Information Extraction with GATE

Angus Roberts
Recap

- Installed and run GATE
- Language Resources - LRs
  - documents
  - corpora
- Looked at annotations
- Processing resources - PRs
  - loading
  - running
Outline

- Introduction to Information extraction
- Example systems
- Hands on tour of ANNIE
  - Build ANNIE step by step
  - Interlude on multilingual IE
  - Introduce JAPE grammars
  - Introduce co-reference
What is information extraction?
IE is not IR

IR pulls **documents** from large text collections (usually the Web) in response to specific keywords or queries. You analyse the **documents**.

IE pulls **facts** and **structured information** from the content of large text collections. You analyse the **facts**.
With traditional query engines, getting the facts can be hard and slow

Where has the Queen visited in the last year?
Which places on the East Coast of the US have had cases of West Nile Virus?

Which search terms would you use to get this?

How to specify you want someone’s home page?

IE returns information in a structured way
IR returns documents containing the relevant information somewhere
IE as an alternative to IR

- IE returns knowledge at a much deeper level than traditional IR
- Constructing a database through IE and linking it back to the documents can provide a valuable alternative search tool
- Even if results are not always accurate, they can be valuable if linked back to the original text
What does IE extract?

- History: MUC
- Entities
- Relations
  - See JAPE, ML sessions
- Events
What is it used for?

☐ An enabling technology for many others

○ Text Mining
○ Semantic Annotation
○ QA
○ Decision Support
○ Rich retrieval and exploration
○ ...
Types of IE systems

- Deep vs shallow
- Knowledge Engineering vs Machine Learning
  - Supervised
  - Unsupervised
  - Active learning
Knowledge Engineering

- rule based
- developed by experienced language engineers
- make use of human intuition
- require only small amount of training data
- development can be very time consuming
- some changes may be hard to accommodate

Learning Systems

- use statistics or other machine learning
- developers do not need LE expertise
- require large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus
Entity Recognition: the cornerstone of IE

- Traditionally, Identification of proper names in texts, and their classification into a set of predefined categories of interest
- Persons
- Organisations (companies, government organisations, committees, etc)
- Locations (cities, countries, rivers, etc)
- Date and time expressions
- Various other types as appropriate to the application
Why is NE important?

- NE provides a foundation from which to build more complex IE systems

- Relations between NEs can provide tracking, ontological information and scenario building

- Tracking (co-reference) “Dr Smith”, “John Smith”, “John”, “he”

- Ontologies “Athens, Georgia” vs “Athens, Greece”
Typical NE pipeline

- Pre-processing (tokenisation, sentence splitting, morphological analysis, POS tagging)
- Entity finding (gazetteer lookup, NE grammars)
- Coreference (alias finding, orthographic coreference etc.)
- Export to database / XML
Ryanair announced yesterday that it will make Shannon its next European base, expanding its route network to 14 in an investment worth around €180m. The airline says it will deliver 1.3 million passengers in the first year of the agreement, rising to two million by the fifth year.

- **Entities**: Ryanair, Shannon
- **Mentions**: it=Ryanair, The airline=Ryanair, it=the airline
- **Descriptions**: European base
- **Relations**: Shannon base_of Ryanair
- **Events**: investment(€180m)
Examples
HaSIE

- Application developed with GATE, which aims to find out how companies report about health and safety information

- Answers questions such as:
  - “How many members of staff died or had accidents in the last year?”
  - “Is there anyone responsible for health and safety?”

- IR returns whole documents
sustainability management system. BAA has received a RoSPA gold award for occupational safety for the fourth year running. The award is given only if a consistently good or continuously improving performance can be demonstrated over a four-year period. The accident frequency ratio for construction projects was 0.4 (0.49) per 100,000 hours worked, less than one third of the national accident frequency rate in the construction sector. The company is running a "One in a Million?" campaign to raise safety consciousness and standards in construction and reduce the accident frequency rate still further to one for every million man hours worked. We have no higher priority than the safety and security of the passengers, staff and organisations that use our airports. In order to ensure that our systems and practices are continually assessed and upgraded, we work...
Obstetrics records

- Streamed entity recognition during note taking
  - Interventions, investigations, etc.
- Has to cope with terse, ambiguous text and distinguish past events from present
- Entirely on Gazeteers and JAPE
- Used upstream for decision support and warnings
1:30pm

Cx: 3cm, contractions q2-3 min. FHR: reassuring, reactive.

4:00pm

BP: 140/90.

PV: 6cm, 60%; -1; soft consistency, anterior position; cephalic; intact membranes; no vaginal bleeding.

