Automatic Information Extraction from Documents: A Tool for Intelligence and Law Enforcement Analysts

Richard Lee

Sterling Software
1650 Tysons Blvd. #800
McLean VA 22102
rlee@mclean.sterling.com

Abstract

We have developed a number of systems to aid analysts in tracking the activities and associations of various criminal elements. These systems each consist of a relational data base containing the details on the relevant entities, activities, and associations, assorted tools for retrieving and analyzing those details, including link analysis, and a tool for automatically extracting those details from documents.

Introduction

On several projects, Sterling Software has been leading the development of software systems for the intelligence analyst. These systems have been designed to assist the analyst in discovering and keeping track of individuals and organizations involved in various activities such as drug trafficking, terrorism, insurgency, weapons proliferation, etc. They also provide the ability to tie these individuals and organizations to the facilities, vehicles, etc, they use, and to the events and activities they are involved in.

Each system is a collection of tools integrated around a relational data base. The tools provide a variety of ways to search the data base and to view selected data from it; they include both a link display tool and a geographic display tool.

Each system has a facility for storing incoming documents (*messages*) containing raw, free-text data. One key tool uses Natural Language Understanding (NLU) technology to automatically extract all the above-listed types of information from those messages and store that information appropriately in the data base.

The Data

As noted, a relational data base forms the core of each of these systems. In that data base, all the pertinent information about each of the *items* of interest is stored. These items consist of *entities*, *events*, and *associations* between the entities and events. In the later systems, the messages are also stored directly in that same data base. Typically, the *entities* of interest include:

- Individuals
- Organizations (government, commercial, military, extra-legal, etc)
- Places (anything from street addresses or coordinates to continents)
- Facilities (factories, airports, hotels, warehouses, etc)
- Documents (passports, driver's licenses, bank books, etc)
- Money
- Vehicles (air, land, or sea)
- Drugs
- Weapons

For each entity, the data base can contain the appropriate detailed information. The exact nature of the details depends of course on the type of entity. For an individual, for example, the data base typically includes name, aliases, gender, marital status, citizenship, race, date of birth, occupation, hair color, eye color, height, and weight.

This set of entities is largely independent of a particular analytical shop's area of interest. The *events*, however, tend to be more closely tied to that area. For Counter-Drug analysts, the events include:

- Processing, purchasing, transporting, etc of drugs
- Planning, meeting, or communicating about any of the above
- Arrest of traffickers or seizure of drugs, money, weapons, etc

For Counter-Terrorism analysts, on the other hand, the events include:

- Killing, kidnapping, hostage-taking, etc of people
- Bombing, hijacking, etc of buildings and vehicles
- Buying, stealing, etc of weapons and money
- Training in weapons and tactics
- Arrest, conviction, punishment etc of terrorists

- Other government actions known to provoke terrorist retaliation

The *associations* can be grouped into three categories:

- Entity-to-entity
- Entity-to-event
- Event-to-event

Entity-to-entity associations indicate *relations*, the nature of which depend on the types of entities involved. Typical relations between two individuals might include blood relative, spouse, employee, neighbor, etc. Typical relations between an individual and an organization might include owner, head, employee, etc. Typical relations between an individual or organization and a vehicle might include owner and user. Typical relations between an individual and a place might include place of birth, residence, current location, etc.

Entity-to-event associations indicate the *roles* the entities play in the event. For example, a bombing event would typically have roles for bomber, victim, object, weapon, place, etc., each of which could be filled by one of more entities of the appropriate type(s). Each event would also have "roles" for any date and time information.

Event-to-event associations include tying each communicating, meeting, planning (etc) event to the events discussed.

When the messages are stored in the same data base, one special type of association ties each detailed item record to the message(s) the information came from.

The Extraction

On each of these projects, we have developed (or are developing) an Information Extraction (IE) tool to run as a background process, for near-real-time extraction of information from incoming messages. The IE processes a single message at a time, finding all the kinds of information described above and generating the appropriate data base entries containing that information.

