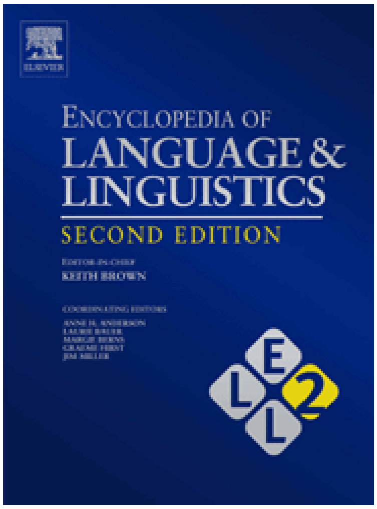


Cunningham H (2006), Information Extraction, Automatic. In: Keith Brown, (Editor-in-Chief) *Encyclopedia of Language & Linguistics, Second Edition*, volume 5, pp. 665-677. Oxford: Elsevier.

Heath S B (1983). *Ways with words*. Cambridge, UK: Cambridge University Press.
Kortmann B (ed.) (2004). *Handbook of varieties of English*. Berlin: Mouton de Gruyter.
Miller J (2003). 'Syntax and discourse in Modern Scots.' In Corbett J, McClure J D & Stuart-Smith J (eds.) *The Edinburgh companion to Scots*. Edinburgh: Edinburgh University Press. 72–109.
Ong W (1982). *Orality and literacy*. London: Methuen.
Perera K (1984). *Children's writing and reading*. Oxford: Basil Blackwell.
Pinker S (1994). *The language instinct*. Harmondsworth, UK: Allen Lane, Penguin Press.
Rosen C & Rosen H (1973). *The language of primary school children*. Harmondsworth, UK: Penguin.

Information Extraction, Automatic

H Cunningham, University of Sheffield, Sheffield, UK

Introduction: Extraction and Retrieval

Information extraction (IE) is a technology based on analyzing natural language in order to extract snippets of information. The process takes texts (and sometimes speech) as input and produces fixed-format, unambiguous data as output. This data may be used directly for display to users, or may be stored in a database or spreadsheet for later analysis, or may be used for indexing purposes in information retrieval (IR) applications such as Internet search engines like Google.

IE is quite different from IR:

- an IR system finds relevant texts and presents them to the user;
- an IE application analyzes texts and presents only the specific information from them that the user is interested in.

For example, a user of an IR system wanting information on trade group formations in agricultural commodities markets would enter a list of relevant words and receive in return a set of documents (e.g., newspaper articles) that contain likely matches. Users would then read the documents and extract the requisite information themselves. They might then enter the information in a spreadsheet and produce a chart for a report or presentation. In contrast, an IE system would automatically populate the spreadsheet directly with the names of relevant companies and their groupings.

There are advantages and disadvantages with IE in comparison to IR. IE systems are more difficult and knowledge-intensive to build, and are to varying degrees tied to particular domains and scenarios. IE is more computationally intensive than IR. However, in applications where there are large text volumes IE is potentially much more efficient than IR because of the possibility of dramatically reducing the amount of time people spend reading texts. Also, where results need to be presented in several languages, the fixed-format, unambiguous nature of IE results makes this relatively straightforward in comparison with providing the full translation facilities needed for interpretation of multilingual texts found by IR.

The rest of this article:

- discusses the origins of IE and the parameters that condition its performance in application settings (see 'Language Computation without Understanding' below);
- presents some application scenarios in which IE is being deployed (see 'Application Scenarios' below);
- gives a canonical definition of five subtasks that collectively characterize the spectrum of IE systems (see 'Five Types of IE' below);
- looks at the development of IE research subsequent to the MUC (message understanding conferences) program of the late 20th century (see 'IE After MUC' below).

See also **Natural Language Processing: Overview.** Other overview sources on IE include: Cowie and Lehnert (1996); Appelt (1999); Cunningham (1999); Gaizauskas and Wilks (1998); Pazienza (2003).

Language Computation without Understanding

Information extraction in the sense discussed here grew out of work in the late 1980s and 1990s the Message Understanding Conferences (MUCs related to Grishman and Sundheim (1996); Sundheim (1995); Chinchor (1998) – *see* **Text Retrieval Conference and Message Understanding Conference**). (Previous work in computation with human language had attempted similar things in some cases, and indeed a related term "fact extraction" was in use as far back as the 1960s. A formal and widely accepted definition of IE did not emerge until the work discussed here, however.) These events were particularly distinctive because they employed a strict quantitative evaluation procedure where different research sites first defined a precise task, then implemented competing systems that were measured against human-annotated data in a controlled environment. Although this type of experimental procedure seems obvious in many other branches of science, its use prior to MUC was uncommon in many language processing contexts (perhaps due to the early association between the work and phenomenological fields such as linguistics, as opposed to empirical experimentation or engineering – Boguraev *et al.*, 1995). The experimental cycle of the MUC competitions drove progress in IE and led to the canonical definition that is presented below in 'Five Types of IE.' The series also contributed much of our understanding of how to measure the performance of IE systems (*see* **Natural Language Processing: System Evaluation**).

The MUCs were held over the decade from 1987; following this IE work took several directions, including commercialization of some of the basic techniques, additional work on portability (see 'Portable IE') and, most recently, connecting with efforts related to the general project of the Semantic Web Berners-Lee (1999) (see 'Ontology-Based IE' below).

An important outcome of the last 15 years of work on IE is that the community now understands reasonably precisely the various parameters that influence the performance of the technology in diverse application settings. (In other words, the scientific work has laid the basis for the engineering of practical systems.) The next section looks at these parameters.

Complexity vs. Specificity

The difference between the general project of language analysis (or understanding) and the more restricted enterprise of extraction is a response to the low performance of analysis in the general case. To illustrate, imagine that we wish to build an application which uses language analysis to support some of its functions. Then consider that items of information to be extracted can vary in complexity (e.g., we may be interested simply in people names or in complex events that involve multiple participants) and in specificity (e.g., we may be interested in general information reported in any way in any text or in specific domains reported in particular ways in certain types of text). If we plot the acceptable performance level of analysis components on a graph of complexity vs. specificity of the information to be extracted (Figure 1), then there is an obvious trade-off to be made between the two: the more complex the data to be extracted, the more specific must be the domain of discourse; the simpler the data, the more generally the extraction algorithms may be applied. This two-dimensional view of IE performance hides a number of subtle issues. Specificity of an IE task is in itself multidimensional, being influenced by both text type and genre. For example, a seminar announcement text (a very specific text type, easy for IE systems to analyze) may be harder or easier to process depending on genre. For example, if it is a seminar series about famous historical figures there may be ambiguity between the speakers' names and those of their subjects. Conversely, if the series relates to mobile telecommunications, the regularity and specificity of the terminology typical of this genre should make structural analysis easier. Therefore we must also consider the parameters text type and subject (or genre, or domain):

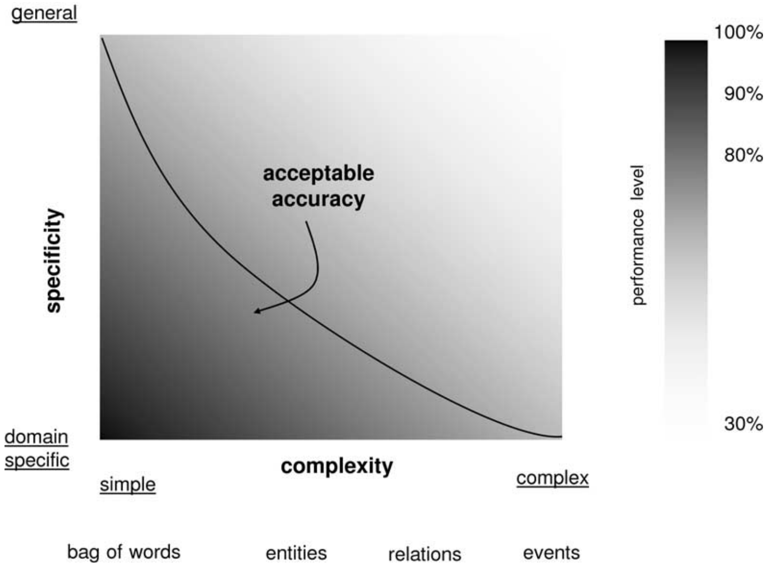


Figure 1 Performance trade-off relative to specificity and complexity.

