INFORMATION EXTRACTION

Lance A. Ramshaw and Ralph M. Weischedel

BBN Technologies
10 Moulton St. MS 2/1B, Cambridge, MA 02138

ABSTRACT

Information Extraction (IE) maps a language stream into database records that capture part of its meaning. Name, entity, and relation extraction are common subtasks, and an event extraction subtask has also been proposed. Various specific target sets have been defined, with results compared in government-sponsored evaluations, but defining 'meaning' for a broad spectrum of applications remains a challenge. Researchers are exploring a wide variety of learning techniques for these tasks. Initial tests have also been performed measuring IE performance on speech recognition output, but current work has only scratched the surface of this important area.

1. GOALS

Perhaps the clearest image for the overall goal of information extraction is filling a database. To extract information, we need to understand enough of the meaning of the input language to be able to make appropriate database entries, of the same sort that humans could make to capture part of the meaning of the documents. Even if IE systems are not yet accurate enough for building useful databases automatically, their output can be very useful for assisting humans in that task, or for collecting evidence from documents that can then support higher-level theories and analysis.

For example, Fig. 1 on the following page shows BBN's demonstration "FactBrowser" interface, a tool that allows analysts to view and work with the database of extraction results from a collection of documents. Using this interface, the analyst can begin by typing a query in the box at the upper left; in this example, the analyst typed "chemical ali". The system then uses textual search to retrieve documents in which those words appear. In this collection, that search selects a couple hundred documents, of which the first 21 are shown in the left-hand panel.

When the analyst selects one of those documents, the text of the document is displayed in the lower right-hand pane. In the bottom center pane next to that, the system lists all of the entities that are mentioned in that document. "Ali, Chemical" happens to be the first entity in that list. When the analyst then selects that entity, the top right-hand pane displays all of the information that the system has extracted about that entity, not only from this document, but from every document in the collection.

The extracted information includes descriptions of the entity, meaning non-name phrases from the text that can supply additional information about it. In this case, the descriptions reveal that Ali is an Iraqi general, and Saddam's cousin. The system also extracts "facts", relations that link this entity to other entities. In this case, we see an employment relation between Ali and Iraq, and a membership relation between Ali and the Revolutionary Command Council. If the analyst then selects one of those relations, the document containing that relation will be displayed below, with the text string from which the relation was extracted highlighted.

This kind of interface shows how IE can help a user to follow informational threads through a large corpus of documents, using relational links, along with cases where they are mentioned in the same document to help determine how entities are connected. While the ultimate goal is IE systems that are accurate enough to fill databases on their own, the immediate criterion is sufficient accuracy to massively increase the efficiency of human analysts.

