incorporate many of the features described by EL participants: Wikipedia aliases, edit-distance, and document relations [9], [10]. Similar features are used for Wikification [11]. However, Wikification systems also typically make heavy use of Wikipedia-specific information.

The ColdStart evaluation is designed to measure performance of the full KB population task and thus most closely mirrors our approach. One participant, [12] uses SERIF's core document-level processing. Our system differs from ColdStart in its larger scale: processing 720K rather than 50K documents. This means not only that AWAKE must grapple with contradictory information, but also that it can harnesses redundancy to improve precision. Such research, one of the core opportunities and challenges of big data, is not represented in ColdStart where only ~5% of the correct answers have more than one justification in the corpus. Furthermore, ColdStart explicitly targets knowledge population without use of external resources; our work focuses on leveraging and fusing information from multiple sources; both structured and unstructured.

Several frameworks and toolkits have been proposed and employed to handle human language technology (HLT) processing. Apache Unstructured Information Management Architecture (UIMA), an open source implementation of UIMA, supports configuring and running complex processing pipelines on large amounts of unstructured text and discovering knowledge that is relevant to users [13]. Many additional frameworks and toolkits, such as, General Architecture for Text Engineering (GATE) [14], Natural Language Toolkit (NLTK) [15], and OpenNLP [16], are also available. Software integration and communication frameworks, such as, ICE [17], Thrift [18], and SOAP [19], support defining data structures in a language-independent manner and protocols to exchange objects. The ADEPT framework defines common data structures for HLT algorithms to represent text documents, outputs, and algorithm interfaces. Moreover, ADEPT provides an API for database and knowledgebase integration along with support to represent custom ontologies for semantic fusion of multiple algorithms. Therefore ADPEPT meets the high throughput processing requirements of AWAKE applications.

## IV. AWAKE Components and Architecture

### A. AWAKE Components

The architecture of AWAKE is shown in Figure **1**. *Ingest* converts the document to a standard xml form identifying text and metadata. For information from speech, ingest includes the BBN BYBLOS speech recognition system [1]. *Deduplication*

[1] http://www.nist.gov/tac/publications/2013/presentations/KBP2013_SF_overview.presentation.pptx

checks that a new document is not largely the same as one that has been processed before.

*BBN SERIF* is a state-of-the-art information extraction engine [2]. It extracts seven kinds of entities (persons, organizations, geo-political entities, locations, facilities, vehicles, and weapons), quantities (dates, times, monetary amounts, percentages, etc.), and about two dozen types of relations (facts) among entities based on the ACE ontology[2].

Whereas SERIF maps text to entities and relations of interest in an ontology, a second document-level extraction engine, the University of Washington's *ReVerb* system [3], maps text to triples of word sequences. For instance, "He is seeking a military victory" produces a triple *<He, released to, his supporters>*. The *SERIF-ReVerb Aligner* then associates ReVerb subject and object strings with entities in the knowledge base, where possible; in the example above correctly associating "He" with the "Ahmed Abdi Godane" mentioned elsewhere in the document.

Next, the *Uploader* loads the (document-level) information discovered by SERIF and ReVerb into the database. *Entity Discovery* adds new entities to the database including alternate spellings, etc., where AWAKE estimates that the confidence is high enough. For instance, two new names might have been seen in the corpus: "Rami Makhluf" and "Rami Makhlouf". Based on a combination of spelling and context, the system will decide that these two names represent the same real-world person and add a new entry to the database to represent him.

*Profile Generation* fuses document-level information to create knowledge base facts and assigns corpus-level confidences to knowledge base facts, creating a "profile" of facts about each individual or organization. The primary goal of this component is to deal with the noise in the facts that are extracted by the upstream systems by assigning confidences from a holistic corpus-level view. For instance, this component will make the decision based on the distribution of all facts in the corpus that Vladimir Putin is extremely likely to be the president of Russia and extremely unlikely to be the president of Syria—even if an extraction error has led to the latter conclusion in a particular document in the corpus.

*Knowledge Base Fusion* augments those facts derived from language sources (text and speech) with facts as represented in a pre-existing outside knowledge base. For the proof-of-concept experiment described here, we use two outside sources to provide additional information about persons and organizations: DBpedia[3] and the CIA World Leaders Lists[4].

The *Graphical User Interface* supports five views into the knowledge base: (1) An overview shows the set of persons and organizations that a user is currently researching. (2) A graph shows all connections between a focus person/organization and others in the database. The connections can be filtered by connection type and graph size. (3) An updates view shows the most recent facts added to the database in reverse chronological order, showing either all facts or just the facts relevant to a particular person or organization. (4) A summary provides information known about a particular person or organization. (5) If the justification for a fact is requested, source documents are displayed, highlighting the sentence/paragraph justifying the fact and its context.