- Text type: the kinds of texts we are working with, for example: *Wall Street Journal* articles, or email messages, or novels, or the output of a speech recognizer.
- Domain: the broad subject matter of those texts, for example: world events, or symptoms of a deficiency in a particular enzyme, seminar announcements, financial news, requests for technical support, or tourist information, and the style in which they are written, e.g., informal, formal.

Finally there is the question of the particular types of information that the IE user is interested in, for example: personal names, rocket launches, mergers between companies, problems experienced with a particular software package, or descriptions of how to locate parts of a city. We return to this issue in 'Five Types of IE' below.

Application Scenarios

This section contains some scenarios that illustrate how IE application software can mediate between the text and the structured information needs of various types of user. In each case, the user specifies an information need to IE development staff, who then prepare a custom extraction system. The input materials are increasingly defined as subsections of the World Wide Web, either the formal materials of news, company sites, etc., or the informal Web of weblogs, mailing lists and wikis. The IE developers must analyze the problem according to the dimensions outlined in 'Complexity vs. Specificity' above, and determine how to combine human and machine processing so as to attain the required performance profile (see 'Portable IE' below).

The types of scenario given here are of increasing relevance due in large part to the explosive growth, dynamic nature and multilingual content of the Web.

Financial Analysts

The Web contains various indications of how a company is viewed, and whether or not it can be expected to perform strongly in the coming period. Some of this data is already highly analyzed, e.g., financial news in English. Other data, in other languages, or from less well scrutinized sources, is both voluminous and obscure. IE can enable analysts to answer questions such as

- How many instances predicting strong performance for a particular company are out there?
- Over the past year how has the profile of predictions for this company changed?
- How many positive/negative sentiments were expressed for the company?

Marketing Strategists

The dynamic adjustment of marketing spending is currently made more difficult by the lack of metrics that indicate the impact of campaign elements. IE can support campaign tuning today based on yesterday's results, producing outputs such as:

> In this morning's IT press 7% of articles discussed your company. The average proportion of the article directly relating to your company was 33%. The figures for the other key players in your sector are summarized in the following table…

Similarly, we can measure the extent of media coverage relative to spend events:

> Company Y exhibited at Comdex. In the week following the exhibition 20% of the press that covered Comdex mentioned Y.

PR Workers

Public relations staff are concerned to identify negative reporting events as quickly as possible in order to respond. IE can be configured to report like this:

> The table below summarizes 12 negative reporting events concerning your company in the last 24 hours of IT news…

Media Analysts

IE can be used to create a range of media metrics, for example the media distance, or extent of collocational association, between concepts and products/companies:

> The media distance between your company and the subject of XML is 0.09; for IBM the value is 0.2.

Five Types of IE

By the time that it ended in 1998 (the final conference was MUC-7 (SAIC 1998)) the MUC programme had arrived at a definition of IE split into five tasks:

- Named entity recognition (NE)
  Finds and classifies names, places, etc.

- Coreference resolution (CO)
  Identifies identity relations between entities.

- Template element construction (TE)
  Adds descriptive information to NE results (using CO).

- Template relation construction (TR)
  Finds relations between TE entities.

- Scenario template production (ST)
  Fits TE and TR results into specified event scenarios.

In simpler terms, NE is about finding entities, CO about which entities and references (such as pronouns) refer to the same thing, TE about what attributes entities have, TR about what relationships between entities there are, and ST about events that the entities participate in. Consider these sentences:

> The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head. Dr. Head is a staff scientist at We Build Rockets Inc.

NE discovers that the entities present are the *rocket, Tuesday, Dr Head* and *We Build Rockets Inc.* CO discovers that *it* refers to the rocket. TE discovers that the rocket is *shiny red* and that it is Head's *brainchild*. TR discovers that *Dr Head* works for *We Build Rockets Inc.* ST discovers that there was a rocket launching event in which the various entities were involved.

The scenarios presented above in 'Application Scenarios' deploy these five types of information invarious combinations (and with various typical performance profiles).

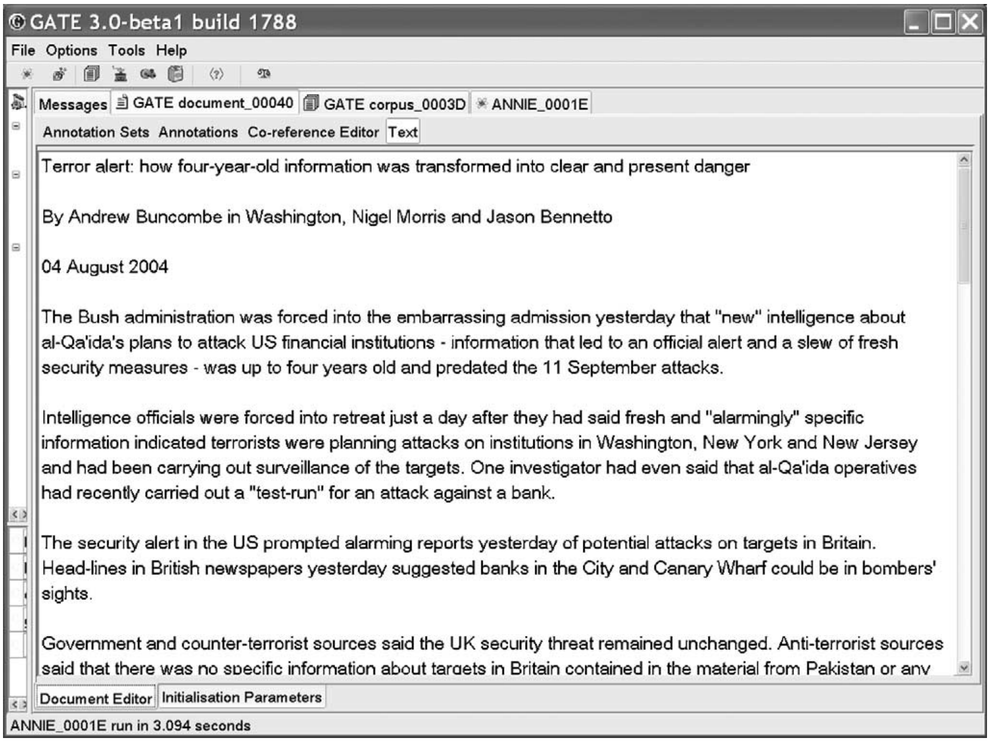
These various types of IE provide progressively higher-level information about texts. They are described in more detail below, with examples (see also the extended example in 'An Extended Example').