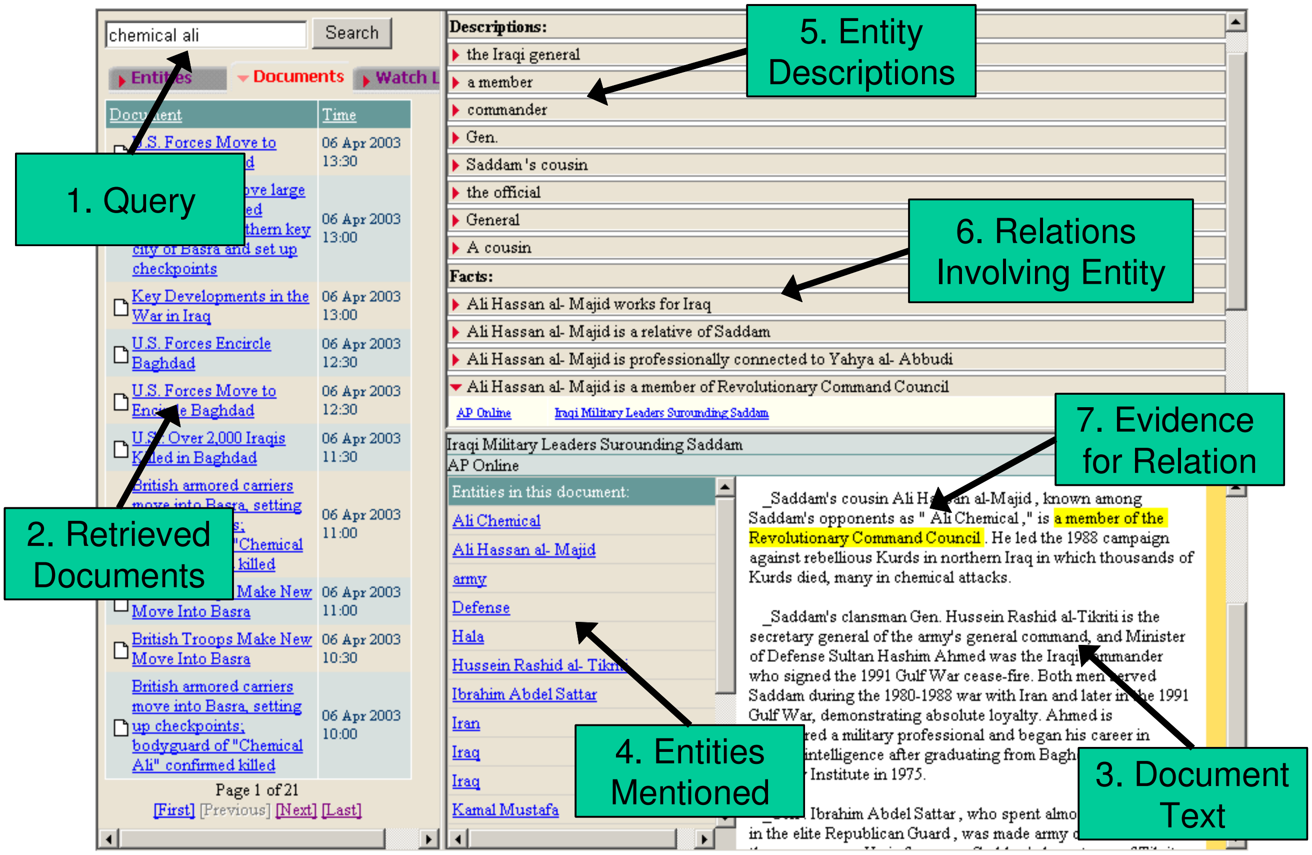
2. TASKS

That overall goal of IE has been broken down into tasks of increasing difficulty, beginning with name finding and moving up through entities, relations, and events. The definitions described here are largely those of the current ACE (Automatic Content Extraction) evaluations [4, 9].

2.1. Names

Extracting names means simply identifying them as substrings in the text and determining their type. This task is usually defined as restricted to non-nested names, allowing it to be approached as a tagging problem in the

0-7803-8874-7/05/$20.00 ©2005 IEEE



chemical ali

Document	Time
.S. Forces Move to d	06 Apr 2003 13:30
ve large ed them key city of Basra and set up checkpoints	06 Apr 2003 13:00
Key Developments in the War in Iraq	06 Apr 2003 13:00
U.S. Forces Encircle Baghdad	06 Apr 2003 12:30
U.S. Forces Move to Enci e Baghdad	06 Apr 2003 12:30
U.S. Over 2,000 Iraqis K led in Baghdad	06 Apr 2003 11:30
British armored carriers move into Basra, setting s; "Chemical killed	06 Apr 2003 11:00
Make New Move Into Basra	06 Apr 2003 11:00
British Troops Make New Move Into Basra	06 Apr 2003 10:30
British armored carriers move into Basra, setting up checkpoints; bodyguard of "Chemical Ali" confirmed killed	06 Apr 2003 10:00

Page 1 of 21
[First] [Previous] [Next] [Last]

Descriptions:
- the Iraqi general
- a member
- commander
- Gen.
- Saddam 's cousin
- the official
- General
- A cousin

Facts:
- Ali Hassan al- Majid works for Iraq
- Ali Hassan al- Majid is a relative of Saddam
- Ali Hassan al- Majid is professionally connected to Yahya al- Abbudi
- Ali Hassan al- Majid is a member of Revolutionary Command Council

AP Online   Iraqi Military Leaders Surounding Saddam

Iraqi Military Leaders Surounding Saddam
AP Online

Entities in this document:
Ali Chemical
Ali Hassan al- Majid
army
Defense
Hala
Hussein Rashid al- Tikri
Ibrahim Abdel Sattar
Iran
Iraq
Iraq
Kamal Mustafa

_Saddam's cousin Ali H an al-Majid , known among Saddam's opponents as " Ali Chemical ," is a member of the Revolutionary Command Council . He led the 1988 campaign against rebellious Kurds in northern Iraq in which thousands of Kurds died, many in chemical attacks.

_Saddam's clansman Gen. Hussein Rashid al-Tikriti is the secretary general of the army's general command, and Minister of Defense Sultan Hashim Ahmed was the Iraqi mmander who signed the 1991 Gulf War cease-fire. Both men erved Saddam during the 1980-1988 war with Iran and later in he 1991 Gulf War, demonstrating absolute loyalty. Ahmed is red a military professional and began his career in intelligence after graduating from Bagh Institute in 1975.

Ibrahim Abdel Sattar , who spent almo in the elite Republican Guard , was made army

Figure 1: BBN's FactBrowser Information Extraction Interface

linear sequence of words, rather than requiring more parse-like tree structures. Standard approaches to measuring name-finding performance were worked out under the earlier MUC (Message Understanding Conference) program [10].