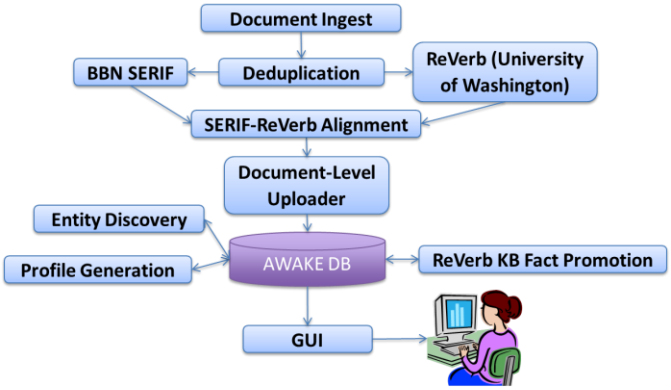


Figure 1: Architecture of the AWAKE System

## B. *AWAKE Software Architecture*

One of the goals of the AWAKE system is to allow for seamless integration of new human language technology components as desired. For this reason, we have developed a scalable, multi-layered, plug-n-play framework named *ADEPT* with capabilities that enable rapid integration of information extraction and natural language processing (NLP) algorithms. We have specifically designed the AWAKE database and architecture to support this generic, flexible representation of extracted information.

The ADEPT framework incorporates standardized definitions for core NLP data structures such as Token, Chunk, Entity, Relation, Mention, etc. and provides a uniform catalog of algorithm interfaces that facilitate semantic interoperability, algorithm fusion and Big Data processing in a scalable, parallel computing environment.

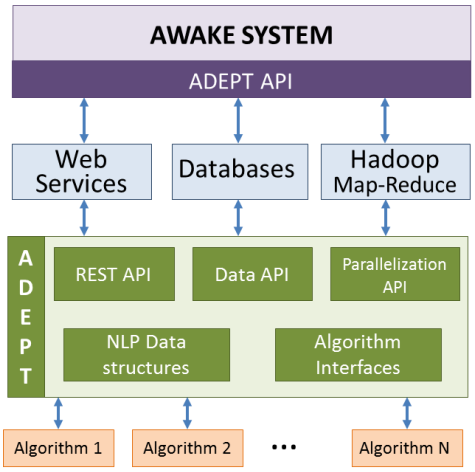


Figure 2: Multi-layered architecture with "RESTful" services enabling a robust, adaptable, distributed system

[2]http://www.itl.nist.gov/iad/mig//tests/ace/ace05/doc/ace05eval_official_results_20060110.htm

[3] http://dbpedia.org

[4] https://www.cia.gov/library/publications/world-leaders-1/

The ADEPT architecture is illustrated in Figure 2. The core elements of the architecture are NLP data structures and algorithm interfaces that enable access to one or more algorithms via multiple access points; either as RESTful web services via the REST API, or as Apache Hadoop Map-Reduce components or as serialized data objects via the Data APIs. The current implementation of the AWAKE system uses the latter. This multi-layered design abstracts the internals of the algorithms from applications such as AWAKE. The algorithm interfaces in ADEPT are designed exhaustively to support sentence-level, document-level, and corpus-level data processing, such that algorithms that fall in any of these categories can be easily plugged into the framework.

Moreover, the ADEPT data structures (and their representation in the AWAKE database) allow for not just a particular system's single-best interpretation of a document but also for each algorithm to propose multiple "n-best" interpretations if desired. For instance, all "facts" added to the database are associated with confidence measures, allowing a system to produce facts of varying quality—even those that contradict each other and even if it is not confident enough to display them to a user; these can still be helpful to the downstream knowledge fusion process. As another example, each document entity can be linked to a set of database entities (rather than just one), again with confidence measures. This also provides valuable information to downstream processes, which may be better positioned to resolve such ambiguity than a document-level extraction engine.

Beyond the data structures themselves, the ADEPT Data API standardizes algorithm access to databases and knowledge bases. Figure 3 shows the ADEPT data layer. The purpose of this is three-fold; it 1) acts as middleware between the ADEPT API and databases, 2) homogenizes access across various types of databases, such as, relational, triple stores, and No SQL data bases, and 3) supports both content from pre-built knowledge bases and content produced by algorithms. While the present AWAKE system uses a relational database, the ADEPT framework enables incorporating Graph or Key-Value databases to support Knowledge Base representation and inference.

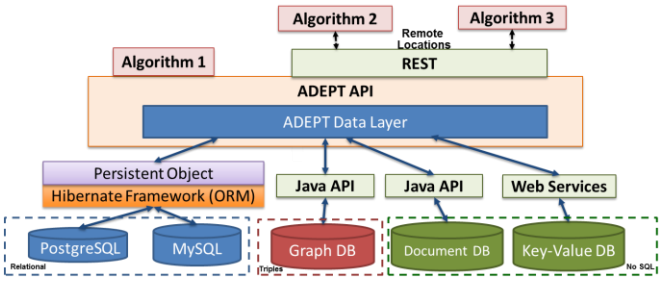


Figure 3: ADEPT Data Layer supporting multiple database access to algorithms via Data APIs

Since different NLP algorithms can employ many different ontologies (e.g. those represented by ACE, TAC-KBP or Freebase) for representing Entity and Relation types, the ADEPT framework also incorporates the Web Ontology language (OWL) to consistently represent inherent hierarchies and sibling relations across multiple ontologies. Figure 4 shows how a part of a hierarchical ontology, such as TAC-KBP, can be represented in terms of various ontology classes and properties using OWL. The solid lines represent hierarchical relationships while the dotted lines represent the properties associated with the classes.