Named Entity Recognition

The simplest and most reliable IE technology is *Named entity recognition* (NE – *see* **Named Entity Extraction**). NE systems identify all the names of people, places, organizations, dates, amounts of money, etc. So, for example, if we run the text in Figure 2 (*The Independent*, London, 4th August 2004.) through an NE recognizer, the result is as in Figure 3. (The results shown here and below are from IE software distributed with the GATE system (Cunningham 2002) – *see* **Computational Language Systems: Architectures**.)

All things being equal, NE recognition can be performed at up to around 95% accuracy (see also 'Complexity vs. Specificity'). Given that human annotators do not perform to the 100% level (measured in MUC by interannotator comparisons), NE recognition can now be said to function at human performance levels, and applications of the technology are increasing rapidly as a result.

The process is weakly domain dependent, i.e., changing the subject matter of the texts being processed from financial news to other types of news would involve some changes to the system, and changing from news to scientific papers would involve quite large changes.


GATE 3.0-beta1 build 1788
File Options Tools Help
Messages
GATE document_00040
GATE corpus_0003D
ANNIE_0001E
Annotation Sets Annotations Co-reference Editor Text
Document Editor Initialisation Parameters

Terror alert: how four-year-old information was transformed into clear and present danger

By Andrew Buncombe in Washington, Nigel Morris and Jason Bennetto

04 August 2004

The Bush administration was forced into the embarrassing admission yesterday that "new" intelligence about al-Qa'ida's plans to attack US financial institutions - information that led to an official alert and a slew of fresh security measures - was up to four years old and predated the 11 September attacks.

Intelligence officials were forced into retreat just a day after they had said fresh and "alarmingly" specific information indicated terrorists were planning attacks on institutions in Washington, New York and New Jersey and had been carrying out surveillance of the targets. One investigator had even said that al-Qa'ida operatives had recently carried out a "test-run" for an attack against a bank.

The security alert in the US prompted alarming reports yesterday of potential attacks on targets in Britain. Head-lines in British newspapers yesterday suggested banks in the City and Canary Wharf could be in bombers' sights.

Government and counter-terrorist sources said the UK security threat remained unchanged. Anti-terrorist sources said that there was no specific information about targets in Britain contained in the material from Pakistan or any

ANNIE_0001E run in 3.094 seconds

Figure 2 An example text.

Coreference Resolution

Coreference resolution (CO) involves identifying identity relations between entities in texts. These entities are both those identified by NE recognition and anaphoric references to those entities. For example, in

> Alas, poor Yorick, I knew him Horatio.

coreference resolution would tie 'Yorick' with 'him' (and 'I' with Hamlet, if sufficient information was present in the surrounding text).

This process is less relevant to users than other IE tasks (i.e., whereas the other tasks produce output that is of obvious utility for the application user, this task is more relevant to the needs of the application developer). For text browsing purposes, we might use CO to highlight all occurrences of the same object or provide hypertext links between them. CO technology might also be used to make links between documents. The main significance of this task, however, is as a building block for TE and ST. CO enables the association of descriptive information scattered across texts with the entities to which it refers. To continue the hackneyed Shakespeare example, coreference resolution might allow TE or TR analysis to situate Yorick in Denmark. Figure 4 shows results for the example text from Figures 2 and 3.

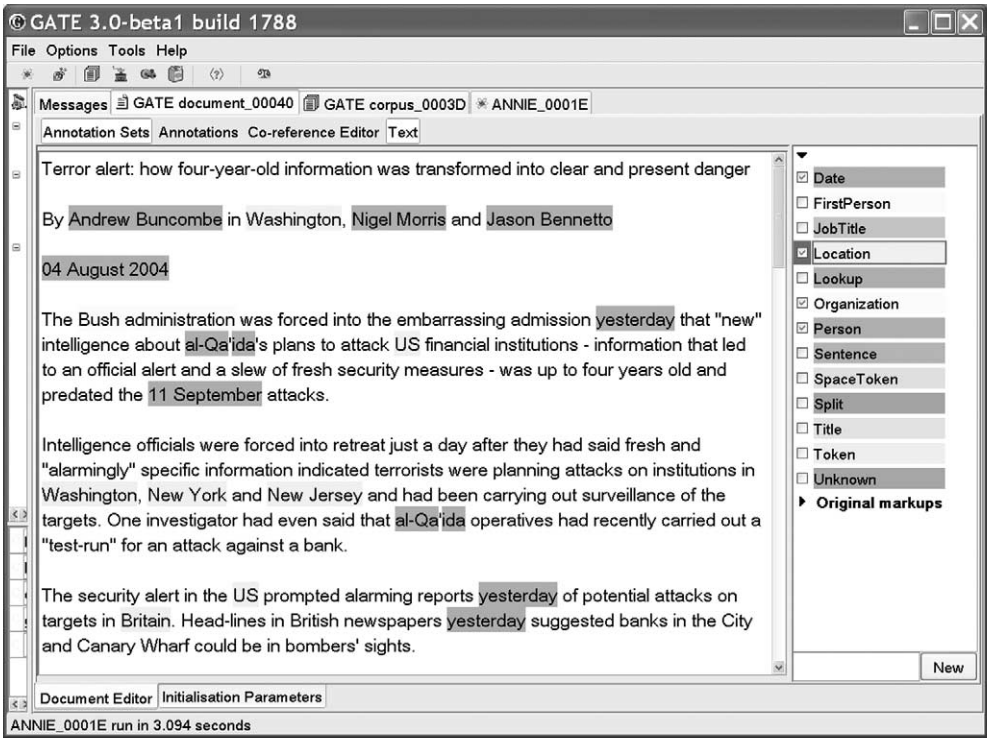
CO breaks down into two subproblems: anaphoric resolution (e.g., 'I' with Hamlet); and proper-noun resolution. Proper-noun coreference identification finds occurrences of same object represented with different spelling or compounding, e.g., 'IBM,' 'IBM Europe,' 'International Business Machines Ltd.,' etc.). CO resolution is an imprecise process, particularly when applied to the solution of anaphoric reference. CO results vary widely; depending on domain perhaps only 50–60% may be relied upon. CO systems are domain-dependent.

Template Element Production

The TE task builds on NE recognition and coreference resolution, associating descriptive information with the entities. For example, from the Figure 2 text the system finds out that the 'Bush administration' is also referred to as 'government officials,' and adds this as an alias.

Template elements for the example text are given in Figure 5. The format is an arbitrary one; the point to note is that it is essentially a database record, and could just as well be formatted for SQL (structured query language) operations, or reading into a spreadsheet, or (with some extra processing) for multilingual presentation. See 'An Extended Example' for an example in a simplified format.

Good scores for TE systems are around 80% (on similar tasks humans can achieve results in the mid-90s, so there is some way to go). As in NE recognition,



Terror alert: how four-year-old information was transformed into clear and present danger

By Andrew Buncombe in Washington, Nigel Morris and Jason Bennetto

04 August 2004

The Bush administration was forced into the embarrassing admission yesterday that "new" intelligence about al-Qa'ida's plans to attack US financial institutions - information that led to an official alert and a slew of fresh security measures - was up to four years old and predated the 11 September attacks.

