1 General information

ANNIE

Gate group[1], Department of Computer Science, University of Sheffield, UK

Status: stable

Languages covered: English

2 Implemented NLP services

ANNIE is an open-source, robust Information Extraction (IE) system. Its output is which relies on finite state algorithms. ANNIE consists of the following main language processing tools: tokeniser, sentence splitter, POS tagger, named entity recogniser.

The named entity recogniser identifies and categorizes entity names (such as persons, organizations, and location names), temporal expressions (dates and times), and certain types of numerical expressions (monetary values and percentages). For this purpose, it uses three types of processing resources: a gazetteer, a part of speech tagger and a rule grammar module. The gazetteer consists of lists such as cities, organizations, days of the week, etc. It not only consists of entities, but also of names of useful indicators, such as typical company designators (e.g. 'Ltd.'), titles, etc. The gazetteer lists are compiled into finite state machines, which can match text tokens. The part of speech tagger attaches morpho-syntactic labels ("noun", "verb", "adjective" etc.) to text elements. The rule grammar component allows the encoding of rules that operate on the output of both the gazetteer and the pos tagger in order to annotate text spans with the relevant named entity types. The text spans and annotations are exported into an RDF ontology, in which the named entity types such as Organization and Person constitute classes, and the text spans instances of these classes.

3 Language resource standards

No language resource standards are required for input.

Output:

The pos tags are Penn Treebank and MAF compliant.

The morphological output (e.g. lemmatization) is also MAF compliant.

4 Linguistic data encoding

The service does not assume any data categories as input.

[1] http://www.gate.ac.uk

It will run over a text corpus containing documents from a large variety of formats:
plain text
HTML
SGML
XML
RTF
PDF (not all)
Microsoft Word (not all)