Introduction to the GATE language analysis system

Wim Peters
Natural Language Processing Group
Department of Computer Science
University of Sheffield
Some scholarly requirements for computational historical linguistics

• Annotation of texts
  manual/automatic ; stand alone/collaborative

• Analysis types include:
  – Tokenization
  – Part of speech tagging
  – Parsing
  – Morphological analysis
  – Ngram extraction/comparison
  – Identification of cognates between languages
  – Semantic comparison of cognates
Requirements continued

• Visualisation: view textual resources and results of annotation processes
  – Lemma/wordform list; frequency counts
  – Bigrams/trigrams etc.
  – Concordancing
  – Evaluation of results
  – Export results in XML
  – Search text and annotations
GATE

- GATE (Generalised Architecture for Text Engineering)
- is a framework for language processing
- under constant development on the basis of EU and national funding
- includes language processing tools, e.g. pos taggers, parsers for various languages
- tools for visualising and manipulating ontologies
- ontology-based information extraction tools
- evaluation tools
- is freely available (www.gate.ac.uk)
GATE Users

- American National Corpus project
- Perseus Digital Library project, Tufts University, US
- Longman Pearson publishing, UK
- Merck KgAa, Germany
- Canon Europe, UK
- a large number of UK, US and EU Universities
- UK and EU projects include
  - EMILLE: S. Asian languages corpus
  - ETCSL ancient Sumerian corpus
  - Old Bailey: 17th century court reports
  - ACE / TIDES: semantic annotation for Arabic and Chinese
GATE modules I: Language Resources

GATE LRs are documents, ontologies, corpora, lexicons.

Documents / corpora:
• GATE documents loaded from local files or the web;
• Diverse document formats: text, html, XML, email, RTF, SGML.

Multilinguality:
• Unicode compliant
• virtual keyboards for language input
Editing Multilingual Data
Displaying Multilingual Data

Arabic: أنا قادر على أكل الرجاح وهذا لا يلمس.
Armenian: Ուզիքը մեկնում է տուփկի մեկնացումը.
Chinese: 我能吞下玻璃而不伤身体。
Farsi / Persian: من می توانم بدون احساس درد شیشه بخورم.
Georgian: ფილდობით ბრწყინვალე ბრწყინვალე.
Hebrew: אני יכול לאכול עגלה חסרת קרח.
Hindi: मैं कोई खा सकता हूँ, मुझे उस से कोई पीड़ा नहीं होती.
Japanese: 私はガラスを食べられます。それは私が傷つけません。
Korean: 나는 유리를 먹을 수 있어요. 그래도 아프지 않아요.
Marathi: मी काह खाऊ शकतो, मला ते दुःखत नाही.
Pashto: زه شیشه خوری شم، همه ما به جووری.
Sanskrit: काच शक्तिशाली। नोपिवनित्त साम्।
Thai: ฉันกินกระจกได้ แต่ยังไม่ทำให้เจ็บ.
Urdu: مس کانج کھا سکتا ہوں اور محسوس نہ کیリング نیس بونا۔
Yiddish: איך קאָן איך אַלט און אָט טפאַר מיר אַראָפֶן? 

Historical Text Mining Workshop,
Lancaster July 20-21, 2006
GATE modules II: Processing Resources

- Programs that run over texts and add annotations.
- All PRs can handle Unicode data by default.
- Quite a few freely available with GATE, e.g.
  - Tokeniser
  - Sentence splitter
  - Part of speech tagger
  - Morphological analyser
  - Parser
  - Semantic annotator
- New programs can be:
  - created within GATE
  - Integrated into Gate by means of wrappers
- Users are constantly providing new modules.
- The results of running PRs over texts can be exported in XML.
• Quick demo?
Historical Text Mining Workshop,
Lancaster July 20-21, 2006

Northern Ireland Secretary Marjorie Mowlam sat down with members of two outlawed Protestant paramilitary groups and delivered a 14-point statement on why they should reverse a vote they took last weekend to condemn the talks. That vote had thrown the talks’ future into question.

After she left, the prisoners did what she asked. The political party that speaks for them at the negotiating table, the Ulster Democratic Party, announced it was no longer considering boycotting the talks, which are set to resume Monday. Another party affiliated with imprisoned Protestant guerillas, the Progressive Unionist Party, said it would decide on Sunday whether to attend.