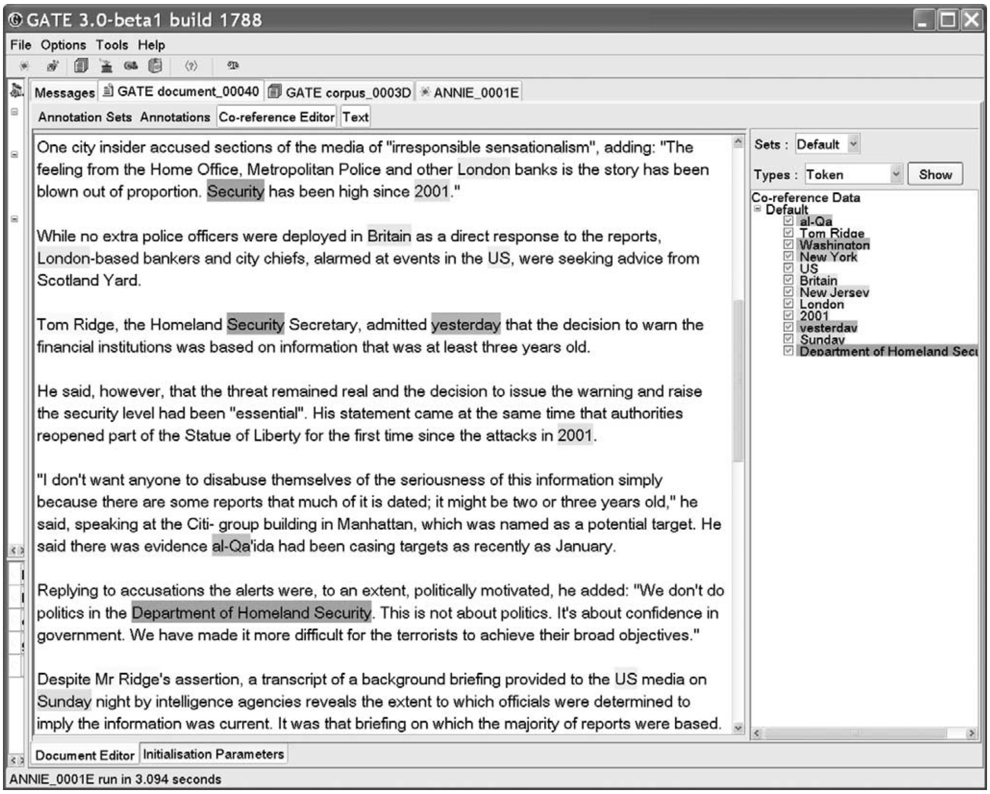
Intelligence officials were forced into retreat just a day after they had said fresh and "alarmingly" specific information indicated terrorists were planning attacks on institutions in Washington, New York and New Jersey and had been carrying out surveillance of the targets. One investigator had even said that al-Qa'ida operatives had recently carried out a "test-run" for an attack against a bank.

The security alert in the US prompted alarming reports yesterday of potential attacks on targets in Britain. Head-lines in British newspapers yesterday suggested banks in the City and Canary Wharf could be in bombers' sights.

☑ Date
□ FirstPerson
□ JobTitle
☑ Location
□ Lookup
☑ Organization
☑ Person
□ Sentence
□ SpaceToken
□ Split
□ Title
□ Token
□ Unknown
▸ Original markups

ANNIE_0001E run in 3.094 seconds

Figure 3 Named entity recognition.



One city insider accused sections of the media of "irresponsible sensationalism", adding: "The feeling from the Home Office, Metropolitan Police and other London banks is the story has been blown out of proportion. Security has been high since 2001."

While no extra police officers were deployed in Britain as a direct response to the reports, London-based bankers and city chiefs, alarmed at events in the US, were seeking advice from Scotland Yard.

Tom Ridge, the Homeland Security Secretary, admitted yesterday that the decision to warn the financial institutions was based on information that was at least three years old.

He said, however, that the threat remained real and the decision to issue the warning and raise the security level had been "essential". His statement came at the same time that authorities reopened part of the Statue of Liberty for the first time since the attacks in 2001.

"I don't want anyone to disabuse themselves of the seriousness of this information simply because there are some reports that much of it is dated; it might be two or three years old," he said, speaking at the Citi- group building in Manhattan, which was named as a potential target. He said there was evidence al-Qa'ida had been casing targets as recently as January.

Replying to accusations the alerts were, to an extent, politically motivated, he added: "We don't do politics in the Department of Homeland Security. This is not about politics. It's about confidence in government. We have made it more difficult for the terrorists to achieve their broad objectives."

Despite Mr Ridge's assertion, a transcript of a background briefing provided to the US media on Sunday night by intelligence agencies reveals the extent to which officials were determined to imply the information was current. It was that briefing on which the majority of reports were based.

☑ al-Qa
☑ Tom Ridge
☑ Washington
☑ New York
☑ US
☑ Britain
☑ New Jersey
☑ London
☑ 2001
☑ yesterday
☑ Sunday
☑ Department of Homeland Secu

ANNIE_0001E run in 3.094 seconds

Figure 4 Coreference resolution highlighted in example text.

the production of TEs is weakly domain-dependent, i.e., changing the subject matter of the texts being processed from financial news to other types of news would involve some changes to the system, and changing from news to scientific papers would involve quite substantial changes.

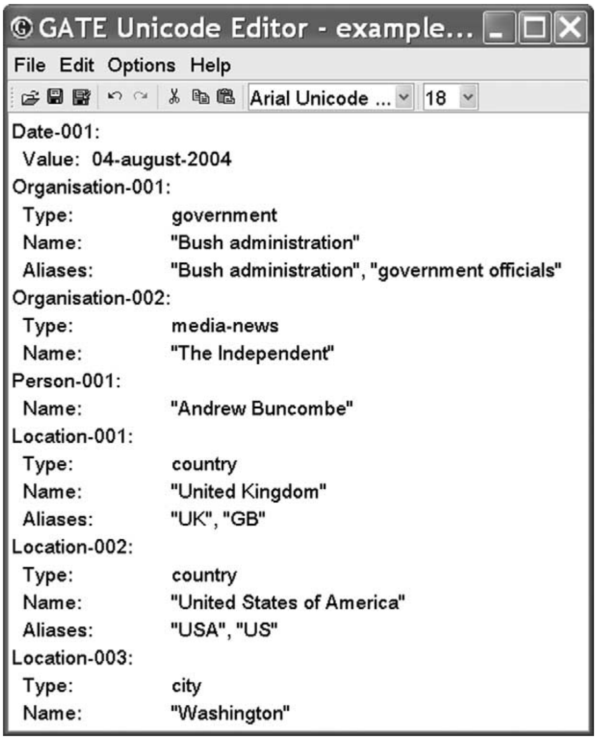
Template Relation Production

Before MUC-7, relations between entities were part of the scenario-specific template outputs of IE evaluations. In order to capture more widely useful relations, MUC-7 introduced the TR task – see Figure 6. As described in Appelt (1999), "The template relation task requires the identification of a small number of possible relations between the template elements identified in the template element task. This might be, for example, an employee relationship between a person and a company, a family relationship between two persons, or a subsidiary relationship between two companies. Extraction of relations among entities is a central feature of almost any information extraction task, although the possibilities in real-world extraction tasks are endless." In general, good TR scores reach around 75%. TR is a weakly domain-dependent task.

Scenario Template Extraction

Scenario templates (STs) are the prototypical outputs of IE systems, being the original task for which the term was coined. They tie together TE entities and TR relations into event descriptions. For example, TE may have identified Isabelle, Dominique and Françoise as people entities present in the Robert edition of Napoleon's love letters. ST might then identify facts such as that Isabelle moved to Paris in August 1802 from Lyon to be nearer the little guy, that Dominique then burned down Isabelle's apartment building and that Françoise ran off with one of Gérard Depardieu's ancestors. A slightly more pertinent example is given in Figure 7, adapted from MUC-6 ARPA (1995).