The IE operates by first looking for phrases containing all the references to the entities of interest, plus any date and time references. It then analyzes the phrases and clauses containing those references to find all the entity-to-entity associations (relations). It then analyzes the clauses for events of interest, assigning each entity reference, date and time in the clause to the appropriate role in the event. For each item found, it constructs a *frame* – a representation which categorizes each piece of pertinent information by putting it into the appropriate *slot*.

Once all the item (entity, event, association) references have been thus identified and represented as frames, those that appear to actually refer to the same item are merged, with the result that (ideally) exactly one frame is left for each distinct extracted item. These frames are then translated into data base records. Typically, each frame is mapped to a single data base record, with each slot mapped to a data base field, but it is often more complicated. Each item in an association record is of course stored in the data base using the key for the corresponding item record.

Technical Approach

The IE's analysis of text is accomplished by applying large collections of patterns. The patterns for recognizing entity references use a combination of internal structures and external contextual clues, each involving both syntactic and semantic information about words and phrases. For example, the phrase "Analyst John Q. Jacobson of Marigold Financial Corp" contains both internal and external indicators for both the individual and the organization. Each reference thus recognized is *reduced* – replaced by a single *token* – with the pieces of recognized information (name, title, occupation, etc) distributed appropriately into slots in the attached frame.

After the entity analysis is done, the relation analysis uses patterns which look for text containing the right combinations of relation-specifying words and phrases, entity tokens, and other bits of text. Simple examples would be "John's wife Mary", "John, an analyst for Foobar Inc.", "members of the Garcia Group Larry, Moe, and Curly", "an aircraft belonging to John", etc. Certain clauses involving state verbs are also recognized as expressing relations; for example, "John owns a Cessna", "John was born in Cleveland", "Boston is the headquarters of Foobar Inc.", etc. Each recognized chunk of text is again reduced, to a token representing the "head" entity, with a frame representing the relation information attached.

Finally, event analysis uses clause-level patterns which look for appropriate combinations of verbs, prepositions, etc, along with the entity (plus data and time) tokens, to recognize references to events of interest. These patterns also indicate how the entities should fill the roles of the events; the analysis constructs the event frames accordingly, with a slot for each role.

Each of these three stages actually involves multiple steps of pattern application and reduction. This is because

- Entity references, particularly places, can be nested inside others ("Japan National Railways" "U.S. dollars", "Miami International Airport")
- Entity references can provide strong contextual clues for others ("Suitomo Bank's Tomiko Hashimoto" "a hospital in Kasumigaseki")
- Relation phrases can be nested ("John Doe, of Boston-based Foobar Corp")

Examples

Here is a very brief example message body fragment, to demonstrate the processing steps and show the wealth of information that is extracted by an IE tool.

PILOT PABLO GARCIA, COLOMBIAN, PPT 2324224, ARRIVED AT MIAMI INTERNATIONAL AIRPORT ON 27 JUN 97. HE WAS ARRESTED BY U.S. CUSTOMS AGENTS WHEN 300 KGS OF COCAINE WAS FOUND IN THE SPARE FUEL TANK OF HIS CESSNA FIREBAT.

The initial processing stages reduce the entity, date, and time references, as well as key verb constructs, into tokens with attached frames. This actually involves recognizing place references ("MIAMI", "U.S.") inside larger entity references; the requisite relation frames are also constructed as part of reducing the larger reference. The result (not showing the contents of the frames) can be represented thus:

<INDIVIDUAL> <ARRIVE> AT <FACILITY> ON <DATE>. <INDIVIDUAL> <ARREST-pas> BY <ORGANIZATION> WHEN <DRUGS> <FIND-pas> IN THE SPARE FUEL TANK OF <INDIVIDUAL> 's <VEHICLE>.

In this example, there is only one additional relation found, reducing the last three tokens to just <VEHICLE>.

The event processing finds a Movement event in the first sentence, and an Arrest event plus an (implied) Seizure event in the second.

