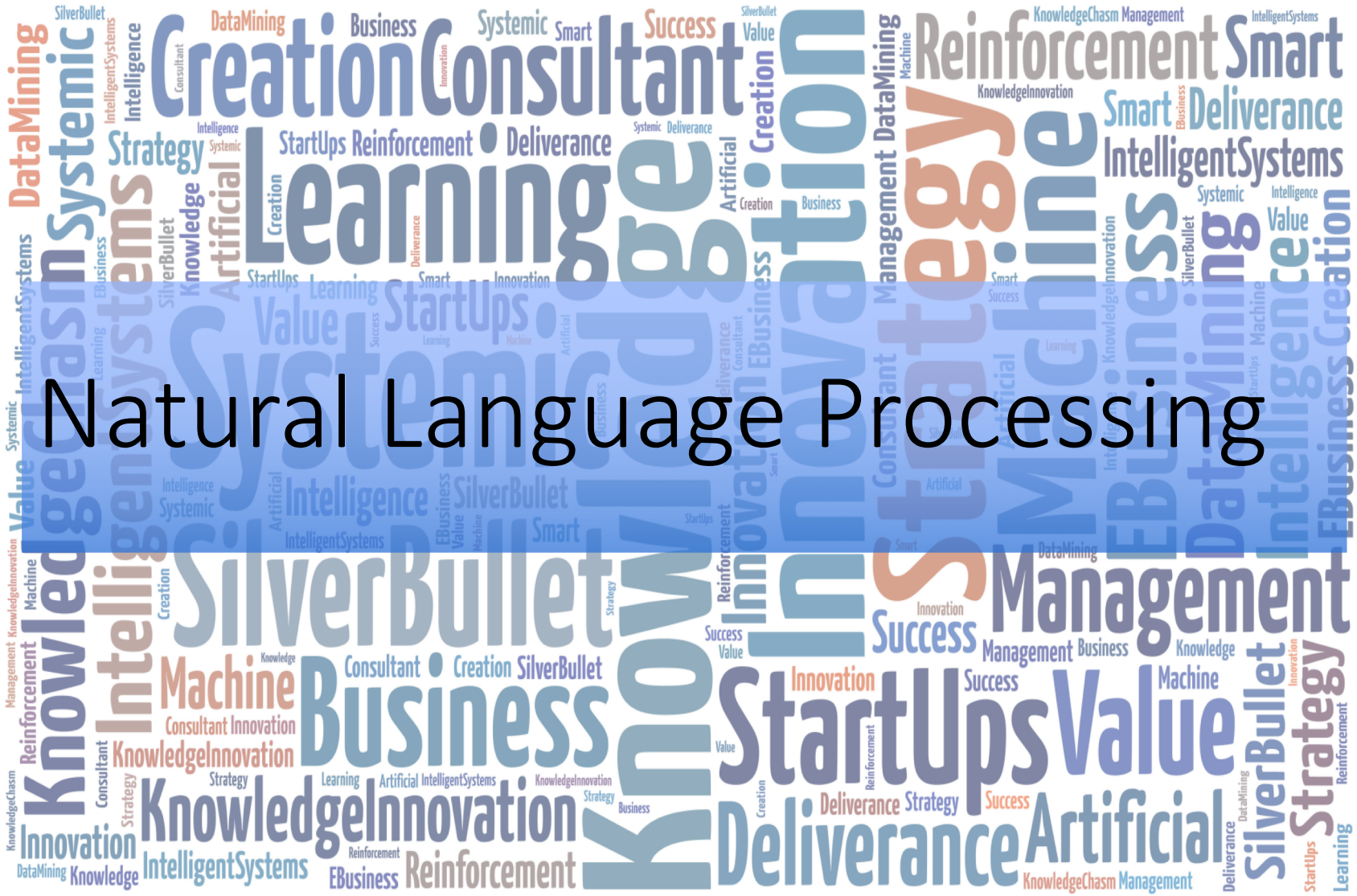