ST is a difficult IE task; the best MUC systems score around 60%. The human score can be as low as around 80+%, which illustrates the complexity involved. These figures should be taken into account when considering appropriate applications of ST technology. Note, however, that it is possible to increase precision at the expense of recall: we can develop ST systems that do not make many mistakes, but that miss quite a lot of occurrences of relevant scenarios. Alternatively, we can push up recall



```
Date-001:
 Value:  04-august-2004
Organisation-001:
 Type:          government
 Name:          "Bush administration"
 Aliases:       "Bush administration", "government officials"
Organisation-002:
 Type:          media-news
 Name:          "The Independent"
Person-001:
 Name:          "Andrew Buncombe"
Location-001:
 Type:          country
 Name:          "United Kingdom"
 Aliases:       "UK", "GB"
Location-002:
 Type:          country
 Name:          "United States of America"
 Aliases:       "USA", "US"
Location-003:
 Type:          city
 Name:          "Washington"
```

Figure 5 Template elements.

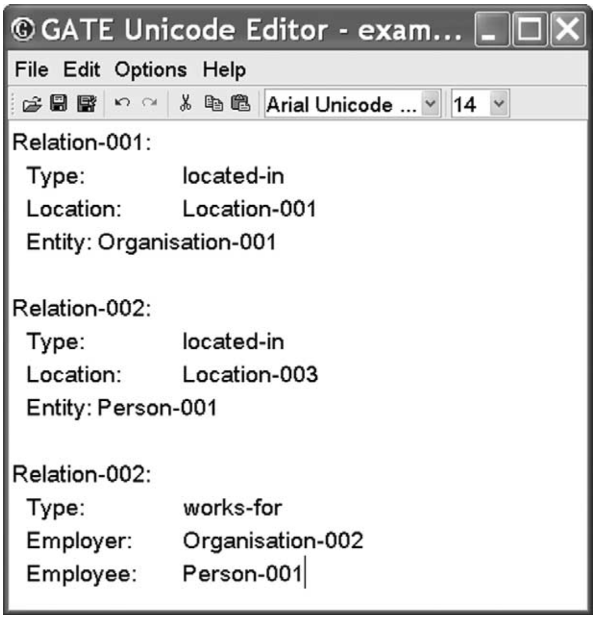
and miss less, but at the expense of making more mistakes.

The ST task is both domain-dependent and, by definition, tied to the scenarios of interest to the users. Note, however, that the results of NE, TR and TE feed into ST. Note also that in MUC-6 and MUC-7 the developers were given the specifications for the ST task only 1 month before the systems were scored. This was because it was noted that an IE system that required very lengthy revision to cope with new scenarios was of less worth than one that could meet new specifications relatively rapidly (see 'Portable IE'). As a result of this, the scores for ST in MUC-6/7 were probably slightly lower than they might have been with a longer development period. Experience from previous MUCs and from subsequent work suggests, however, that current technology has difficulty attaining scores much above 60% accuracy for this task.

An Extended Example

So far we have discussed IE from a general perspective. In this section we look at the capabilities that might be delivered as part of an application designed to support analysts tracking international drug dealing.

When the system is specified, our imaginary analyst states that "the operational domains that user interests are centered upon are … drug enforcement, money laundering, organized crime, terrorism, legislation." The entities of interest within these domains are cited as "person, company, bank, financial entity, transportation means, locality, place, organization, time, telephone, narcotics, legislation, activity." A number of relations (or links) are also specified, for example between people, between people and companies, etc. These relations are not typed, i.e., the kind of relation involved is not specified. Some relations take the form of properties of entities – e.g., the location of a company – while others denote events – e.g., a person visiting a ship.



```
Relation-001:
 Type:          located-in
 Location:      Location-001
 Entity: Organisation-001

Relation-002:
 Type:          located-in
 Location:      Location-003
 Entity: Person-001

Relation-002:
 Type:          works-for
 Employer:      Organisation-002
 Employee:      Person-001
```

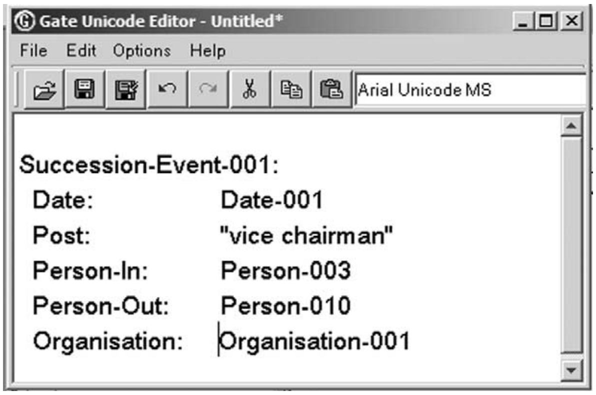
Figure 6 Template relations.

Working from this starting point, an IE system is designed that:

- is tailored to texts dealing with drug enforcement, money laundering, organized crime, terrorism, and legislation;
- recognizes entities in those texts and assigns them to one of a number of categories drawn from the set of entities of interest (person, company, etc.);
- associates certain types of descriptive information with these entities, e.g., the location of companies;
- identifies a set (relatively small to begin with) of events of interest by tying entities together into event relations.

For example, consider the following text:

> Reuter – New York, Wednesday 12 July 2005.
>
> New York police announced today the arrest of Frederick J. Thompson, head of Jay Street Imports Inc., on charges of drug smuggling. Thompson was taken from



Succession-Event-001:

Date:	Date-001
Post:	"vice chairman"
Person-In:	Person-003
Person-Out:	Person-010
Organisation:	Organisation-001

Figure 7 Scenario template.

> his Manhattan apartment in the early hours yesterday. His attorney, Robert Guliani, issued a statement denying any involvement with narcotics on the part of his client. “No way did Fred ever have dealings with dope,” Guliani said.
>
> A Jay Street spokesperson said the company had ceased trading as of today. The company, a medium-sized import-export concern established in 1989, had been the main contractor in several collaborative transport ventures involving Latin-American produce. Several associates of the firm moved yesterday to distance themselves from the scandal, including the midwestern transportation company Downing-Jones.
>
> Thompson is understood to be accused of importing heroin into the United States.

From this IE might produce information such as the following (in some format to be determined according to user requirements, e.g., SQL statements addressing some database schema).

First, a list of entities and associated descriptive information. Relations of property type are made explicit. Each entity has an ID, e.g., ENTITY-2, which can be used for cross-referencing between entities and for describing events involving entities. Each also has a type, or category, e.g., company, person. Additionally, various type-specific information is available, e.g., for dates, a normalization giving the date in standard format.

Reuter	
id:	ENTITY-1
type:	company
business:	news
New York	
id:	ENTITY-2
type:	location
subtype:	city
is_in:	US
Wednesday 12 July 2005	
id:	ENTITY-3
type:	date
normalization:	12/07/2005
New York police	
id:	ENTITY-4
type:	organization
location:	ENTITY-2
Frederick J. Thompson	
id:	ENTITY-5
type:	person
aliases:	Thompson; Fred
domicile:	ENTITY-7
profession:	managing director
employer:	ENTITY-6
Jay Street Imports Inc.	
id:	ENTITY-6
type:	organization
aliases:	Jay Street
business:	import-export
Manhattan	
id:	ENTITY-7
type:	location
subtype:	city
is_in:	ENTITY-2
Robert Guliani	
id:	ENTITY-8
type:	person
aliases:	Guliani
1989	
id:	ENTITY-9
type:	date
normalization:	?/?/1989
Latin-America	
id:	ENTITY-10
type:	location
subtype:	country
Downing-Jones	
id:	ENTITY-11
type:	organization
business:	transportation
heroin	
id:	ENTITY-12
type:	drug
class:	A
United States	
id:	ENTITY-13
type:	location
subtype:	country

(These results correspond to the combination of NE and TE tasks; if we removed all but the type slots we would be left with the NE data.)