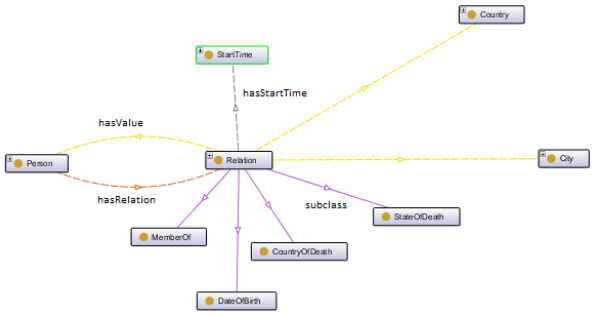


Figure 4: TAC-KBP Ontology represented in OWL format

Finally, algorithms are often interdependent and perform certain preprocessing steps on input documents that are redundant across algorithms. For this proof of concept employing only a few separate NLP algorithms, we used a simple form of distributed computing to manage the interdependencies among components. However, to process large quantities of data efficiently, the ADEPT framework additionally provides a robust processing pipeline that enables, (1) adding new interdependent algorithms, (2) reusing output from algorithms before them in the pipeline, (3) parallelizing independent algorithms, and finally, (4) logical fusion of output. Through this framework, the processing pipeline can be spawned either as distributed web services components communicating via RESTful interfaces, or as Map-Reduce jobs on an Apache Hadoop parallel computing infrastructure as shown in Figure 5.

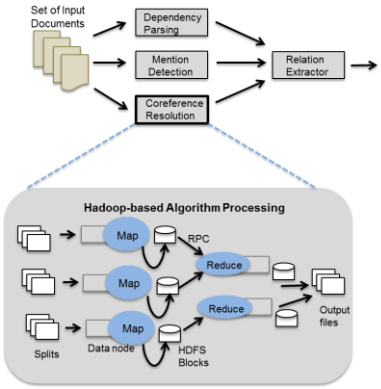


Figure 5: Hadoop-based pipeline shows the parallel processing of independent algorithms on a set of documents, with dependent algorithms running serially on the output from previous step in the pipeline

## V. PROOF OF CONCEPT

The SERIF information extraction engine recently processed roughly 600 million documents to provide document-level analysis of the English documents in the Text Retrieval Conference (TREC) Knowledge Base Acceleration

track [5]. However, for our proof-of-concept experiment in knowledge base construction (as opposed to uncorrelated document-level facts) we chose to process a smaller corpus of 725,000 English newswire documents from August, 2013 through April, 2014. We supplement this text-corpus with automatic transcriptions of 1.7K hours of broadcast news. The resulting knowledge base included roughly 550K persons, 275K organizations, 365K locations, 2.3 million triples from ReVerb, and 1.6 million corpus-level facts from SERIF. To explore the fusion of information across multiple modalities, we augmented this knowledge base with 120K facts from DBPedia and 6500 facts from the World Leaders Lists [6]. Further discussion of the distribution and overlap among these facts is included in section VI.F below.

## VI. Meeting the Challenges

### A. Deduplication

Several components in the AWAKE pipeline rely on an analysis of the distribution of entities and facts across the corpus. To take the simplest example, a fact that is frequently repeated across the corpus is more likely to be considered correct by the system. However, the conclusions drawn from this kind of analysis can be deeply unreliable when the corpus contains multiple copies of the same document (or sentence), since that repetition no longer represents multiple sources of validation for a particular fact, but rather a single source that happened to be collected multiple times.

When dealing with large corpora, comparing each document to each other is often out of scope, unless one requires an exact match (which enables various shortcuts). However, exact match is rarely sufficient to detect most duplication in the sources we have examined: it is much more likely that a repeat of an article will differ by a sentence from the original than that it will be an exact duplicate.

To reduce the scope of the problem, the AWAKE system considers only duplicates within a seven-day sliding window based on document publication or broadcast date. The de-duplication itself is performed by comparing the percentage of trigram word overlap between two documents; when one document has more than 80% of its trigrams covered by another, it is considered a duplicate and is discarded. Future work beyond this proof of concept could involve treating sub-regions of documents as duplicates instead, so as to preserve the non-duplicated portion of a document that is currently discarded. However, since most downstream processing requires the treatment of documents as a coherent whole, this would involve changes to multiple stages of the processing pipeline.