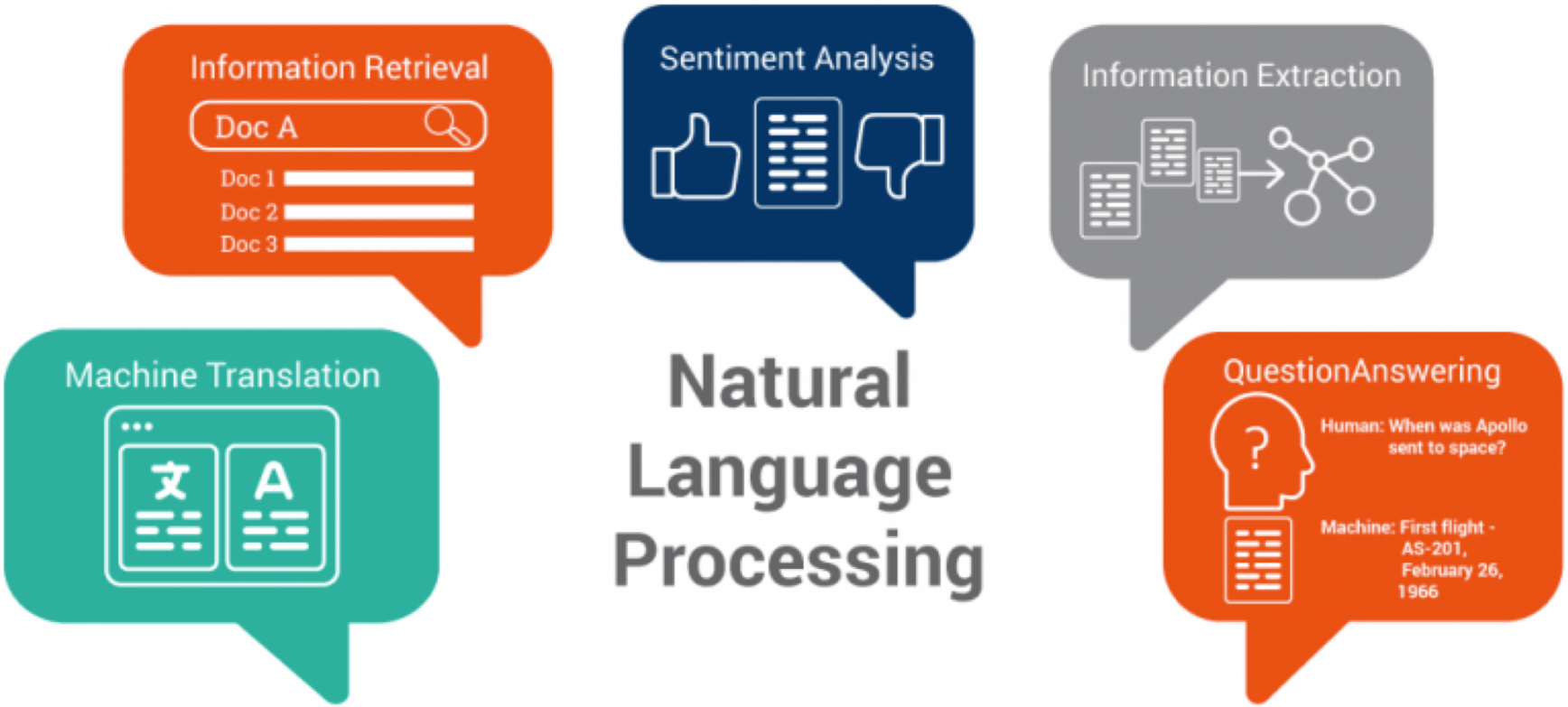