Second, relations of event type, or scenarios:

narcotics-smuggling	
id:	EVENT-1
destination:	ENTITY-13
source:	unknown
perpetrators:	ENTITY-5, ENTITY-6
status:	on-trial
joint-venture	

id:	EVENT-2
type:	transport
companies:	ENTITY-6, ENTITY-11
status:	past

(These results correspond to the ST task.)

Multilingual Extraction

The results described above may be translated for presentation to the user or for storage in existing databases. In general, this task is much easier than translation of ordinary text, and is close to software localization, the process of making a program's messages and labels on menus and buttons multilingual. Localization involves storing lists of direct translations for known items. In our case these lists would store translations for words such as 'entity,' 'location,' 'date,' 'heroin.' We also need ways to display dates and numbers in local formats, but code libraries are available for this type of problem.

Problems can arise where arbitrary pieces of text are used in the entity description structures, for example the descriptor slot in MUC-6 TE objects. Here a noun phrase from the text is extracted, with whatever qualifiers, relative clauses, etc. happen to be there, so the language is completely unrestricted and would need a full translation mechanism.

IE After MUC

In this section we bring the story up-to-date by looking at developments in IE system portability, the ACE program which succeeded MUC, and the project of automated annotation for semantic web applications.

Portable IE

A particular IE application might be configured to process financial news articles from a particular news provider (written in the house style) and find information about mergers between companies and various other scenarios. The performance of the application would be predictable for only this conjunction of factors. If it was later required to extract facts from the love letters of Napoleon Bonaparte as published on wall posters in the 1871 Paris Commune, performance levels would no longer be predictable. Tailoring an IE system to new requirements is a task that varies in scale depending on the degree of variation in the parameters discussed in 'Complexity vs. Specificity.'

Therefore a central track of IE research addresses the issue of portability, and this has been one of the most fruitful areas of development since the end of the MUC program in the late 1990s.

We can distinguish three broad currents of work in this area:

1. Learning extraction rules of models from annotated examples.
2. Embedding learning systems within end-user systems.
3. Supporting the development of rules/models by skilled specialist staff.

The first approach is to learn part or all of the extraction system from annotated training data. The advantage is a reduction in the need for skilled staff to perform system porting. The disadvantages are:

- only simple data can be extracted, or complex data from simple texts, such as seminar announcements (in fact many of the algorithms currently common in this area were developed for screen scraping which is a simpler task than most language analysis). (Screen scraping is the process of deleting the presentational elements of data or text, e.g., as displayed on a web page, in order to allow further processing of the data. Sites that provide comparative shopping lists are often based on scraping the pages of competing suppliers and representing the data for comparison purposes.);
- large volumes of training data may be required.

Surveys of work in this area are to be found in Cardie (1997); Daelemans and Osborne (2003).

Secondly, researchers have attempted to embed learning within end-user tools where the users correct IE suggestions. This approach addresses the problem of the costs of producing training data associated with the learning approach by speeding up the annotation process. A cyclical method known as mixed-initiative learning is used, where the user begins by doing all the annotation work manually while the system learns in the background. When the quality of the learned models is high enough, the system can then propose annotations to the user; correction of mistakes is fed back into the learning algorithm. See Day *et al.*, (1997); Grishman (2001). Mixed initiative learning was reinvented as adaptive IE in later work.

Lastly, IE system portability can be maximized by providing a development environment for skilled staff to adapt a core system. The advantages are:

- the core system can be designed for robustness and portability;
- extraction data complexity is not limited by a learning algorithm;
- all the engineering aspects of the process can be taken care of by the infrastructure (from data visualization to performance evaluation).

The disadvantage is that the adaptation process is labor intensive, and it is difficult for end-users to acquire the necessary skills. A survey of work is available (Cunningham and Scott, 2004); *see* **Computational Language Systems: Architectures**.

ACE: Automatic Content Extraction

The automatic content extraction program (ACE, 2004; Maynard *et al*., 2003) is a successor to MUC that has been running since a pilot study in 1999, which has continued the competitive quantitative evaluation cycles of its predecessor. ACE differs from MUC in three significant ways:

1. Several of the five MUC tasks defined in 'Five Types of IE' are conflated in ACE. The NE and CO tasks become a single entity detection and tracking (EDT) task in ACE, and the TE and TR tasks a single relation detection and tracking task. The ST task is renamed event detection and characterization.
2. The ACE tasks are more complex than their MUC counterparts in several ways. In the EDT task, for example, there is a more fine-grained taxonomy of entities, and it is necessary for systems to interpret metonymic entity mentions, requiring semantic analysis of the texts under consideration. Also, multiple domains and sources are used, including materials output from automatic speech transcription and optical character recognition programs.
3. The evaluation results from the ACE program are not public. Whereas in MUC all the competition results were made public, ACE results are restricted to the participants. The utility of the program for nonparticipants is therefore much lower than for MUC.

Ontology-Based IE

The Semantic Web aims to add a machine-tractable, re-purposeable layer of annotation relative to ontologies to complement the existing web of natural language hypertext (Fensel *et al*., 2002; Davies *et al*., 2002; Bechhofer *et al*., 2003). In order to realize this vision, the creation of semantic annotation, the linking of web pages to ontologies, and the creation, evolution and interrelation of ontologies must become automatic or semiautomatic processes, and a significant body of recent work has looked at the application of ontology-based IE (OBIE) (Bontcheva, 2004) in this context. Figure 8 illustrates the way in which IE and other language technologies can be used to bring together the natural language upon which the current web is mainly based and the formal knowledge needed for a semantic web. Figure 9 illustrates OBIE in action: the ontology hierarchy appears on the right, with an annotated document in the center of the pane. OBIE poses two main challenges:

- the identification of concept instances from the ontology in the text;

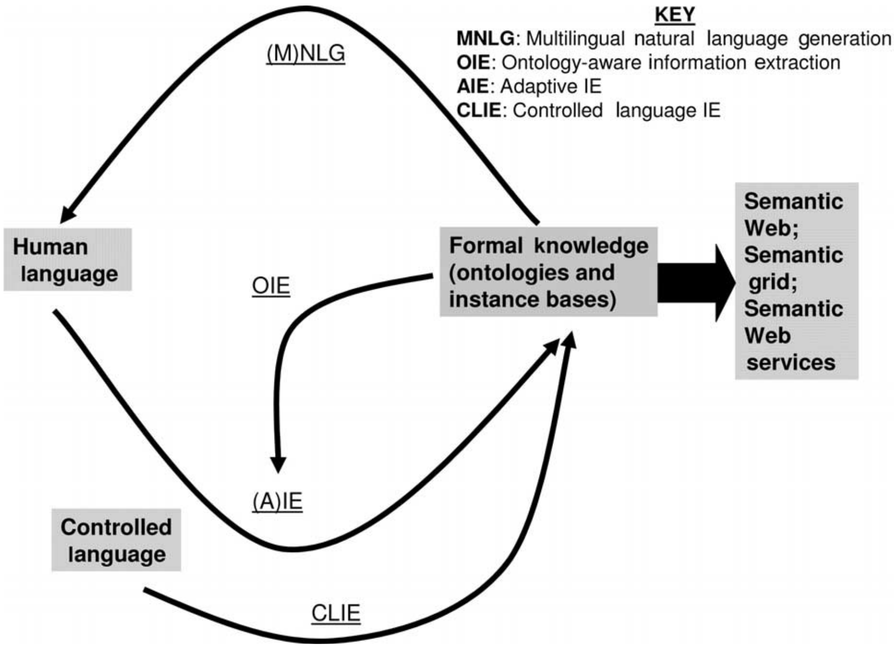

KEY
MNLG: Multilingual natural language generation
OIE: Ontology-aware information extraction
AIE: Adaptive IE
CLIE: Controlled language IE
(M)NLG
Human language
OIE
Formal knowledge (ontologies and instance bases)
Semantic Web; Semantic grid; Semantic Web services
(A)IE
Controlled language
CLIE

Figure 8 Closing the language loop.