Among the 725,000 news documents, approximately 50,000 were found as duplicates by this process. Approximately three-quarters of these were found as duplicates of a document published on the same day, while 94% were found as duplicates of documents published within three days. When we expanded the sliding window to fourteen days (from seven), we detected only 7% more duplicates, with a significant cost in increased processing time. We therefore used the seven-day window for the rest of our experiments. For purposes of this proof-of concept, deduplication was performed only within a source type, anecdotal evidence from review of system output suggests that overlap also occurs between text and broadcast news sources, for example a story being published on a new-channels website as well as read during the evening news.

### B. Name Variation & Entity Disambiguation

The name variation and entity disambiguation challenge is a significant current area of research in the field of natural language processing and knowledge base curation. The AWAKE system addresses this challenge in several ways.

First, the AWAKE knowledge base is seeded with entities from existing structured knowledge sources. For this proof of concept, we began with a seed database based on a snapshot of DBpedia (itself a crowd-sourced effort to extract structured information from Wikipedia) and the geonames gazetteer [7]. From DBpedia, we created seed entries for 510K persons and 230K organizations; from geonames we extracted 365K named geographical locations around the world. Both source databases are imperfect, despite being constructed manually. In places the DBpedia ontologies are inconsistent and incomplete; properties sometimes violate their ontological definitions. Even conclusively determining a list of persons and organizations is sometimes challenging (for instance, the entry for "*Illinois Senate career of Barack Obama*" is tagged as type Person in DBpedia, and "*Al Qaeda*" is typed as a Country).

Nevertheless, both DBpedia and the geonames database are an incredibly valuable source of information about persons, organizations, and places; they prove helpful in various ways for the name variation and entity disambiguation problem. First, they contain already within themselves some information about valid name variations for each entity. Second, the existence of multiple seed database entries with the same name is a good (though not perfect) indication of the need for disambiguation when encountering said name in the corpus. Finally, they contain factual information about their entries that can be used to help disambiguate among similarly named entities (e.g. that person A is affiliated with Russia and person B is affiliated with Jamaica).

After initializing a seed database, the next step in the AWAKE pipeline is to match each named entity found in each document to this database, where possible. SERIF performs this process, producing a set of possible database entity matches with confidence measures for each document-level named entity. The system will also produce its estimation of the likelihood that this named entity is actually not yet a part of the seed database.

The first source of information used by SERIF is the name string itself. This primarily addresses the core name variation problem and relies on a combination of several algorithms that examine edit distance between names (both at the character and the token level). It also draws on a variety of additional

[5] http://trec-kba.org/trec-kba-2014/

[6] https://www.cia.gov/library/publications/world-leaders-1/

[7] http://www.geonames.org/

external sources (e.g. Wikipedia redirect links) to expand its set of possible name variations for each entity in the database.

Where possible, SERIF also leverages additional analyses that it has produced about each document, for example facts extracted about a named entity in the local context (e.g. *this person is a Russian*), and other named entities mentioned in the same neighborhood (e.g. *this person is mentioned near Vladimir Putin and Dmitry Medvedev*). Given the facts already associated with existing database entries, e.g. via DBpedia, this information can be used to improve entity disambiguation. For instance, the system might know that database entity #3 and database entity #72 are both reasonable candidates for this document's named entity based on the name string used, but which is actually more likely? When identifying possible matches for geographical locations (where name collision is very frequent), this type of additional analysis is particularly useful; a document author is less likely to suddenly refer to a small town in Oklahoma when the rest of the document is discussing persons and places in the Ukraine.

Naturally, the two challenges of name variation and entity disambiguation are not distinct: entity disambiguation is a challenge when two entities share the exact same name, but it is just as likely that the sources of information used to distinguish between two "John Smith"s are the same sources that the system will use to decide whether "Bob Doll" and "Bob Dole" are valid name variations for each other or whether they more likely represent two separate people (despite differing by only one character).

### C. *Entity Discovery*

Even with a large seed database, there will be a wide array of novel entities that appear in any corpus, especially a corpus made up of real-time streaming data. In our experiments, we found that only about 25% of named persons and organizations in our corpus were judged by the AWAKE system to map to a seed database entry with reasonable confidence.

The goal of the *Entity Discovery* component is to identify new entities that can be confidently added to the AWAKE database. There is an obvious trade-off of coverage and accuracy: the more aggressive this component is, the more likely it is to add a new entry for an entity that really should have been linked to an existing entry instead (for instance, adding a new entry for an entity named "*Islamic Resistance Movement*" when this really should have been an alternate name for the existing entry for Hamas). For this reason the Entity Discovery component exposes several parameters to users which allow them to directly tune its level of aggressiveness.

The goal of the Entity Discovery component is related to that of the name variation and entity disambiguation component described above, but rather than being given a single named entity and asked which of the existing database entries it best belongs to, it is given a set of document-level named entities and asked to cluster them together. To perform this task it relies on many of the same algorithms and sources of information described in the section above. For instance, name variation detection algorithms are crucial in clustering together two different spellings of a novel entity's name, and analysis of the local context for each name also provides information about which entities are more likely to refer to the same novel real-world entity.