The all-party talks, chaired by former U.S. senator George Mitchell (D-Maine), seek a political solution in Northern Ireland between Protestants, most of whom want to remain part of Britain, and Catholics, who want greater political rights, including, for some, a political union with the Republic of Ireland to the south.

Throughout the conflict, the British government has held to the line that it talks to people who renounce violence, not to killers. To many people in Britain, it seemed today that Mowlam was being summoned by men convicted of crimes that include murder and arson.

"We are very unhappy about it," said Dave Roberts, development officer for a Northern Ireland peace group called Families Against Intimidation and Terror. Mowlam spoke directly with terrorists, he said, "which many victims felt was grossly insulting."
Viewing data: identifying patterns in corpora

- ANNIC – ANNotations In Context
- Provides a keyword-in-context-like interface for identifying annotation patterns in corpora
- Concordancing possible not only over text elements, but over any type of annotation that has been added to the text
- \{Token.string="Baroness"\}\{Token.string="Thatcher"\}: find all occurrences of the string “Baroness Thatcher”
- \{Token.category="NN"\} \{Token.category="NN"\}: Find all noun-noun combinations
- \{Person\}: find all Person annotations
Historical Text Mining Workshop, Lancaster July 20-21, 2006

ANNIC example

New Query: (Mention.class &="Person")

Total Found Patterns: 313

Export Patterns XML HTML All Patterns Selected Patterns

Annotation Types: Token Features: All Add Annotation Type

Pattern Text: Conservative vice chairman, said Baroness Thatcher had a "very small"

Token.category
JJ NN NNP NN ...

Mention class
Executive Person Woman

Token.orth
upperinitial lower...

Token

Text: Conservative vice chairman, said Baroness Thatcher had a "very small"

Document
f-extremists-07-oct-2001.xml_0003

Pattern
Iain Duncan Smith on Sunday night signalled his

Right Context
has turned on the group

f-extremists-07-oct-2001.xml_0003

under Thatcherism. Gary Streeter, a Conservative vice chairman

f-extremists-07-oct-2001.xml_0003

drawing a line under Thatcherism. Gary Streeter, a Conservative vice chairman

f-extremists-07-oct-2001.xml_0003

Conservative vice chairman, said Baroness Thatcher had a 'very small'

f-extremists-07-oct-2001.xml_0003

should break with her as Tony Blair had ditched clause 4 –

f-extremists-07-oct-2001.xml_0003

commitment to full-scale socialism. Michael Howard, shadow chancellor, said

f-extremists-07-oct-2001.xml_0003

shadow chancellor, said that Mrs Thatcher had "saved this country was

f-extremists-07-oct-2001.xml_0003

got to move on". Mr Streeter also presaged the move against

f-extremists-07-oct-2001.xml_0003

a multiracial, multicultural society. Mr Howard told the BBC's On

f-extremists-07-oct-2001.xml_0003

Club's deputy chairman, Andrew Rosindell, and Angela Watkinson –

f-extremists-07-oct-2001.xml_0003

Andrew Rosindell, and Angela Watkinson – all of who supported

f-extremists-07-oct-2001.xml_0003

Mr Duncan Smith's leadership bid - have

f-extremists-07-oct-2001.xml_0003

some confusion yesterday morning when Mr Hunter said he was "considering

f-extremists-07-oct-2001.xml_0003

were told to quit by David Maclean, the chief whip.
XYZ was established on 03 November 1978 in London. It opened a plant in Bulgaria in ...
Evaluation metrics and tools

- Evaluation metrics mathematically define how to measure the system’s performance against human-annotated gold standard
- Scoring program implements the metric and provides performance measures
  - for each document and over the entire corpus
  - for each type of semantic class
  - may also evaluate changes over time
- A gold standard reference set needs to be provided – this may be time-consuming to produce
- Visualisation tools show the results graphically and enable easy comparison
Conclusion

- Gate is a text mining tool that contains synchronic language analysis modules that are of direct interest to diachronic linguistic research;
- As a platform architecture, it can incorporate any text processing algorithm;
- Therefore it allows extension and adaptation to any type of research specific historical text mining
- It enables integration of research effort and data sharing.
- Manuals, tutorials and support available.