Contractions: 3/10min; regular; moderate

On urinalysis: Protein > 300mg

BP before 20 weeks gestation: 120/80

Plan: monitor Vital Signs by protocol for elevated BP

5:15pm
Multiflora


Ranunculus acris var. latiscutus Beck

Stems erect from short caudex or rhizome, never rooting nodally. Hispid, strigose, or glabrous, base not bulbous. Roots never tuberous. Basal leaf blades pentagonal in outline, deeply 3-5-parted, 1.8-5.2 x 2.7-9.8 cm, segments 1-2 X-lobed or -parted, ultimate segments narrowly elliptic or oblong to lanceolate, margins toothed or lobulate, apex acute to rounded. Flowers: receptacle glabrous; sepals spreading, 4-6(-9) x 2-5 mm, hispid; petals 5, yellow, 8-11(-17) x 7-13 mm. Heads of achenes globose, 5-7(-10) mm wide; achenes 2-3 X 1.8-2.4 mm, glabrous, margin forming narrow rib 0.1-0.2 mm wide; beak persistent, deltate, usually with lip short or long, straight or curved, subulate, 0.2-1 mm, 2n = 14.
Indictment.

William Mills, of the Parish of St. Sepulchres, was indicted for stealing a dark grey Gelding, value 12 l. out of the Grounds of George More, Esq; on the 5th of October last. It appear'd, That the Horse was lost out of the Prosecutor's Grounds at Newark Trens, and sold by the Prisoner at the G in Smithfield, and he not being able to give an Account how he came by it, was found Guilty of the Indictment.

Laurance Singleton, Mary Singleton, and _bert, were indicted for breaking the W house of Joseph Wives, and stealing thence 16 Foot Wels out, Blank 60 Foot of Wainscot, and Foot of Deal, on the 29th of September last. It appear'd, that an Evidence who Swore, he saw him being brought in them half a Yard long) and burn them at Singleton's House; which not being the Cause for an Indictment, they were acquitted.

Andrew James, (a little Boy) of the Parish of St. Dunstun in the West, was indicted for breaking the W house of George Mac, on the 8th instant. It was prov'd that the Kerchief taken upon him; whereupon he was found Guilty to the Value of 10 d.

Mary was indicted for Assaulting ) with infection to on the 2nd of November last. It appear'd by the Prosecutor, that the Child was to be School'd, the Child being sick, the Child
رقمي: 51-7 (في🏾) 0 ادى نادي غروتيتا الإيطالي اهتمامه بضم مهاجم ستنجي.

و كانت لانديةة برايسلا الإنجليزية وكريفا الديكروز تواجه كريستيان سوكادا الإسبان اعترفت عن رغبتها في ضم غريفيش الذي تقدر قيمة انتقاله بحوالي 61 مليون دولار.

ووقع رابيفيل لاعب وسط منتخب تشيكي ونادي التشكيل الإسباني الذي هبط إلى الدرجة الثانية عقدا انتقاله للنادي الفرنسي لمدة 3 سنوات من دون أن تعفي قيمة الصفقة.

وكان رافيل (72 عاما) واحداً من منتخب بلاده في كأس الأمم الأوروبية الأخيرة. لكنه خاض 31 دقيقة فقط في المباريات الثلاث التي خاضها منتخب بلاده في البطولة لأنها خرج من الدور الأول.

ووقع الكروات ميلان رابيفيل مهاجم يرتدي الإيطالي عقدا لمدته 3 سنوات مع فريق 0 في غريفيش التشكيل.

وبلغت قيمة انتقال رابيفيل (72 عاما) نحو 61 مليون دولار.

ولعب رابيفيل 31 مباراة دولية مع منتخب بلاده وكان في صفوف نادي هايديدوك سبيلين الكروات قبل انتقاله إلى إيطاليا.
My name is Anil Roy. I live in Lancaster. My father lives in Liverpool. My father’s name is Rajesh Roy. Lancaster University is my place of study.
What is ANNIE?

- ANNIE is a vanilla information extraction system
- NER and simple coreference
- Primarily developed over newswire
- A set of core PRs:
  - Tokeniser
  - Sentence Splitter
  - POS tagger
  - Gazetteers
  - Named entity tagger (JAPE transducer)
  - Orthomatcher (orthographic coreference)
Core ANNIE components

Input: URL or text

Document format (XML, HTML, SGML, email, )

Gate Document

Unicode Tokeniser

FS Gazetteer Lookup

Sentence Splitter

Hepple POS Tagger

Character Class Sequence Rules

Lists

JAPE Sentence Patterns

Brill Rules Lexicon

Semantic Tagger

Ortho Matcher

Pronominal Coreferencer

JAPE IE Grammar Cascade

JAPE Grammar

Gate Document XML dump of IE Annotations

Output:
Hands on exercises

- We will build ANNIE in several steps

- For each step
  - Introduce the concepts
  - Briefly describe what you need to do
  - Time to follow the instructions hands on

- The first few steps will be very short “see one, do one”

- Later, we will give brief descriptions only, and time for hands on will get longer
Hands-on (1) – the corpus