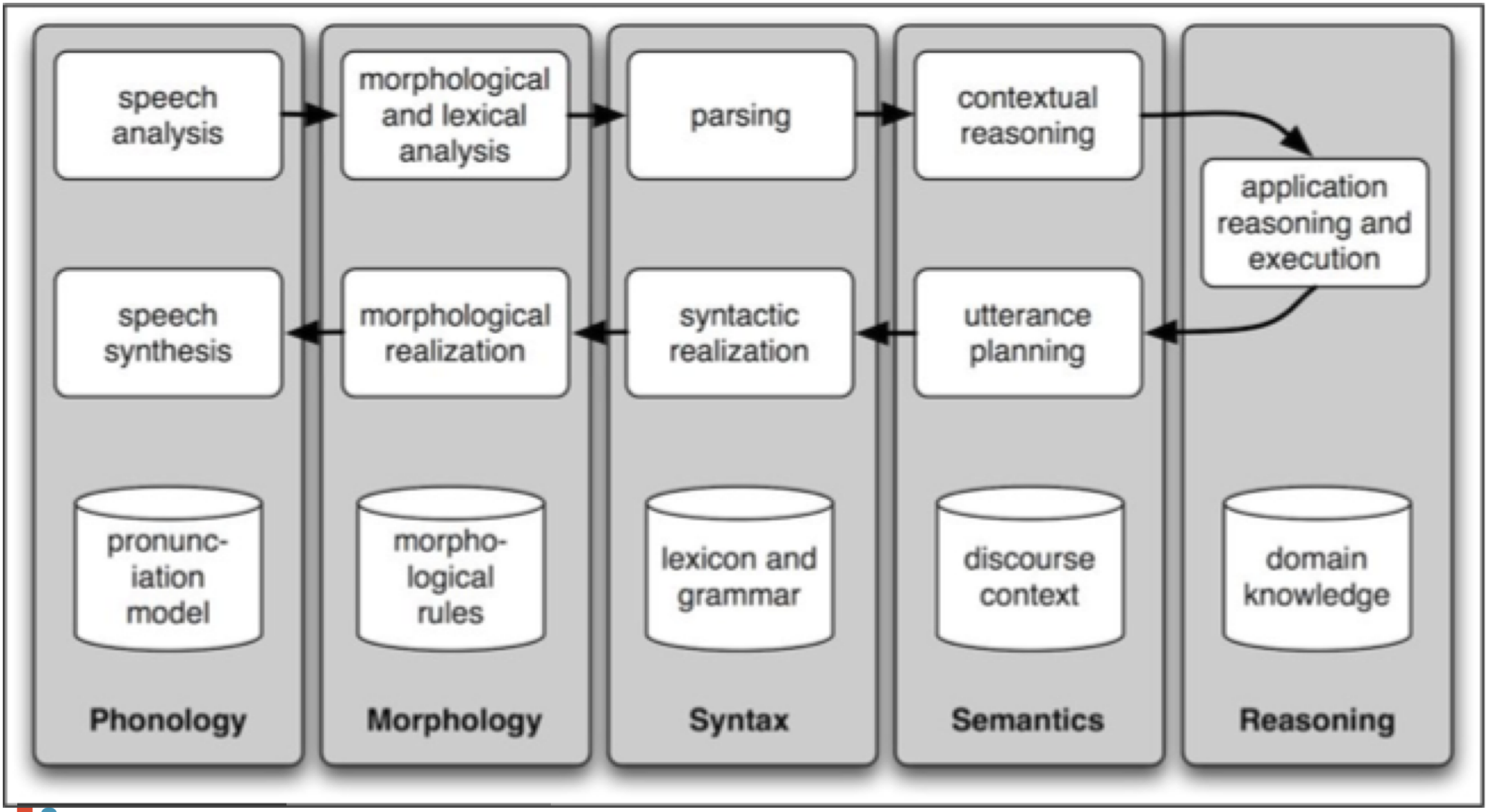

