Entity and Event Recognition in ARCOMEM

University of Sheffield, UK
What is Entity and Event Recognition?

- Entity and Event recognition are two of the main components of an Information Extraction (IE) system.

- IE is about the automatic discovery of new, previously unknown information, by automatically extracting information such as entities and events from different textual resources.

- A key element is the linking together of this extracted information to form new facts or to allow new hypotheses to be explored further.

- For example, entities and events can be used to help find opinions expressed in the text about people, objects, or things that have happened.
Information Extraction with GATE

- In ARCOMEM, we use GATE for performing the information extraction and opinion mining tasks.
- GATE is an architecture for language processing, in development at the University of Sheffield since 2000, and used by hundreds of thousands of researchers, scientists and organisations all over the world.
- It includes components for language processing, e.g. parsers, machine learning tools, stemmers, IR tools, IE components for various languages...
- It also includes tools for visualising and manipulating text, annotations, ontologies, parse trees, etc., and tools for evaluation.
Entity Recognition

- The application for Entity Recognition is based on (modified versions of) ANNIE and TermRaider, which are part of GATE
- It consists of a set of Processing Resources (PRs) executed sequentially over the corpus of documents
  - Document Pre-processing
  - Linguistic Pre-processing
  - Named Entity Recognition
  - Term Extraction
  - RDF generation
Linguistic Pre-Processing

- We use the following PRs:
  - Tokeniser
  - Sentence Splitter
  - Language Identification
  - POS tagger
  - Morphological analyser
- For social media such as twitter, we use a specially adapted version of some of these
Named Entity Recognition

- The following entity types are recognised by the NER component, corresponding to the data model:
  - Persons (e.g. artists, politicians, web 2.0 users)
  - Organizations (e.g. companies, music bands, political parties)
  - Locations (e.g. cities, countries)
  - Dates and times (of events and content publication)
- NER consists of:
  - Gazetteer lookup
  - JAPE grammars
  - Co-reference
In so far as a political party in the United States can "decide" anything, the party decided not to have the fight it needed to have between reality-based Republicans and the other kind. And so the election season, but in disguised form. The party understood with a theory like this. There is no need to explain some of the key phrases.

Given the state of our political parties, I mean those who recognize the danger in trying to make descriptions of the world conform to their wishes. By the "other kind" I mean those who don't. Or: members of the Republican coalition who exhibit certain behaviors.

F.A. Hayek wrote about in 1960. This quotation was dug up by Chris Mooney, author of The Republican War on Science. It is from Hayek's essay, "Why I am Not a Conservative."

Personally, I find that the most objectionable feature of the conservative attitude is its tendency to negatively distort things that are inherently good, particularly those of the
The Greek Crisis: Trichet rejects ECB role as lender of last resort

The Greek Crisis

Tuesday, October 4, 2011
Trichet rejects ECB role as lender of last resort
Financial Times
October 4, 2011

Jean-Claude Trichet has dashed hopes that the European Central Bank will ride to the rescue of the eurozone by pledging to backstop crisis-hit member states.

In one of his last appearances as ECB president, Mr Trichet rejected the idea of the ECB acting as lender of last resort to governments. It was up to eurozone political leaders to restore investor confidence in Europe’s monetary union, he told the European Parliament.

"It is their responsibility, individually and collectively, to ensure financial stability. It is the way Europe has been constructed and it is the way, it seems to all of us, we must proceed. If it is not done by governments it will not be credible," said Mr Trichet, whose non-renewable eight-year term finishes on October 31.

items in the co-reference chain
German Entity Extraction

- System for German NER is similar to the English one
- It uses German-specific resources (TreeTagger for lemmatisation and POS tagging, tailored gazetteers and grammars)
- We use a language identification module (TextCat) to determine which documents or sentences are in German
Entities in the Rock am Ring forum
Term Recognition
Term Recognition in ARCOMEM

- Terms generally consist of important nouns or noun phrases that are not usually proper names (unlike most Named Entities) but reflect the content or topic of the text.

- For example, important terms in a political text might be "government", "minister", "financial situation", "Chinese economy".

- These are used in a similar way to Named Entities, for things like event recognition, topic detection and opinion mining.