- Create a new corpus
- Populate it from
  - `hands-on-resources/ie/business`
- Set:
  - extensions to xml
  - encoding to windows-1252
    - Important for currencies!
- Take a look at some documents and their annotations
Tokeniser and sentence splitter

- Tokenisation based on Unicode classes
- Language-dependent or -independent tokenisation
- Declarative token specification language, e.g.:
  
  "UPPERCASE_LETTER" LOWERCASE_LETTER"* >
  
  Token; orthography=upperInitial; kind=word

Sentence splitter

- The default one finds sentence based on tokens
- A faster regular expression splitter is also available
Hands on (2.1) – a simple application

- From the left hand pane, load a couple of processing resources, with default parameters:
  - ANNIE English Tokeniser
  - ANNIE Sentence Splitter

- Create a new Application
  - Use a Corpus Pipeline

- Double click the application to open it in a tab

- On the left, choose each of the loaded resources, and place in the right hand selected resources list in this order
  - Tokeniser
  - Sentence Splitter
In the middle of the tab, set the corpus

Click the “Run this application” button

- Also available from menu bars and context menus

Examine the document

- Annotations have been added to the “default” set
- Look at the annotations and their features

What happens if you run the app a few more times?

Add a Document Reset PR to your application

- Create a new processing resource
- Add it to the start of the application

Now run the app a few times and examine
POS tagger and morphological analysis

- Hepple POS tagger
- Java implementation of Brill's transformation based tagger
- Trained on WSJ
- Default ruleset and lexicon can be modified manually
- Penn Treebank tag set
- Morphological analyser
  - Not usually part of ANNIE
  - Flex based
  - Rules for regular verbs, and some common irregular verbs
Hands on (3) – shallow lexico-syntactic features

- Add an ANNIE POS Tagger to your app
- Add a GATE Morphological Analyser after the POS Tagger
  - This may not be in your list of available resources
  - If not, first open the plugins dialog
    - File > Manage creole plugins
    - Make sure that the “Tools” directory is set to “Load Now”
  - It should now appear in the list of available resources
- Examine the features of the Token annotations
  - New features of Category and root have been added
Multi-lingual IE
Language plugins

Languages supported:

- German
- French
- Italian
- Arabic
- Cebuano
- Chinese
- Hindi
- Romanian

Varying degree of sophistication and functions

- see user guide for details
Language Independent PRs

- Unicode tokeniser
- Sentence splitter
- Gazetteer PR
  - but do localise the lists
- Orthomatcher
### Multilingual components

- **Stemmer plugin**
  - Consists of a set of stemmer PRs for:
    - Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Portuguese, Russian, Spanish, Swedish
  - Requires Tokeniser first (Unicode one is best)
  - Language is init-time param, which is one of the above in lower case

- **TreeTagger**
  - A language-independent POS tagger
GATE uses standard (and imperfect) Java rendering engine for displaying text in multiple languages.
Displaying multilingual data

All visualisation and editing tools use the same facilities.

In the image, there is a screenshot of a software interface, likely related to natural language processing (NLP) tools. The interface shows a text editor with annotations and highlight options, indicating the capability to handle and visualize multilingual data.
Editing multilingual data

- Java provides no special support for text input (this may change)
- GATE Unicode Kit (GUK) plugs this hole
- Support for defining additional Input Methods; currently 30 IMs for 17 languages
- Pluggable in other applications (e.g. MPI’s EUDICO)
- Can use virtual keyboard or standard layouts over QWERTY
- IMs defined in plain text files
- GUK comes with a standalone Unicode editor
- To run it: bin/ant gukdemo
Gazeteers
ANNIEs Gazetteers are plain text files containing lists of names (e.g. rivers, cities, people, …)

More sophisticated lookup components are also available
  ○ Faster hash gazeteer
  ○ Ontology based gazeteers

Each text gazeteer set has an index file listing all the lists, plus features of each list (majorType, minorType and language)

Lists can be modified either internally using Gaze, or externally in your favourite editor

Generates Lookup results of the given kind

Information used by JAPE rules
ANNIE’s Gazetteer Lists

- Set of lists compiled into Finite State Machines

- 60k entries in 80 types, inc.:
  - organization; artifact; location; amount_unit; manufacturer; transport_means;
  - company_designator; currency_unit; date; government_designator; ...

- Each list has attributes MajorType and MinorType and Language):
  - city.lst: location: city: english
  - currency_prefix.lst: currency_unit: pre_amount
  - currency_unit.lst: currency_unit: post_amount

- List entries may be entities or parts of entities, or they may contain contextual
  information (e.g. job titles often indicate people)
Load a new PR: ANNIE Gazeteer

- Set the listsURL to
  - hands-on-resources/ie/gazeteer/lists.def

Add it at the end of your application

Run the application and examine the Lookup annotations

- majorType and minorType features
- Look for ambiguities e.g. “second” and “cent”
  - in-shell-cirywire-03-aug-2001.xml_00065

In the PR list, left hand pane, double click the Gazeteer to open Gaze.

