Index

A
AKTiveMedia, 45
Application programming interface (API), 16

B
Bosque v7.3 corpus, 41–42
Boundary detection. See Sentence boundary detection

C
CoreNLP, 53
cURL/Wget, 51

D
Data generation, 13–15
Dependencies, 31–32
Document object model (DOM), 55
Document society, 1–3
Domain representation, 44–46

E
e-Government. See Information extraction (IE), electronic government
Epic parser, 21–22

G
GATE, 24
Generic relation extraction, 31–32

Global positioning system (GPS), 65–67
Google Maps API, 65–66
Graphical user interface (GUI), 16
GrETUA, 65–66

I
Identifying entities. See also Named entity recognition (NER)
generic named entities, 30
goals, 27
named entity recognition, 28
relation, 30–32
website pages, 28
wikipedia categories, 29

Information extraction (IE)
approaches, 6–7
architecture, 8
challenges, 5–6
documents and information retrieval, 5
electronic government
documents, 59–60
goals, 58–59
maps, 65–67
natural language documents, 59
ontology creation, 61–65
semantic information queries, 67–68
extraction tasks, 5
genetic pipeline, 71
identifying entities, 27
information extraction systems, 8–9
key concepts, 71
natural language texts, 5
Information extraction (IE) (cont.)
NLP, 72
OBIE, 7, 39
performance measures, 7–8
process overview, 13–15
iSentenizer, 16–17

K
Knowledge representation (KR), 3–4

L
Lemma/lemmatization, 17–18

M
MaltParser, 23–24, 44
Markov models. See TreeTagger algorithm
Millikan, Robert Andrews, 51–52
Morphological analysis
lemma/lemmatization, 17–18
POS tagging, 17–18
Stanford POS Tagger, 19
SVMTool, 19–20
tools, 18–19
TreeTagger algorithm, 19–20
word stemming, 17–18

N
Named entity recognition (NER)
identifying entities, 27–28, 30
natural language processing, 4, 43–44
Natural language processing (NLP)
arithmetic, 39–40
Bosque sentences, 41–42
cohesive data representation, 23
definition, 4
documents, 59
GATE, 24
HAREM categories, 43–44
named entity recognition,
43–44
NLP steps, 13–14
NLTK interfaces, 24
POS tags, 43
sentence splitting, 43
Stanford NLP, 23–24
syntactic parsing, 44
tasks in, 4–5

Natural Language Toolkit (NLTK)
Punkt, 16
text processing libraries, 24
NER. See Named entity recognition

O
OBIE. See Ontology-Based Information
Extraction (OBIE)
Ontologies
annotation, 64–65
classes, 62
composition, 61–62
definition, 4
perspective, 61, 63
reasons, 61
relational database, 32–33
relevant information, 38–39
Ontology-Based Information Extraction
(OBIE), 7
approaches, 33–34
arbitrary ontology, 39
architecture, 39–40
identifying entities, 27
IE process, 33
OpenNLP, 53
OWL. See Web Ontology Language (OWL)

P
Parse tree, 31
Part of speech (POS)
natural language processing, 43
relation extraction, 31
Stanford POS Tagger, 19
tags, 17–18, 43
Portuguese, 40–41
Probabilistic context-free grammar (PCFG)
parser. See Epic parser
Prototype implementation. See State-of-the-art
tools
Punkt approach, 15–16

R
r-millikan.txt.xml, 54–55
Relation extraction
approaches, 30–31
dependencies, 31–32
entities, 30
methods, 31
named entity recognition, 30
patterns, 31
POS tags, 31
types, 30
Resource Description Framework Schema (RDFS), 4, 33

S
SAXON, 56–58
Semantic extraction and integration
creation, 46–48
ontology relation, 45–46
partial syntactic structures, 47
use of, 48
Semantics
advantages of semantic search, 3
architecture, 39
knowledge representation, 3–4
linguistic expressions, 3–4
queries, 68
Sentence boundary detection.
See Tokenization
Sentence splitting, 43
Simple Knowledge Organization System (SKOS), 63
Software architecture, OBIE
domain representation, 39–40
NLP component, 39
semantic extraction and integration, 39–40
SPARQL query, 68
Stanford CoreNLP, 53–54
Stanford NLP, 23–24
Stanford POS Tagger, 19
StanfordParser, 22–23
State-of-the-art tools
architecture instantiation, 40–41
domain representation, 44–46
natural language processing, 41–44
semantic extraction and integration, 45–48
Support vector machine (SVM) tool, 19–20
Syntactic parsing
computational intensive, 20
Epic parser, 21–22
goals, 20
MaltParser, 23–24
natural language processing, 44
phrase structure grammar, 20–21
StanfordParser, 22–23
syntax theories and techniques, 21
TurboParser, 23–24

T
Tokenization
boundary detection, 15
definition, 15
representative tools, 16–17
tools, 15–16
TreeTagger algorithm, 19–20
TurboParser, 23–24
Tutorials
applications, 72
scenarios, 51
software tools, 53
syntactic parsing, 58
target document, 54–58
tools setup, 53–54
wikipedia page, 51–52

U
Uniform resource locator (URL), 51–52
Unstructured documents, 1–2

W
Web Ontology Language (OWL), 4, 33
Wget command, 59–60
Wikipedia page, 51–52
Word stemming, 17–18

X
XML stylesheet language for transformations (XSLT), 55–57