- In GATE, we use a tool called TermRaider for finding terms.
TermRaider

- GATE plugin for detecting single and multi-word terms
- Based on a simple web service developed originally in the EU NeOn project
- This version runs in GATE, with visualisation tools, and extended functionality such as new scoring systems
- We have developed an adaptation for German.
- Terms are ranked according to three possible scoring systems:
  - \( \text{tf.idf} = \) term frequency (nbr of times the term occurs in the corpus) divided by document frequency (nbr of documents in which the term occurs)
  - augmented \( \text{tf.idf} \) = after scoring \( \text{tf.idf} \), the scores of hypernyms are boosted by the scores of hyponyms
  - Kyoto domain relevance = document frequency \( \times (1 + \text{nbr of hyponyms in the corpus}) \), Bosma and Vossen 2010
TermRaider: Methodology

• After linguistic pre-processing (tokenisation, lemmatisation, POS tagging etc.), nouns and noun phrases are identified as initial term candidates.
• Noun phrases include post-modifiers such as prepositional phrases, and are marked with head information for determining hyponymy. Nested nouns and noun phrases are all marked as candidates.
• Term candidates are then scored in 3 ways.
• The results can be viewed in the GATE GUI, exported as RDF according to the ARCOMEM data model, or saved as CSV files.
• The viewer can be used to adjust the cutoff parameter. This is used to determine the score threshold for a term to be considered valid.
• Terms can also be shown as a tag cloud.
Term candidates in a document

3.) “For a representative figure among reality-based Republicans, I go with David Frum, the former speechwriter for George W. Bush and a conservative who cannot stomach what has happened to his party. Rather than become a Democrat or claim some sort of ideological conversion, Frum has taken up his pen, as with: When Did the GOP Touch With Reality? Â There he writes:

Few of us have the self-knowledge and emotional discipline to say one thing while meaning another. If we say something often enough, we begin to believe it. We donâ€™t usually delude others until after we have first deluded ourselves. Some of the smartest and most sophisticated people knowâ€”canny investors, erudite authorsâ€”sincerely and passionately believe that President Barack Obama has gone far beyond conventional American liberalism and is willfully and relentlessly driving the United States toward a truly socialist and possibly fascist state.”
Top terms from Greek Financial Crisis corpus

- government agency
- organic solvent
- draconian cut
- transatlantic relationship
- organic solvent in certain paint
- French bank
- achievement of labour movement
- bad debt
- financial system
- class pay
- confidence trick
- draconian cut in benefit
- Chinese economy
- key partner
- EU country
- NATO crimes
- debt crisis
- security policy
- foundation of global security
- future of the transatlantic relationship
Terms can be exported as a tag cloud
Event Recognition
Expression of events

- Events can be expressed by:
  - verbal predicates and their arguments (e.g. “The committee dismissed the proposal”);
  - noun phrases headed by nominalizations (e.g. “economic growth”);
  - adjective-noun combinations (e.g. “governmental measure”; “public money”);
  - event-referring nouns (e.g. “crisis”, “cash injection”).
Rule-based event extraction in GATE

- Recognition of entities and the relations between them in order to find domain-specific events and situations.
- In a (semi-)closed domain, this approach is preferable to an open IE-based approach which holds no preconceptions about the kinds of entities and relations possible.
- Application for event recognition is designed to follow the entity recognition application, so no linguistic pre-processing is necessary.
- Basic approach involves finding event-indicative seed words (e.g. “downturn” might indicate an economic event) and some linguistic relation to existing entities (e.g. “Greece”) in the sentence.
Event extraction application

- The application in GATE consists of the following Processing Resources:
  - event gazetteer
  - verb phrase chunker
  - event recognition JAPE grammars
  - RDF generation
- As with the NE recognition module, a German version is also available, and conditional processors decide, based on the language identified for each sentence, which version to run.
Events annotated in GATE
Machine Learning Applications

- We have also developed machine learning applications for performing both entity and event extraction in GATE.
- These make use of the language detection and linguistic pre-processing components, and then replace the JAPE grammars with a learning algorithm based on annotated training data.
- These provide an alternative mechanism for achieving the same result, but require individual training for each different domain of text we want to apply them to.
More information

- More details on entity, event and opinion mining from a technical point of view are available from http://gate.ac.uk/projects/arcomem/training.html

- Comprehensive training material about using GATE for text mining is available at http://gate.ac.uk/family/training.html