- Look at the lists. Compare linear definitions with major- and minorTypes
We will right a new entry to find words that might signal changes in share prices

- Right click in the Gaze left hand linear definition list, and insert a new list “change.lst” with a majorType of “change”
- Save the new linear definition
- Edit the “change” list in the right hand pane
- You might include rise, fall, up, down, rose, fell etc. Save afterwards.

Run the application, and check the annotations
A few optional ideas

- Make two change gazetteers with the same majorType as before, but different minorTypes.
  - Changes-up
  - Changes-down

- Add a gazetteer that matches the root of a token, not the string. With this, you could match “rose”, “rise”, “rising” etc. to the single gazetteer entry, “rise”
  - You will need to create a new gazetteer PR
  - Next, create a Flexible Gazeteer PR, and tell that the name of your root gazetteer and the feature to match
  - Add the flexible gazetteer to your pipeline
NE transducer

- Gazeteers find terms that suggest entities, and their context
- These terms may be ambiguous:
  - Mrs. May Jones vs 1st May 2006
  - Mr. Parkinson vs Parkinson's disease
- Handcrafted grammar are used to define patterns over the Lookups and other annotations
  - Disambiguate
  - Combine annotations: numbers, dates, money, names
- JAPE: regular expressions over annotation graphs
Add a new PR: ANNIE NE Transducer

Add it to the end of the application

Run the application

Look at the annotations

The pattern grammars that match annotations from the previous step to find Named Entities are written in JAPE – this is covered in the next session.
Co-reference
Using co-reference

- Coreference will be covered more fully in a separate session

- Orthographic co-reference module matches proper names in a document

- Improves results by assigning entity type to previously unclassified names, based on relations with classified entities

- May not reclassify already classified entities

- Classification of unknown entities very useful for surnames which match a full name, or abbreviations, e.g. [Bonfield] will match [Sir Peter Bonfield]; [International Business Machines Ltd.] will match [IBM]

- A pronominal PR is also available
Hands on (6) - coreference

- Add a new PR: ANNIE OrthoMatcher
- Add it to the end of the application
- Run the application
- In a document view, open the co-reference editor by clicking the button above the text
- Examine the co-references
  - You will find some in-shell-cirywire-03-aug-2001.xml_00065
- Optional: add and try the ANNIE Pronominal Coreferencer
What next?
Creating a new application from ANNIE

- Typically a new application will use most of the core components from ANNIE
- The tokeniser, sentence splitter and orthomatcher are basically language, domain and application-independent
- The POS tagger is language dependent but domain and application-independent
- You may also require additional PRs (either existing or new ones – e.g. morphological analyser
- The gazetteer lists and JAPE grammars may act as a starting point but will almost certainly need to be modified
- We will look at JAPE next
Spare
GATE Developer supports the separation of algorithms, data and use interface.

GATE components are one of three types:
- Language Resources (LRs), e.g. lexicons, corpora
- Processing Resources (PRs), e.g. parsers, taggers
- Visual Resources (VRs), i.e. visualisation, editing

Algorithms are separated from the data:
- the two can be developed independently by users with different expertise.
- alternative resources of one type can be used without affecting the other
Linguistic information in documents is encoded in the form of annotations.

The annotations associated with each document are a structure central to GATE.

Each annotation consists of:
- start offset
- end offset
- a set of features associated with it
- each feature has a name and a relative value (arbitrary Java object, incl. String)

Annotations are grouped in annotation sets.

Documents and corpora also have features, which describe them.
Annnotations Example

Text: Cyndi savored the soup.

Nodes: [0...5...10..15..20]

Annotation spans:

Annotation descriptions:

<table>
<thead>
<tr>
<th>Id</th>
<th>Type</th>
<th>Start</th>
<th>End</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>token</td>
<td>0</td>
<td>5</td>
<td>pos=NP</td>
</tr>
<tr>
<td>2</td>
<td>token</td>
<td>6</td>
<td>13</td>
<td>pos=VBD</td>
</tr>
<tr>
<td>3</td>
<td>token</td>
<td>14</td>
<td>17</td>
<td>pos=DT</td>
</tr>
<tr>
<td>4</td>
<td>token</td>
<td>18</td>
<td>22</td>
<td>pos=NN</td>
</tr>
<tr>
<td>5</td>
<td>token</td>
<td>22</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>name</td>
<td>0</td>
<td>5</td>
<td>type=person</td>
</tr>
<tr>
<td>7</td>
<td>sentence</td>
<td>0</td>
<td>23</td>
<td></td>
</tr>
</tbody>
</table>