- the automatic population of ontologies with instances in the text.

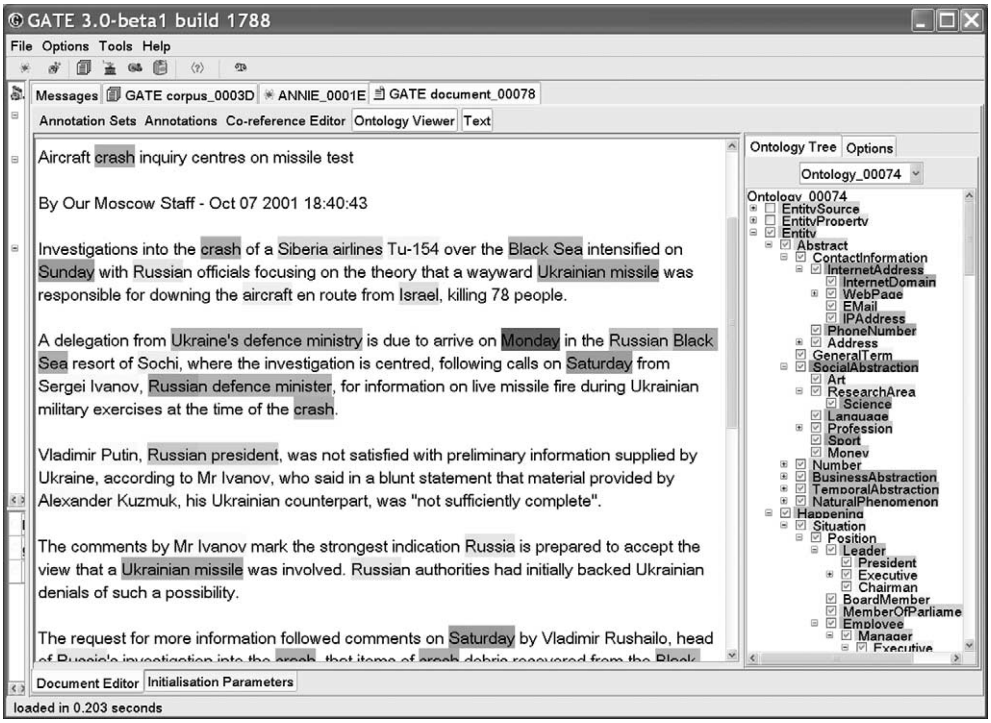
Identification of instances from the ontology If the ontology in question is already populated with instances, the task of an OBIE system may be simply to identify instances from the ontology in the text. Similar methodologies can be used for this as for traditional IE systems, using an ontology rather than a flat gazetteer. For rule-based systems, this is relatively straightforward. For learning-based systems, however, this is more problematic because training data is required. Collecting such training data is, however, likely to be a bottleneck. Unlike traditional IE systems for which training data exist in domains like news texts in plentiful form, thanks to efforts from MUC, ACE and other program, there is a dearth of material currently available for Semantic Web applications. New training data need to be created manually or semiautomatically, which is a time-consuming and onerous task, although systems to aid such metadata creation are currently being developed.

Automatic ontology population In this case an OBIE application identifies instances in the text belonging to concepts in the ontology, and adds these instances to the ontology in the correct location. It is important to note that instances may appear in more than one location in the ontology, because of the multidimensional nature of many ontologies and/or ambiguity, which cannot or should not be resolved at this level.

Example systems The knowledge and information management system (KIM, Popov *et al.*, 2004) is an extendable platform for knowledge management which offers IE-based facilities for metadata creation, storage, and semantic-based search. It also includes a set of front-ends for online use that offer semantically enhanced browsing.

Magpie (Domingue *et al.*, 2004) is a browser add-in that uses IE to facilitate the interpretation of webpages and collaborative sense-making. It annotates web pages with metadata in a fully automatic fashion and needs no manual intervention. It automatically populates an ontology from relevant web sources, and can be used with different ontologies. The principle behind it is that it uses an ontology to provide a very specific and personalized viewpoint of the web pages the user wishes to browse.

The SemTag system Dill *et al.* (2003) uses IE to perform large-scale semantic annotation with respect to the TAP ontology. It first performs a lookup phase annotating all possible mentions of instances from the TAP ontology. In the second, disambiguation phase,



Aircraft crash inquiry centres on missile test

By Our Moscow Staff - Oct 07 2001 18:40:43

Investigations into the crash of a Siberia airlines Tu-154 over the Black Sea intensified on Sunday with Russian officials focusing on the theory that a wayward Ukrainian missile was responsible for downing the aircraft en route from Israel, killing 78 people.

A delegation from Ukraine's defence ministry is due to arrive on Monday in the Russian Black Sea resort of Sochi, where the investigation is centred, following calls on Saturday from Sergei Ivanov, Russian defence minister, for information on live missile fire during Ukrainian military exercises at the time of the crash.

Vladimir Putin, Russian president, was not satisfied with preliminary information supplied by Ukraine, according to Mr Ivanov, who said in a blunt statement that material provided by Alexander Kuzmuk, his Ukrainian counterpart, was "not sufficiently complete".

The comments by Mr Ivanov mark the strongest indication Russia is prepared to accept the view that a Ukrainian missile was involved. Russian authorities had initially backed Ukrainian denials of such a possibility.

The request for more information followed comments on Saturday by Vladimir Rushailo, head of Russia's investigation into the crash, that items of crash debris recovered from the Black

Figure 9 Ontology-based IE in action.

SemTag uses a vector-space model to assign the correct ontological class or determine that this mention does not correspond to a class in TAP.

Acknowledgments

Thanks to Valentin Tablan, Diana Maynard, and Kalina Bontcheva for important contributions to this article. The author was partly supported by EPSRC grant GR/N15764/01 (AKT) and by EU grants SEKT, PrestoSpace and KnowledgeWeb.

See also: Computational Language Systems: Architectures; Human Language Technology; Language Processing: Statistical Methods; Named Entity Extraction; Natural Language Processing: Overview; Natural Language Processing: System Evaluation; Text Retrieval Conference and Message Understanding Conference.

Bibliography

ACE (2004). 'Annotation guidelines for entity detection and tracking (EDT).' Available at http://www.ldc.upenn.edu/Projects/ACE/.

Appelt D (1999). 'An introduction to information extraction.' *Artificial Intelligence Communications 12(3),* 161–172.

