1 General information

GATE English POS tagger and lemmatizer

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

Status: in preparation

Languages covered: English

2 Implemented NLP services

First, the service preprocesses the text input (for required text format see below) with domain- and application-independent techniques.

- Tokenization: the tokeniser splits text into simple tokens, such as numbers, punctuation, symbols, and words of different types (e.g. with an initial capital, all upper case, etc.).
- The sentence splitter segments the text into sentences.
- The tagger produces a part-of-speech tag as an annotation on each word or symbol.
- The lemmatizer produces text annotated with lemma information for nouns and verbs in xml format.

3 Language resource standards

The output is in the form of xml annotated text. The pos annotations are Penn Tree-bank and MAF compliant.

4 Linguistic data encoding

The service does not assume any data categories as input.

It will run over a text corpus containing documents from a large variety of formats:

plain text

HTML

SGML

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

XML
RTF
PDF (not all)
Microsoft Word (not all)