At the end of its clustering process, the Entity Discovery component adds all of the resulting novel entities to the database and links those novel entities to the document entities from which they were derived.

In our proof-of-concept study, approximately 39k unique persons and 37k unique organizations from the seed database were judged by AWAKE to be mentioned in our corpus. Some were mentioned thousands of times, e.g. *Hezbollah* (mentioned more than 10,000 times); some were very rare, e.g. *Bahir Dar University* (mentioned exactly once). The Entity Discovery system then considered all unmatched named entities whose name strings occurred at least five times in the corpus. From these, it identified approximately 40k new persons and 23k new organizations to add to the entity database.

### D. *Mapping text to knowledge*

BBN's general language understanding component SERIF [Ram11] has achieved state-of-the-art performance in applications such as information extraction and machine reading; it processes English, Arabic, Chinese, and Spanish text. It includes four dimensions to interpreting text: syntactic parsing, creating a simple propositional logical form called text graphs, understanding coreference (e.g., pronouns or definite descriptions), and disambiguation in context.

The semantic interpretation of a sentence is a text graph abstracted from the sentence's exact syntactic realization. All intermediate nodes in a text graph represent propositions; all leaves represent arguments of propositions that cannot be expanded. The labels on the arcs between nodes are argument labels. Text graphs are enhanced with coreference, e.g., if the phrase "his brother" is mentioned by name earlier in the document, the name is associated with that name.

These automatically-generated and automatically-augmented text graphs are used to improve the system's ability to map diverse textual representations of a piece of information to a single fact in a knowledge base. As an example, the text graph shown in **Figure 6** would be part of the semantic interpretation of each of the following: "*John accused Mary," John, a friend of Sheila, accused Mary", Mary was accused by John," John accused his British friend, Mary,"* and *John was sorry to accuse his friend Mary."* All of these constructions should map to the core knowledge-base fact that "John" accused "Mary" assuming that a target ontology includes Accusation as a fact type. At the same time, none of the following sentences would generate that fragment of a text graph, despite using similar words, so none of the following would produce that (in this case incorrect) Accusation fact: "*Bob, a friend of John, accused Mary," "Bob forgave John but accused Mary," "Mary was accused by critics of copying John,"* and *"John accused someone other than Mary."*

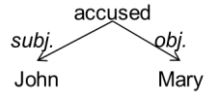

Figure 6: A simple text graph without nested structures

As described above, we supplement the relations extracted by SERIF with ReVerb's openIE extraction of facts. AWAKE uses the ADEPT interfaces to align ReVerb's triples with SERIF's mentions. AWAKE maintains those facts for which at least one argument is aligned with an actor in the AWAKE KB. To represent the open-set of predicates produced by ReVerb, the fact is uploaded with a type of *open* with three arguments: one for each argument extracted by ReVerb and one for the predicate. In contrast facts from SERIF are given one of the types listed in **Table 1** and typically have two arguments.

### *E. Erroneous Extractions*

As discussed above, the document-level information provided to the corpus-level knowledge fusion component is noticeably noisy. This is particularly true since the infrastructure specifically allows for document-level algorithms to produce facts or entity links with low confidence. The goal of the corpus-level knowledge fusion component (*Profile Generator*) is to mitigate the effect of this noise—to make use of all the low- and high-confidence extractions to produce a single coherent set of facts about each entity (each still with its own corpus-level confidence). For instance, the document-level facts provided to Profile Generator for Hamid Karzai include not just a great number of facts supporting his position as the president of Afghanistan but also a smaller number supporting his position as the president of Tajikistan and also of the United States. However, one cannot simply discard all but the most frequent fact for any category, since a person can have multiple positions in multiple organizations, some of which are more frequently mentioned than others; a more nuanced line of attack is required.

The primary approach of Profile Generator is to combine a number of different sorts of confidence and sources of information for each fact or set of facts. These confidences include (1) the document extractor's confidence about the fact, (2) the document extractor's confidence about the link between the entity-mentions and the document entities involved, (3) the document extractor's confidence about the link between the document entities and the knowledge base actors involved, (4) the frequency of the fact, (5) the system's confidence about the source of the information, to allow a reduction in the system's belief in facts that may be the result of a cascade of errors and (6) the fact's consistency with other facts proposed at the corpus level (what else do we know about Hamid Karzai? what else do we know about who the president of Tajikistan might be?).

With respect to the last source of information, Profile Generator does incorporate soft (non-binding) relation-specific constraints to enforce consistency, for instance, a person is unlikely to have more than two spouses or more than a few organizational affiliations. ProfileGenerator also allows for various forms of "fuzzy" match when fusing facts, for instance allowing a fact of the form (ACME, Smith, leader) to match one of the form (ACME, Smith, president).

The end goal of the profile generation stage is a set of corpus-level facts for each person or organization entity in the knowledge base, each with a corpus-level confidence assigned by this component.