2.2 Entities

In ACE terms, extracting an "entity" involves recognizing all of the places in the text where it is mentioned and then correctly predicting the coreference between those mentions, that they all do refer to the same entity. The mentions themselves can be either names (like "George Bush"), nominals (like "the US President" or "the occupant of the White House"), or pronouns (like "him" or "his"). The fact that coreference is central to the definition of entities means that even systems that correctly find and type all of the mentions of an entity can still score poorly due to coreference mistakes.

ACE currently targets 7 entity types, divided into a total of 44 subtypes. ACE also provides for entity attributes, like the birth date, gender, and nationality of a person, but the only attributes in the current task definition are the name strings of the entity. ACE is also moving toward cross document annotation and scoring of entities, where systems are required to correctly determine the coreference of mentions across the whole corpus, but the current task is still defined at the document level.

2.3 Relations

Relations in ACE terms connect two entities in some logical relationship. For example, "located" relations can apply between two geographical regions, locations, or facilities, "employment" or "membership" relations between people and organizations, and "family" relations between people. ACE currently targets 7 general relation types divided into a total of 23 such specific relation sub-types.

For a relation to be annotated in a document, it must be stated explicitly at least once. For example, a "family" relation between George and Laura Bush could be conveyed by phrases like "the president's wife" or "George married Laura in 1977". Thus, to identify relations, systems typically look for local evidence surrounding mentions of the two entities. Note that getting a particular relation correct requires making five correct choices. Both of the mentions must be found correctly, both must be coreferenced correctly, so that they represent the correct entity, and finally the type of the relation itself must be recognized correctly.

2.4 Events

The ACE program is currently working on extending its coverage to include events. Possible target events for organizations might include “start-org”, “merge-orgs”, or “declare bankruptcy”, while events involving people might include “be-born”, “marry”, “start-employment”, “arrest”, or “attack”. Unlike relations, events can involve more or less than two entities, and they typically have additional properties like times and locations. Current work is focusing on defining clear guidelines for identifying mentions of particular event types.

3. APPROACHES

While hand-written rules were at one time a common technique for IE, current research under the ACE program is directed largely toward trained models, given their obvious advantages in domain and language portability.

In a typical system design, the process begins with any necessary preprocessing, like identifying the word tokens if the input language does not mark word boundaries and perhaps predicting sentence breaks. Systems typically then process the text to identify entity mentions, named, nominal, and pronominal. A separate model is then used to predict coreference, meaning how that set of mentions should be partitioned into entities. Yet a third model is then often used to predict relations between entities based on textual evidence surrounding mentions of the two entities.

A wide range of machine learning techniques are being explored for these tasks. For finding names and nominal mentions, a common approach is to treat the task as a tagging problem, assigning tags to each word like “person-name-start” or “organization-name-continue” or “not-a-name”. HMMs [2], Maximum Entropy models [1, 5], Conditional Random Fields [6], and max margin techniques [3] have all been used here.

For the coreference task, the model must decide for each mention whether it belongs together with other, earlier mentions, or if the phrase is introducing an entity that has not been mentioned before. Many systems use a greedy strategy [11], scanning through the text and choosing for each mention whether or not it links back, and if so, to which previous entity. Other researchers have explored graph-partitioning approaches [8] that consider the whole document at once.

As noted previously, relations are typically predicted based on local evidence from those locations in the text where two entities are mentioned together. For example, in “the Bush ranch in Texas”, the location relation is conveyed by the facility and state being connected with “in”. One common approach is to reduce that context to a vector of features and then classify the instance based on those features as to which relation it conveys, if any [12].

4. EVALUATION

The speech field was able, fairly early on, to settle on shared specification for transcribing the words as the common goal against which to evaluate. Word error rate against that transcript has served as the standard measure of performance.

Evaluation for information extraction is a much more challenging issue, since there is no common definition of the meaning elements that are to be extracted. Different users may use different database schemas, causing the same bit of data to be encoded either, say, as an entity attribute or as a relation. Even when the schema is the same, defining the desired output class can be very difficult. For example, while an entity class like “country” would seem initially to be pretty straightforward, examples like “Palestine”, “Europe”, and “Kurdistan”, which are in some ways countries and in other ways not, show the kinds of problems that quickly arise. Annotation for IE always requires development of substantial guidelines documents describing the desired targets with sufficient examples to guide the annotation process.

These difficulties in pinning down IE targets can be seen in the human inter-annotator agreement rates for the various tasks. When comparing two sets of human annotation, the LDC (Linguistic Data Consortium) [7] reported value scores of 92.6% for entities, but only 70.2% for the more complex task of relations. The scores achieved by systems need to be considered in that light.

