Information Extraction of Swedish text

Background

Information Extraction is a Natural Language Processing (NLP) task which is responsible for the extraction of pre-defined information from unstructured text. Potential pieces of information to be extracted include [1]:

- Entities (names, dates/times, etc)
- Relationships between entities
- Events described in the text

Information Extraction techniques have applications in:

- The curation of medical journals
- Military intelligence
- World Wide Web (search engines)

Aim and Objectives

The overall aim of this project is to produce a novel tool which performs information extraction on Swedish text using Python.

The specific research objectives are:

- To review the relevant Information Extraction literature
- To implement the solution as a package using Python
- To evaluate the effectiveness of the solution with training models

Swedish has different grammatical rules to English, and there are no specialised Swedish-based IE systems, which is a problem which my solution is attempting to solve.

Methodology

Natural language is unstructured, with various inconsistent rules which can also be dependent on context. Despite this, these rules can be represented mathematically. All natural language text can be modelled through language models by assigning probabilities, and using patterns to match text against [2].

An example of pattern matching is shown below through Named Entity Recognition (NER):

$GCSE_{[qualification]}$:n var dock inte tillgängliga, och $Dan_{[name]}$ var tvungen att nöja sig med Law. $Tio\ år\ senare_{[time]}$ läser $Dan_{[name]}$ sitt sista år av $Computer\ Science\ BSc_{[qualification]}$ vid $Bangor\ University_{[school]}$ i $Wales_{[country]}$.

Information Extraction produces an output which contains all the relevant pre-defined information, in a structured format:

TAKE-OVER-1:	
Relationship:	TAKE-OVER
Entities:	"Microsoft"
	"Bethesda Softworks"
	"ZeniMax Media"
Company:	Xbox (Microsoft)
Activity:	ACTIVITY-1
Amount:	$750000000000

ACTIVITY-1:	
Activity:	ACQUISITION
Company:	"ZeniMax Media"
Product:	"Bethesda Softworks"
Date:	"Monday"

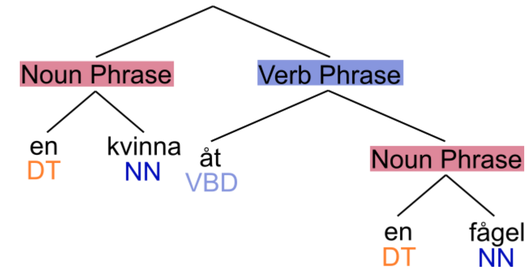
Results

Information extraction for English is a mature aspect of natural language processing, but remains in its infancy for other languages, such as Swedish.

An Information Extraction system is comprised of many different processes, with one being a pattern-based Chinker, which removes a sequence of tokens from a Part-of-Speech tagged chunk [3]:

```
(S
  (NP den/DT lilla/JJ gula/JJ hunden/NN
  (skällde/VBD
  (NP på/PP katten/NN))
```

An example phrasal decomposition tree:



References

[1] Appelt, Douglas E. "Introduction to information extraction." AI Communications 12, no. 3 (1999): 161-172.

[2] Rosenfeld, Roni. "A maximum entropy approach to adaptive statistical language modeling." (1996).

[3] Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.", 2009.

Author: Dan Stoakes
dns18sxx@bangor.ac.uk

PRIFYSGOL BANGOR UNIVERSITY

Supervisor: Dr. W. J. Teahan
w.j.teahan@bangor.ac.uk