### *F. Fusion with External Sources*

We have already described how our approach to entity linking fuses textual sources with an external database (here DBpedia). We capitalize on these links to further augment our knowledge about the entities in our knowledge base. For those entities for which we have identified an outside link, we import supplemental information. This importation requires an additional set of deduplication and conflict resolution. SERIF's ontology is not completely aligned with DBpedia's. Furthermore, while the fact that DBpedia is a structured source could suggest that its accuracy is quite high, it contains significant amounts of noise in the form of contradictions in the upper levels of the ontology as well as errors in a manually created resource. For instance, if a person is in relationship to France, "France" might sometimes be represented as a DBpedia country, or sometimes just as an unnormalized string. Or, as discussed above, "*Al Qaeda*" is typed as a Country and "*Illinois Senate career of Barack Obama*" as a Person.

Name variation is also a significant challenge. For instance, the first country in the World Leaders lists is Afghanistan. Looking at the first three leaders after the president on this list, all three had different spellings in the World Leaders lists compared to DBpedia: Mohammad Fahim Khan vs. Mohammad Fahim, Abdul Karim Khalili vs. Karim Khalili, and Mohammad Asif Rahimi vs. Mohammad Asef Rahimi. There are also many name collisions, e.g. in DBpedia there exists a Tim Johnson (U.S. Senator), a Tim Johnson (Illinois politician), a Tim Johnson (film director), and a Tim Johnson (baseball).

There are also challenges in fusing relations when their definitions are related but not identical. For instance, "birth place" and "country of origin" are clearly related, but also have potential differences in both granularity ("Minnesota" vs. "United States") and in definition (an American born abroad would likely still identify "United States" as country of origin).

For this proof of concept, we rely on a fairly conservative approach, resolving across sources only when our name variation and entity disambiguation algorithms could be reasonably certain of their decisions. So, in this case, it is likely that no facts about "Tim Johnson" would be added to the knowledge base, since there is so much ambiguity involved.

**Table 1** represents the distribution of fact types across the corpus (extracted by SERIF) and across DBpedia and the World Leaders lists. Note that some fact types are targeted in only one source. For instance, the World Leaders lists only provide information about persons and their affiliation (job position with a national government) and their nationality.

Table 1: Fact Type Distribution

| Fact Type | SERIF | DBpedia | World Leaders |
|---|---|---|---|
| Quotation | 729364 | 0 | 0 |
| Related news | 206821 | 0 | 0 |
| Quotation about | 179271 | 0 | 0 |

| | | | |
|---|---|---|---|
| Description | 60845 | 0 | 0 |
| Affiliation | 22230 | 14306 | 1409 |
| Contacts | 15098 | 14760 | 0 |
| Nationality | 10135 | 166 | 1855 |
| Location of operation | 9502 | 39 | 0 |
| Headquarters | 2036 | 144 | 0 |
| Place visited | 1740 | 0 | 0 |
| Family | 1548 | 4383 | 0 |
| Spouse | 487 | 2814 | 0 |
| Founding date | 446 | 0 | 0 |
| Education | 415 | 13567 | 0 |
| Death date | 387 | 11869 | 0 |
| Birth date | 376 | 46800 | 0 |
| Founder | 355 | 0 | 0 |
| Award | 0 | 9492 | 0 |
| Creative Work | 0 | 2443 | 0 |
| Subsidiary | 0 | 1432 | 0 |
| Parent Company | 0 | 931 | 0 |
| Cause of death | 0 | 358 | 0 |

Table 2 demonstrates the complementary nature of the facts found by SERIF in the corpus and the facts imported from external sources. For each fact type targeted by both SERIF and the external-source importers, the table shows the number of unique facts found only in the corpus, the number found only in an external source, and the number found in both.

**Table 2: Overlap between SERIF and External Sources**

| FactType | SERIF only | External only | BOTH |
|---|---|---|---|
| Affiliation | 20632 | 14117 | 1598 |
| Birth date | 197 | 46621 | 179 |
| Contacts | 15082 | 14744 | 16 |
| Death date | 140 | 11622 | 247 |
| Education | 357 | 13509 | 58 |
| Family | 1344 | 4179 | 204 |
| Headquarters | 1929 | 37 | 107 |
| Location of operation | 9467 | 4 | 35 |
| Nationality | 8869 | 743 | 1266 |
| Spouse | 259 | 2586 | 228 |

Excerpts from the end result of profile generation and external-source fusion are shown in the box below. Items that are singly starred (*) come solely from an external source; items that are doubly starred (**) are supported both by an external source and SERIF's extraction from the input corpus. Unstarred facts were detected only by SERIF, from the input corpus. Information in [brackets], e.g. "*The group [Al-Shabaab]*" or "*Wednesday [2013-09-25]*"is information that has been automatically added by the system to enable a user to understand the information without reading the full source text.