Natural Language Processing
iknowlation
knowledge embedded in innovation

Sub – areas of NLP



Applications

- Summarization
- Reference Resolution
- Machine Translation
- Language Generation
- Language Understanding
- Document Classification
- Author Identification
- Part of Speech Tagging
- Question Answering
- Information Extraction
- Information Retrieval
- Speech Recognition
- Sense Disambiguation
- Topic Recognition
- Relationship Detection
- Named Entity Recognition



Steps in NLP

Steps in NLP

- **Phonetics, Phonology**:
  - how Word are prononce in termes of sequences of sounds
- **Morphological Analysis**:
  - Individual words are analyzed into their components and non word tokens such as punctuation are separated from the words.
- **Syntactic Analysis**:
  - Linear sequences of words are transformed into structures that show how the words relate to each other.
- **Semantic Analysis**:
  - The structures created by the syntactic analyzer are assigned meanings.
- **Discourse integration**:
  - The meaning of an individual sentence may depend on the sentences that precede it and may influence the meanings of the sentences that follow it.

Morphology

- The study of the forms of things, words in particular.

Consider pluralization for English:

- Orthographic Rules: puppy → puppies
- Morphological Rules: goose → geese or fish

Major parsing tasks:
stemming, lemmatization and tokenization.

Tokenization

Process of breaking text into defined segments (usually using regexes or simple delimiters).

By Stanford Def :- Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens , perhaps at the same time throwing away certain characters, such as punctuation.

Input: Friends, Romans, Countrymen, lend me your ears;

Friends	Romans	ears	lend	me

Stop words

Dropping common terms: stop words

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

English has common list of around 512 words

Normalization

USA -> U.S.A

Car -> Cars

Care -> Caring

Color (British) -> Colour (American)

Stemming and lemmatization

The goal of both stemming and lemmatization is to **reduce inflectional forms** and sometimes **derivationally related forms** of a word to a common base form.

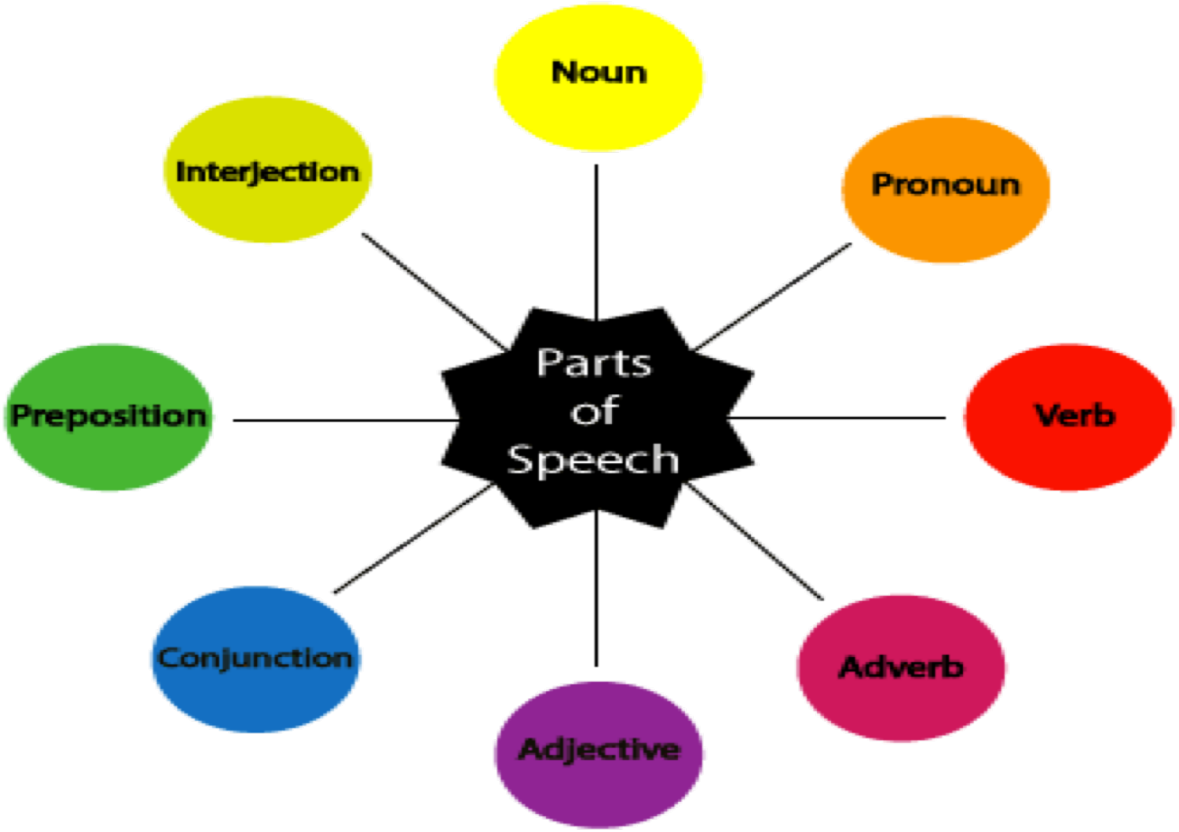
For instance:

the boy's cars are different colors
→
the boy car be differ color

Natural Language Processing Terminology

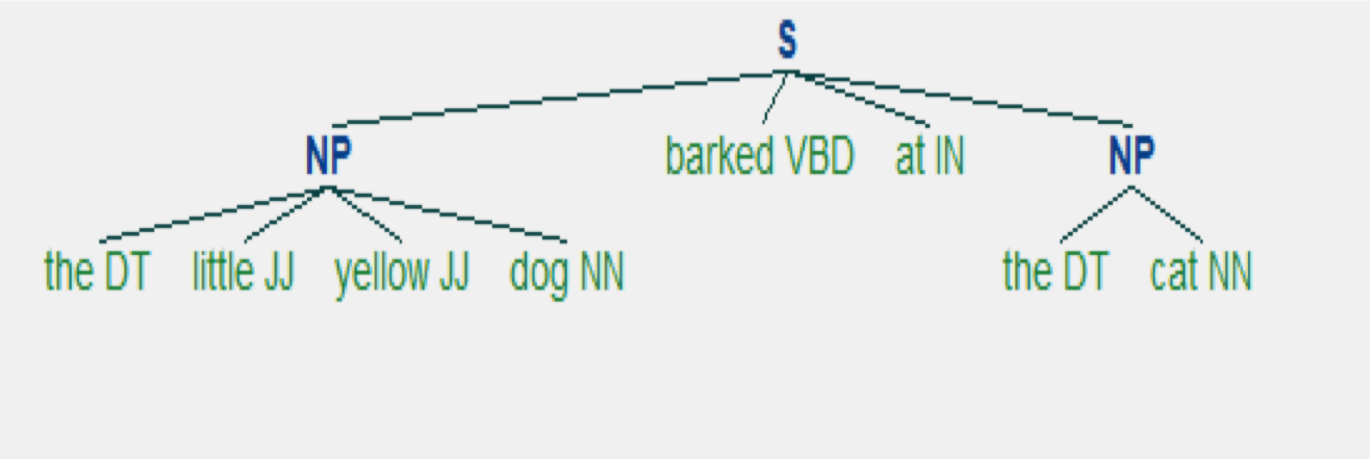
Tokenization,
Corpus or Corpora,
Stemming,
Bag of Words,
Stop Words,
Tf-idf,
Disambiguation,
Topic Models ,
Word Boundaries

Syntactic Analysis



Syntactic Analysis

The study of the rules for the formation of sentences.



Major tasks:

chunking, parsing, feature parsing, grammars

POS Tagging (Penn Treebank - 36 tags

POS tagging is a supervised learning solution that uses features like the previous word, next word, is first letter capitalized etc.