ARPA (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6).* San Francisco, CA: Defense Advanced Research Projects Agency, Morgan Kaufmann.

Bechhofer S, van Harmelen F, Hendler J, Horrocks I, McGuinness D L, Patel-Schneider P F & Stein L A (2003). OWL web ontology language reference. Tech. rep., W3C Proposed Recommendation 15 December 2003, http://www.w3.org/TR/2003/PR-owl-ref-20031215/.

Berners-Lee T (1999). *Weaving the web.* London: Orion Business Books.

Boguraev B, Garigliano R & Tait J (1995). Editorial. *Natural Language Engineering 1(Part 1),* 1–7.

Bontcheva K (2004). 'Open-source tools for creation, maintenance, and storage of lexical resources for language generation from ontologies.' In Lino M T *et al.* (eds.) *Proceedings of 4th language Resources and Evaluation Conference (LREC'04).* Lisbon, Portugal. Paris: ELDA.

Cardie C (1997). 'Empirical methods in information extraction.' *AI Magazine 18(4),* 65–80.

Chinchor N A (1998). 'Overview of Proceedings of the seventh message understanding conference (MUC-7)/MET-2.' In *Proceedings of the Seventh Message Understanding Conference (MUC-7).* Fairfax, VA. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.

Cowie J & Lehnert W (1996). 'Information extraction.' *Communications of the ACM 39(1),* 80–91.

Cunningham H (1999). *'Information extraction: a user guide (revised version).'* Research memorandum CS–99–07, Sheffield: Department of Computer Science, University of Sheffield.

Cunningham H (2002). 'GATE, a general architecture for text engineering.' *Computers and the Humanities 36,* 223–254.

Cunningham H & Scott D (eds.) (2004). *Natural Language Engineering 10(3–4).* Special issue on software architecture for language engineering.

Daelemans W & Osborne M (eds.) (2003). CoNLL-2003, 7th Conference on computational natural language learning. Edmonton, Canada.

Davies J, Fensel D & van Harmelen F (eds.) (2002). *Towards the semantic web: ontology-driven knowledge management.* Chicester: Wiley.

Day D, Aberdeen J, Hirschman L, Kozierok R, Robinson P & Vilain M (1997). 'Mixed-initiative development of language processing systems.' In Grishman R (ed.) *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97).* Washington DC: Association for Comoputational Linguistics.

Dill S, Eiron N, Gibson D, Gruhl D, Guha R, Jhingran A, Kanungo T, Rajagopalan S, Tomkins A, Tomlin J A & Zien J Y (2003). 'SemTag and Seeker: bootstrapping the Semantic Web via automated semantic annotation.' In Hencsey G & White B (eds.) *Proceedings of the Twelfth International World Wide Web Conference.* Budapest, Hungary.

Domingue J, Dzbor M & Motta E (2004). 'Magpie: Supporting browsing and navigation on the semantic web.' In Nunes N J & Rich C (eds.) *Proceedings ACM Conference on Intelligent User Interfaces (IUI).* Funchal, Portugal: ACM. 191–197.

Fensel D, Hendler J, Wahlster W & Lieberman H (eds.) (2002). *Spinning the Semantic Web: bringing the World Wide Web to its full potential.* Cambridge, MA: MIT Press.

Gaizauskas R & Wilks Y (1998). 'Information extraction: 'beyond document retrieval.'' *Journal of Documentation 54(1),* 70–105.

Grishman R (2001). 'Adaptive information extraction and sublanguage analysis.' In Kushmeric N (ed.) *Proceedings of Workshop on Adaptive Text Extraction and Mining at Seventeenth International Joint Conference on Artificial Intelligence.* WA: Seattle. http://www.smi.ucd.ie/ATEM2001/proceedings/toc.html.

Grishman R & Sundheim B (1996). 'Message understanding conference-6: a brief history.' In Tsujii J (ed.) *Proceedings of the 16th International Conference on Computational Linguistics.* Copenhagen: ICCL.

Maynard D, Bontcheva K & Cunningham H (2003). 'Towards a semantic extraction of named entities.' In Nikolov N, Bontcheva K & Angelova G (eds.) *Recent advances in natural language processing.* Borovetz, Bulgaria, Amsterdam: John Benjamins.

Pazienza M T (ed.) (2003). *Information extraction in the Web era.* Berlin: Springer-Verlag.

Popov B, Kiryakov A, Kirilov A, Manov D, Ognyanoff D & Goranov M (2004). 'KIM – semantic annotation platform.' *Natural Language Engineering, 10(3–4)*, 375–392.

SAIC (1998). *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html.

Sundheim B (ed.) (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Columbia, MD: Morgan Kaufmann.

Information Structure in Spoken Discourse

T Kotschi, Institut für Romanische Philologie, Berlin, Germany

When producing spoken discourse, it is important for speakers, in order to transmit their message properly, to give it a structure that may support the appropriate understanding of the message on the part of the listener. This is normally done by marking some words or word groups as more important than others and by 'packaging' the message into units of different sizes. The listener, on the other hand, will perceive words or word groups as being more or less important and will detect different types of boundaries that will enable him or her to interpret the utterance on the basis of a division into units to be identified as paragraphs, subparagraphs, sentences, clauses, phrases, and others. The message as uttered by the speaker and interpreted by the listener will thus have a structure that can be characterized in terms of important information and of boundaries between units of different extension. This type of structure is commonly referred to as the information structure of a discourse.

Although the essence of these considerations seems to be uncontroversial, there are nevertheless divergent opinions on what may be considered to belong to information structure, and it is not always clear which categories should be used for the description of its different forms of manifestation. The reasons for this may be that the concept of 'important information' is interpreted differently, descriptions may be based on divergent theoretical assumptions, or either the sentence (or an equivalent linguistic expression) or the discourse may be in the focus of description. Regarding information structure in (spoken or written) discourse, research on its forms and functions is confronted with the problem that a discourse is a complex system of structural spheres within which information structure has to be assigned its theoretically and methodologically adequate place.

For the current purpose, it is assumed that some of the problems implied by the subject of this article may be adequately treated by referring to a global, modular approach to discourse analysis. Following an introduction to some frequently used basic concepts (topic, comment, active concept, semiactive concept, new concept, focus and presupposition), the central aspects of the subject are presented based on a distinction between elementary and complex organization forms in discourse, particularly between information structure and topical organization. Then, the phenomenon of 'punctuation,' which is an inextricable part of the subject under discussion, is discussed.

Basic Concepts

The concepts most frequently used in research work on information structure in discourse can be categorized into three different points of view.

Topic and Comment

First is the distinction between 'topic' and 'comment.' In using these terms allowance is made to the fact that in uttering a (minimal) discourse unit the speaker 'says something about something'; in other words, there is something that has to be regarded as the already established 'matter of current concern' about which new information is added. The added information is named 'comment,' whereas the information that has already been established and thus can serve as an anchoring point for the new information is designated as 'topic.' It is this 'aboutness' relation that connects the two categories. As Lambrecht (1994: 127) stated,

> A referent is interpreted as the topic of a proposition if in a given discourse the proposition is construed as being about this referent, i.e. as expressing information which is relevant to and which increases the addressee's knowledge of this referent.

In the discourse, *I saw John yesterday. He was angry*, the pronoun *he* in the second sentence refers to the topic (it is the 'topic expression'), whereas *was angry* designates the comment that is about this topic. A distinction has to be drawn between sentence topic and discourse topic. A sentence topic (*he* in the cited example) does not necessarily have to be congruent with the grammatical subject of the sentence in