The merging of frames then decides that the three <INDIVIDUAL> frames all refer to the same entity, which requires doing the pronominal reference resolution.

At this point, the frames would look something like this:

INDIVIDUAL:
  NAME: PABLO GARCIA
  OCCUPATION: PILOT
  CITIZENSHIP: COLOMBIA

DOCUMENT:
  TYPE: PASSPORT
  NUMBER: 2324224

PLACE:
  CITY: MIAMI
  COUNTRY: US

FACILITY:
  TYPE: AIRPORT
  NAME: MIAMI INTERNATIONAL AIRPORT

DATE:
  YEAR: 1997
  MONTH: 06
  DATE: 27

ORGANIZATION:
  NAME: CUSTOMS
  TYPE: LEA
  COUNTY: US

DRUGS:
  TYPE: COCAINE
  QUANTITY: 300
  UNIT: KG

VEHICLE:
  TYPE: AIR
  MANUFACTURER: CESSNA
  MODELNAME: FIREBAT
  MODELNUMBER: XJ3

RELATION:
  TYPE: HASDOC
  ENT1: <INDIVIDUAL>
  ENT2: <DOCUMENT>

RELATION:
  TYPE: LOCATED
  ENT1: <FACILITY>
  ENT2: <PLACE>

RELATION:
  TYPE: OWNS
  ENT1: <INDIVIDUAL>
  ENT2: <VEHICLE>

EVENT:
  TYPE: MOVEMENT
  AGENT: <INDIVIDUAL>
  DESTINATION: <PLACE>
  DESTFAC: <FACILITY>
  ENDDATE: <DATE>

EVENT:
  TYPE: ARREST
  ARRESTEE: <INDIVIDUAL>
  ARRESTER: <ORGANIZATION>
  PLACE: <PLACE>
  FACILITY: <FACILITY>
  BEGINDATE: <DATE>
  ENDDATE: <DATE>

EVENT:
  TYPE: SEIZURE
  SEIZEE: <DRUGS>
  SEIZER: <ORGANIZATION>
  PLACE: <PLACE>
  FACILITY: <FACILITY>
  BEGINDATE: <DATE>
  ENDDATE: <DATE>

Notice that while realizing the implied seizure is not difficult for IEC, the cross-sentence reasoning required to realize that the Movement event was actually a Transshipment involving the drugs and the aircraft is currently beyond its capabilities.

More Examples

Here are a few more examples of relation phrases, to demonstrate the variety and nesting complexity that is handled by the IEC:

1. ANALYST JOHN DOE OF XYZ CORP
2. JOHN DOE, WHO IS A MEMBER OF THE GARCIA-CARTAGENA GROUP,
3. JOHN'S WIFE MARY, A SECRETARY AT XYZ CORP,
4. THE GUZMAN ORGANIZATION, LOCATED IN MEDELLIN,
5. A CESSNA OWNED BY BOSTON-BASED XYZ CORP

These would result in frames (and then data base records) for the obvious entities, and for relations:

1. Employee (individual – organization)
2. Employee (individual – organization)
3. Spouse (individual – individual); Employee (individual – organization)
4. Location (organization – place)
5. Location (organization – place); Owner (organization – vehicle)

... respectively. (Obviously, the relation types are not cast in stone. If the analysts' needs and the data base design warrant it, the distinction between "member" and "employee", or the more specific "analyst" and "secretary", can be maintained by the IEC.)

Note that the result of reducing phrases such as the above is a token representing the head of the phrase, which is still available for use in event patterns. For example, if the larger text of the 3rd fragment was "JOHN'S WIFE MARY, A SECRETARY AT XYZ CORP, WAS ARRESTED FRIDAY FOR ..." then, in addition to the relation frames described above an Arrest Event frame with the Individual frame for "Mary" filling the Arrestee role would be produced.

