Information Extraction

The IE Application

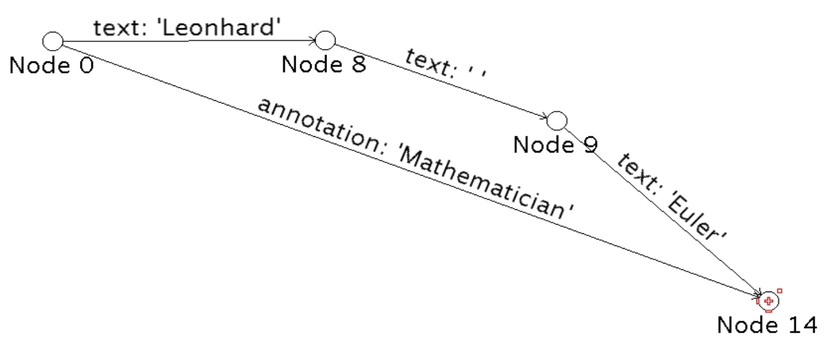
The IE Application used most of the components of the ANNIE system, while also having extended its gazetteer list to identify names of mathematicians and added JAPE grammars `mathematician.jape` and `mathematician_context.jape`, which identify instances of mathematicians and the protagonist's parents in a biography. In addition, the MaxEnt machine-learning annotator, which uses the maximum entropy model to learn which texts to annotate, was trained to recognize the protagonist. Once the annotations were made, the annotated document was outputted in an XML format.

Code snippet: `mathematician_context.jape`

```
(
 ({Token.string == "father"})
 ({Token})?
 ({Token.string == "was"})?
 (
  {FirstPerson}
  ({Mathematician}):lastname
 ):father
)
```

Annotation Graphs

The intuitive representation of named-entity recognition output is to annotate the document, i.e. identify specific phrases as being an instance of a particular object type. Formally, an annotated text may be considered to be a directed acyclic graph, with a linear sequence of nodes that represent specific locations in the text, and where the literal text and the annotations are arcs pointing from a start node to an end node. Naturally, between two consecutive nodes there is a literal text arc, while an annotation arc leading from a start node to an end node means the annotation of the text between the two nodes, i.e. the concatenation of the sequence of text arcs that lead from the start to the end node.



GATE

The General Architecture for Text Engineering (GATE), first proposed by Cunningham, is a comprehensive architecture for the development and execution of NLP applications. GATE's Document class represents a document with annotations as an annotation graph. NLP applications are created as pipelines, i.e. a string of smaller tools known as Processing Resources (PRs) that perform a particular task, such as string tokenising or performing gazetteer lookup. Several useful built-in tools include A Nearly New Information Extraction (ANNIE) system, and a machine learning PR that trains a maximum entropy model for automatically creating annotations.

Prolog Knowledge Base

Document Frames

We propose a compromise representation, document frames, to represent what the machine knows about the text itself, i.e. the fact that a certain string in the document received some annotation by treating a document as an object, and annotations as slots. Hence, we give rise to a simple Prolog scheme for representing annotations: ***Document(Annotation, Text).*** For example,

```
'Euler.html'('Mathematician',
  'Leonhard Euler').
'Euler.html'('Father', 'Paul Euler').
```

Intelligent Querying

The ability to present more complex and intelligent queries to our knowledge base with little difficulty is a great advantage provided by the use of Prolog.

```
| ?- 'Galois.html.xml'('Mathematician', X),
     'Cauchy.html.xml'('Protagonist', X).

X = Cauchy;

no
```

The query asks for the name of the individual, if he exists at all, who appears in Galois' biography while being a protagonist of Cauchy's biography (i.e. himself). Since as seen above, Cauchy appears in Galois' biography, the correct answer is obtained. One could define a subject-matter expert (SME) rule that defines when two documents may be considered to be linked together in this same fashion.

```
link(DocumentA, DocumentB) :-
  Query1 =.. [DocumentA,
    'Mathematician', Mathematician],
  call(Query1),
  Query2 =.. [DocumentB,
    'Protagonist', Mathematician].
  call(Query2).
```

Conclusions

The efficacy of the proposed two-stage process was demonstrated; the pair of applications successfully process natural language documents into a Prolog knowledge base that can be queried intelligently. The advantages of a document-based frame representation were immediately manifest in the ability to easily find out the relationships between documents, and the ability of the Prolog application to work with a GATE IE application designed to work with any subject matter.

Possible directions for further research in the area include the development of a document frame representation that is more expressive and the addition of SME rules that are able to infer even more facts not directly specified in the KB. Use of nondeterministic and/or statistical methods for the analysis of language context is another promising direction of research. Computational power, however, will continue to be a limitation.