ACE evaluation is currently reported in terms of value scores, that take account of the estimated value of each data element to the end user or application. At the entity level, the answer key includes the entities of each of the specified target types that are mentioned in the document. For each entity, its type and all of its mentions (name, nominal, and pronominal) are specified. The scorer maps the entities in the system output onto those in the answer key and then scores the system’s output in terms of correctly found entities, missed entities, and false alarms.

The value of a correctly found entity depends on the type (people are viewed as more valuable extraction targets than countries) and on the number and type of its mentions (name mentions count more than nominal or pronominal mentions). If the system finds only some of the mentions, perhaps due to incorrect coreference choices, it gets partial credit for those that it did find, and loses credit for those it missed. Systems can also loose partial credit if they predict the entity type incorrectly. The total value

	Entities	Relations
English	79.9	49.0
Chinese	72.4	40.1
Arabic	74.5	29.7

Table 1: Current IE Performance Levels (ACE 2004 best system value scores)

	Entities	Relations
ASR	47.5	16.0
Transcripts	78.7	44.4

Table 2: Comparing IE on ASR vs. Transcripts (ACE 2004 value scores)

score is then divided by the value of the answer key's answers, resulting in a percentage of the maximum possible value that extraction could have achieved. Note that this value score can be negative, if the penalties for misses and false alarms are larger than the credit for correctly-found entities.

Relations are similarly value scored based on an alignment of their argument entities between the system output and the answer key.

Table 1 shows the value scores of the top-scoring system in the 2004 ACE evaluation for entities and relations, when scoring on a combination of newswire data and clean transcriptions of broadcast news. Results are given for English, Chinese, and Arabic.

5. COMBINING EXTRACTION AND SPEECH RECOGNITION

One subtask within ACE involves doing extraction directly on the ASR output for broadcast news, rather than on clean transcribed text. Table 2 shows the entity and relation value scores of the highest-performing system in the 2004 ACE evaluation on ASR, and compares them to the top scores when running on clean transcripts. The word error rate of the ASR system used to generate the test data was estimated at 8%. Note that the loss in IE value terms is more than proportional to the word error rate, as would be expected, given that systems have to get multiple items correct in order to get an entity or relation correct.

This evaluation on ASR output was not a central focus in ACE, and the IE systems were only supplied with a single hypothesis word string from the recognizer. Tighter integration between the speech recognition model and the extraction model should be able to improve those scores significantly.

6. SUMMARY

Information extraction is driven by the application need to find the meaning in the text, beyond just the words. Because its targets are more semantic, it faces greater challenges in defining the exact boundaries of its target concepts, and in laying out accepted evaluation measures for system performance. The model structures and features used also tend to be shaped by the particular semantic targets, whether names, entities, or relations, and by the linguistics of how those are conveyed in text. Still, many of the same statistical modeling techniques that are used for recognition are also useful for extraction, and closer integration between the two tasks offers an opportunity to further improve performance of extraction from speech.

7. REFERENCES

[1] Berger, A., S. Della Pietra, and V. Della Pietra. "A Maximum Entropy Approach to Natural Language Processing", *Computational Linguistics 22*, pp. 39-71, 1996.

[2] Bikel, D., R. Schwartz, and R. Weischedel. "An Algorithm That Learns What's in a Name", *Machine Learning 34,* pp. 211-241, 1999.

[3] Collins, M., and N. Duffy. "New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures and the Voted Perceptron", *Proc. of ACL 2002,* pp. 235-270, 2002.

[4] Doddington, G., et al., "The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation", *Proceedings of LREC 2004*, pp. 837-840, 2004.

[5] Ittycheria, A., L. Lita, N. Kambhatla, N. Nicolov, S. Roukos, and S. Stys. "Identifying and Tracking Entity Mentions in a Maximum Entropy Framework", *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, 2003.

[6] Lafferty, J., A. McCallum and F. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", *Proceedings of the 18th International Conference on Machine Learning,* pp. 282-289, 2001.

[7] Linguistic Data Consortium, *ACE website*, http://www.ldc.upenn.edu/Projects/ACE/

[8] McCallum, A, and B. Wellner. "Conditional Models of Indentity Uncertainty with Application to Noun Coreference", In *Advances in Neural Information Processing Systems 17,* MIT Press, 2005.

[9] NIST, *ACE website*, http://www.itl.nist.gov/iaui/894.01/tests/ace/

[10] NIST, *MUC-7 Proceedings*, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

[11] Ramshaw, L., E. Boschee, S. Bratus, S. Miller, R. Stone, R. Weischedel, and A. Zamanian. "Experiments in Multi-Modal Automatic Content Extraction", In *Proceedings of HLT*, Morgan Kaufman, pp. 110-113, 2001.

[12] Zelenko, D., C. Aone, and A. Richardella. "Kernel Methods for Relation Extraction", *Journal of Machine Learning Research 3,* pp. 1083-1106, 2003.