Example Systems

CDIS

The first system Sterling Software developed an IE capability for is called the Counter-Drug Intelligence System (CDIS), which was developed to support analysts tracking the entire spectrum of narcotics-related activities from crop cultivation and precursor chemical production to delivery of the drugs into the United States. CDIS was designed around a Sybase relational data base which stores detailed information on all the entity types listed above, a dozen narcotics-related event types as outlined above, and the full assortment of relations and roles.

We developed and integrated into CDIS an IE tool called the Automated Templating System (ATS) which extracted all the relevant information from the free-text portion of four different types of intelligence messages. On a test of entity extraction skills, the ATS scored much better than several counter-drug analysts on both speed and accuracy.

Alta Analytics integrated their commercial link analysis tool, Netmap, into CDIS. That tool displays myriads of entities, events, and associations placed in the data base by analysts and ATS.

MUC6

The second system was developed for the DARPA-sponsored Sixth Message Understanding Conference (MUC-6). It extracted and templated information on Individuals, Organizations, Locations, Money, Dates, and Times, from Wall Street Journal articles. Its accuracy on those tasks compared very well with those of other participants.

MDITDS

The third system we developed an IE tool for was the Migration Defense Intelligence Threat Data System (MDITDS). This system is being built around a Memex data base – not, strictly speaking, a relational data base engine, but the design has separate tables for the primary entity types, a table for events, and the crucial Association table.

The Information Extraction Component (IEC) was developed to extract the usual details on all the types of entities listed above, the various event types relevant to Counter-Terrorism analysts, and the usual assortment of relations and roles. We expect that the IEC will be delivered shortly, with the knowledge bases needed for extracting from newswire articles; work on knowledge bases for intelligence messages may be continued at a future time.

The link analysis tool for MDITDS is being developed by Orion Scientific.

Unnamed

The fourth system, as yet unnamed, is in the early stages of design and prototype. It will be built around an Oracle relational data base, the design of which will be similar to that of the data bases for CDIS and MDITDS and will reflect the lessons learned on those projects.

The IE tool development will be a continuation of the work begun under MDITDS; that continuation will consist primarily of:

- Reworking of the code for greater modularity and portability
- Knowledge base development for intelligence messages
- Knowledge base and code development for additional event types
- Work on improving the handling of coreference

Conclusions for Link Analysis

Integrating an information extraction tool and a link analysis tool with a relational data base is a natural and proven way to provide the latter with plenty of usefully structured data to work with. The association information automatically placed in the data base by the former tool is the natural source of link data to drive that analysis.

The bottom-up nature of our approach to information extraction means that entity extraction happens first, without regard to event extraction. Since the set of entity types varies little across problem domains, it is relatively easy to port that portion of an IE from one application to another, as long as the style of documents being processed is similar.

One limitation of the IE tool is that it makes no effort to decide whether an item it has found in a message is the "same" as another item already in the data base; it errs on the side of assuming it is not. Thus, each message that refers to a "Pablo Garcia", or a "Cessna Firebat", for example, will result in a new record for that item being generated. It is up to the analyst to use various tools – including link analysis – to decide that two records refer to the same item. This leads to a useful synergy between the two tools and the analyst.

We have demonstrated that the state of the art in Artificial Intelligence – specifically, Natural Language Understanding – is advanced enough that we can implement a practical Information Extraction tool which populates relational data bases with detailed information from free-text messages. This information includes not just entities but events and associations. The information is structured in a way which enables useful analysis by means of link analysis and other tools.

Acknowledgments

With the exception of the few weeks work on MUC-6, all work described herein was carried out on various US DoD contracts. This is published with the sponsoring agencies' permissions.

All versions of the IE tool were developed using the NLToolset, which is a product of Lockheed-Martin.

References

Osterholtz, L.; Lee, R.; and McNeilly, C. 1995. The Automated Templating System for Database Update from Unformatted Message Traffic. In *1995 ONDCP International Technology Symposium.*

Lee, R. 1995. An NLToolset-Based System for MUC-6. In *Proceedings of the Sixth Message Understanding Conference*, 249-261. Columbia, Maryland: DARPA; Morgan Kaufmann Publishers.