**Moktar Ali Zubeyr**

Birthdate: July 10, 1977 *

Nationality: Somalia**

Affiliation: Al-Shabaab (Leader)**

**Descriptions**:

*top commander*

*the hardliner most interested in a strong relationship with international jihadists and Al Qaida*

*Al-Shabaab chief*

**Statements by**:

*Shebab chief Ahmed Abdi Godane said the Westgate mall carnage was retaliation for Kenya's military intervention in Somalia.*

*Al-Shabab leader Shaykh Muqtar Abu-Zubayr, aka Ahmad Abdi Godane, has congratulated the attackers of Nairobi's Westgate Shopping Mall, saying they carried out the attack because they were against the injustices meted out against their people*

*The group [Al-Shabaab]'s leader, Ahmed Godane Shaykh Mukhtar Abu Zubayr, warned the Kenyan public there was no way they could "withstand a war of attrition inside your own country", in a statement posted on the internet late on Wednesday [2013-09-25].*

**Statements about:**

*Robow has previously accused Godane of being an apostate,*

*Abdisamad explained that Godane is a supporter of global jihad who believes that Somalia belongs to all Muslims across the world.*

**ReVerb triples:**

*(Ahmad Godane, was previously accused of being behind, the assassination)*

*(Al-Hijra, was not sanctioned by, Godane)*

*(Godane, instituted, a purge)*

## VII. Conclusion and Future Work

In this paper we reported a pilot study in automatic reading of text to build a knowledge base designed to free researchers from reading masses of documents to compile facts so that those researchers can focus primarily on analysis and synthesis. To do so, we have provided proof-of-concept solutions to several challenges including: deduplication, name variation & entity disambiguation, entity discovery, mapping text to knowledge, and coping with erroneous extractions. We have demonstrated multiple types of fusion: (a) integrating facts

extracted from the modalities text, speech, and structured data and (b) integrating ontology-based extraction and OpenIE.

There are two significant next steps. One is to expand the set of components that add to the knowledge base. A second is to fuse information from additional modalities, including non-English text, image processing of videos, and character recognition in videos.

### A. *Adding Additional Components*

Many other capabilities could be added. Additional relationship extraction algorithms could provide coverage for new fact types (or simply new facts in the currently targeted categories, where some were missed by the current algorithms). They could also provide new confidence information about already extracted facts to the Profile Generation / Fusion components—for instance, if three separate algorithms identify a particular fact in the corpus, it may be more likely to be correct.

Especially if alternative, less formal sources of data are incorporated (e.g. blogs, twitter) components that understand the more implicit concepts in text such as sarcasm detection [21], belief tagging [22], and opinion analysis [23] could add an additional layer of information. For example, sentiment detection could be used to distinguish between positive/negative/neutral statements. The ADEPT framework is designed to support a range of NLP components to support such extensions

### B. *Fusing Additional Modalities*

Our long-term goal is illustrated in Figure 7. Multiple modalities offer the potential of different insights, corroborating evidence, disambiguating evidence, and for combining individually inconclusive evidence that in combination could support a valid conclusion. Furthermore, evidence from a prior knowledge base and/or from other sources should improve detection and extraction from each source, since redundancy across sources improves accuracy

Yet, the challenges addressed in the pilot study reported here are magnified. The component technologies operate at different levels of maturity. As we integrate less mature technologies, the importance of using confidences derived from the corpus as a whole and characteristics of big data to resolve inconsistencies is raised. Furthermore the component technologies may have been developed targeting very different ontologies or data types: video-event detection does not necessarily operate at the same level of granularity as event-detection developed for text-processing.

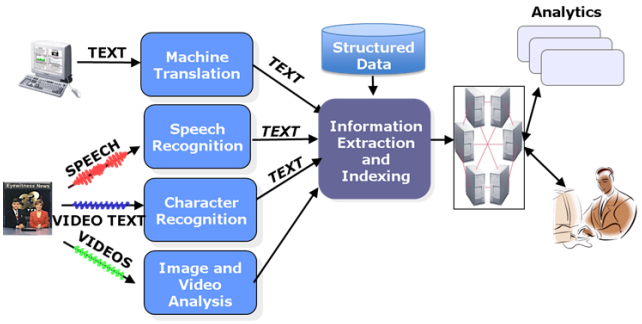


Figure 7: Vision for Fusing Multiple Modalities

## ACKNOWLEDGMENT

This paper is based upon work supported by the DARPA DEFT Program. The views expressed are those of the author(s) and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This research was developed with funding from the Defense Advanced. Approved for Public Release, Distribution Unlimited. The authors would also like to thank Alex Zamanian, Manaj Srivastava, and Connor Stokes, Artan Sameqi, Fred Choi, Jonathan Watson, Paul Martin, and Ray Tomlinson.

## REFERENCES

[1] P. Natarajan, E. Macrostie, R. Prasad and J. . "Analysis of Multimodal Natural Language Content in Broadcast Video." In Multimedia Information Extraction: Advances in Video, Audio, and Imagery Analysis for Search, Data Mining, Surveillance, and Authoring,. M. Maybury, Ed. Wile Online, 2012.