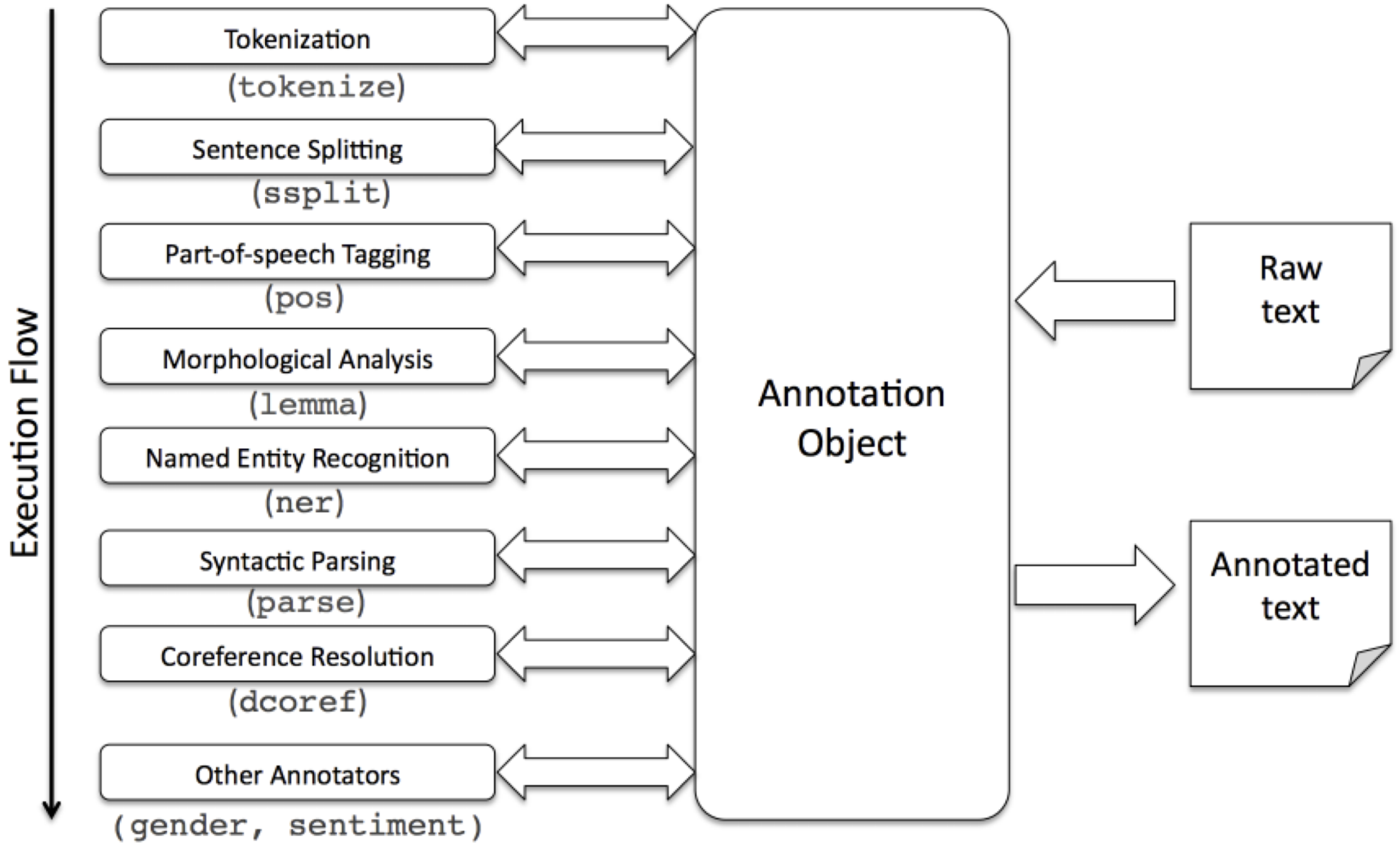
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Chunking

Chunking is a process of extracting phrases from unstructured text. Instead of just simple tokens which may not represent the actual meaning of the text, its advisable to use phrases such as “South Africa” as a single word instead of ‘South’ and ‘Africa’ separate words.

Chunking works on top of POS tagging

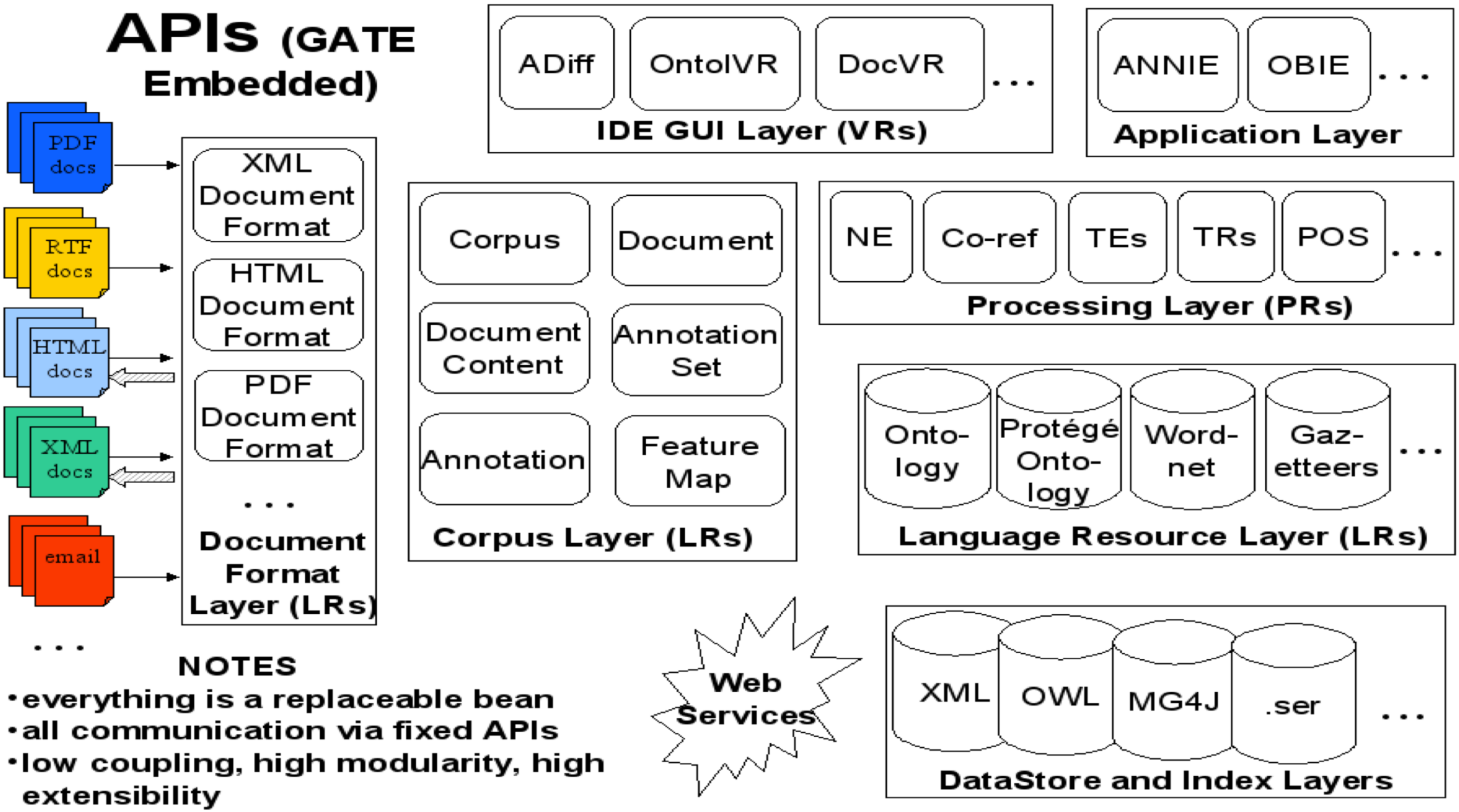
https://medium.com/greyatom/learning-pos-tagging-chunking-in-nlp-85f7f811a8cb


Execution Flow
Tokenization
(tokenize)
Sentence Splitting
(ssplit)
Part-of-speech Tagging
(pos)
Morphological Analysis
(lemma)
Named Entity Recognition
(ner)
Syntactic Parsing
(parse)
Coreference Resolution
(dcoref)
Other Annotators
(gender, sentiment)
Annotation Object
Raw text
Annotated text

Stanford CoreNLP Pipeline

Process of breaking text into defined segments (usually using regexes or simple delimiters).

("annotators",
"tokenize,ssplit,pos,lemma,ner,parse,depparse,coref,kbp,quote");


APIs (GATE Embedded)
PDF docs
RTF docs
HTML docs
XML docs
email
. . .
XML Document Format
HTML Document Format
PDF Document Format
. . .
Document Format Layer (LRs)
ADiff
OntoIVR
DocVR
. . .
IDE GUI Layer (VRs)
ANNIE
OBIE
. . .
Application Layer
Corpus
Document
Document Content
Annotation Set
Annotation
Feature Map
Corpus Layer (LRs)
NE
Co-ref
TEs
TRs
POS
. . .
Processing Layer (PRs)
Onto-logy
Protégé Onto-logy
Word-net
Gaz-etteers
. . .
Language Resource Layer (LRs)
NOTES
•everything is a replaceable bean
•all communication via fixed APIs
•low coupling, high modularity, high extensibility
Web Services
XML
OWL
MG4J
.ser
. . .
DataStore and Index Layers

Hidden Markove Model (POS tags)