[2] L.Ramshaw, E. Boschee, M. Freedman, J. MacBride, R. Weischedel, A. Zamanian. "SERIF Language Processing — Effective Trainable Language Understanding," In Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation, J. Olive et al. Eds., pp.626-631, Springer, 2011.

[3] A. Fader, S. Soderland, and O. Etzioni. "Identifying Relations for Open Information Extraction," Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1535–1545, Edinburgh, Scotland, UK, July 27–31, 2011.

[4] J. Ellis, J. Getman, J. Mott, X. Li, K. Griffitt, S. Strassel, J. Wright . Linguistic Resources for 2013 Knowledge Base Population Evaluations, In Proceedings of the Sixth Text Analysis Conference, 2013.

[5] Y Li, Y Zhang, D Li, X Tong, J. Wang, N. Zuo, Y. Wang, W. Xu, G. Chen, J. Guo, "PRIS at Knowledge Base Population 2013," In Proceedings of the Sixth Text Analysis Conference. 2013.

[6] B. Roth, T. Barth, M. Wiegand, M. Singh, D. Klakow. "Effective Slot Filling Based on Shallow Distant Supervision Methods," In Proceedings of the Sixth Text Analysis Conference, 2013.

[7] B. Roth, Grzegorz Chrupala, Michael Wiegand, Mittul Singh, and Dietrich Klakow. Generalizing from freebase and patterns using distant supervision for slot filling. In Proceedings of the Text Analysis Conference, 2012.

[8] M. Surdeanu. Overview of the TAC2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling. In Proceedings of the Sixth Text Analysis Conference, 2013.

[9] G. Pink, A. Naoum, W. Radford, W. Cannings, J. Nothman, D. Tse, J. Curran. SYDNEY_CMCRC: English Entity Linking. In Proceedings of the Sixth Text Analysis Conference, 2013

[10] D. Yu, H. Li, T. Cassidy, Q. Li, H. Huang, Z. Chen, H. Ji, Y. Zhang, D. Roth. RPI-BLENDER TAC-KBP2013 Knowledge Base Population System. In Proceedings of the Sixth Text Analysis Conference, 2013.

[11] L. Ratinov and D. Roth and D. Downey and M. Anderson, Local and Global Algorithms for Disambiguation to Wikipedia. ACL 2011.

[12] P. McNamee, J. Mayfield, T. Finin. HLTCOE Participation at TAC 2013.I n Proceedings of the Sixth Text Analysis Conference, 2013.

[13] D. Ferrucci, A. Lally. "UIMA: an architectural approach to unstructured information processing in the corporate research environment." Natural Language Engineering 10, no. 3-4 (2004): 327-348.

[14] K. Bontcheva, H. Cunningham, D. Maynard, V. Tablan, and H. Saggion. "Developing reusable and robust language processing components for information systems using gate." In 2012 23rd International Workshop on Database and Expert Systems Applications, pp. 223-223. IEEE Computer Society, 2002.

[15] S. Bird, E. Klein, E. Loper, J. Baldridge. "Multidisciplinary instruction with the natural language toolkit." In Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics, pp. 62-70. Association for Computational Linguistics, 2008.

[16] T. Morton, J. Kottmann, J. Baldridge, and G. Bierner. "Opennlp: A java-based nlp toolkit." 2005.

[17] M. Henning. "A new approach to object-oriented middleware." Internet Computing, IEEE 8, no. 1 (2004): 66-75.

[18] M. Slee, A. Agarwal, M. Kwiatkowski. "Thrift: Scalable cross-language services implementation." Facebook White Paper 5 2007.

[19] E. Newcomer. Understanding Web Services: XML, Wsdl, Soap, and UDDI. Addison-Wesley Professional, 2002.

[20] Marjorie Freedman, Lance A. Ramshaw, Elizabeth Boschee, Ryan Gabbard, Gary Kratkiewicz, Nicolas Ward, Ralph M. Weischedel: Extreme Extraction - Machine Reading in a Week. EMNLP 2011: 1437-1446.

[21] R. Gonzalez-Ibanez, S. Muresan, N. Wacholder. "Identifying Sarcasm in Twitter: A Closer Look". Proceedings of ACL-HTL 2011 (short paper). 2011.

[22] M. Diab, L. Levin, T. Mitamura, O. Rambow, V. Prabhakaran, and W. Guo. "Committed belief annotation and tagging. In Proceedings of the Third Linguistic Annotation Workshop" (ACL-IJCNLP '09). Association for Computational Linguistics, Stroudsburg, PA, USA, 68-73. 2009.

[23] Y. Choi, E. Breck, and C. Cardie. "Joint extraction of entities and relations for opinion recognition". In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06). Association for Computational Linguistics, Stroudsburg, PA, USA, 431-439. 2006.

.