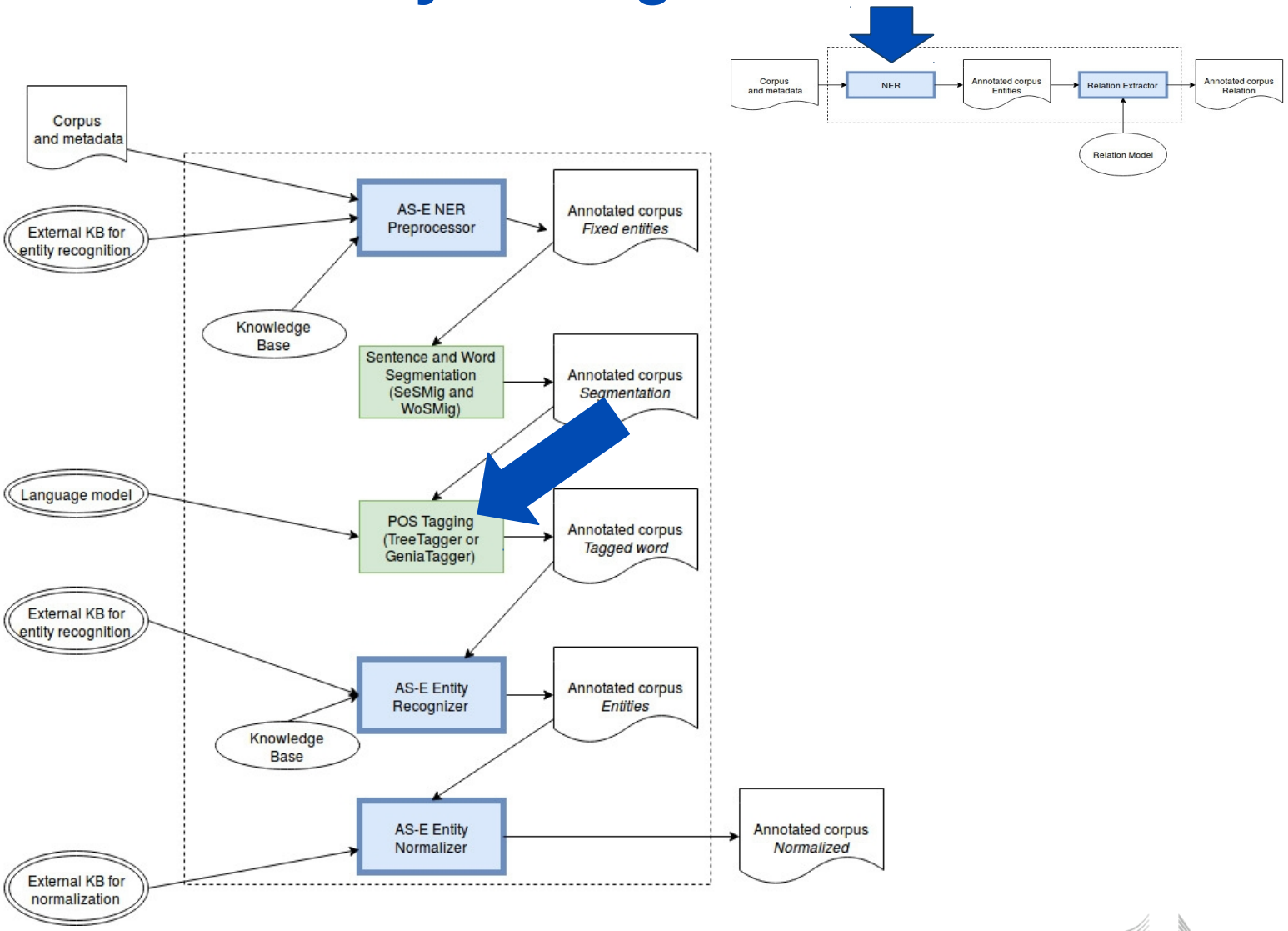
Text-mining methods used for information extraction in plant scientific papers

3.From words to entity recognition

NER: Named Entity Recognition



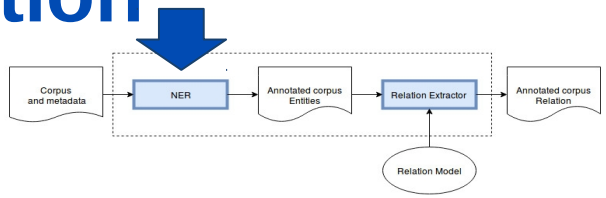
Word Tagger

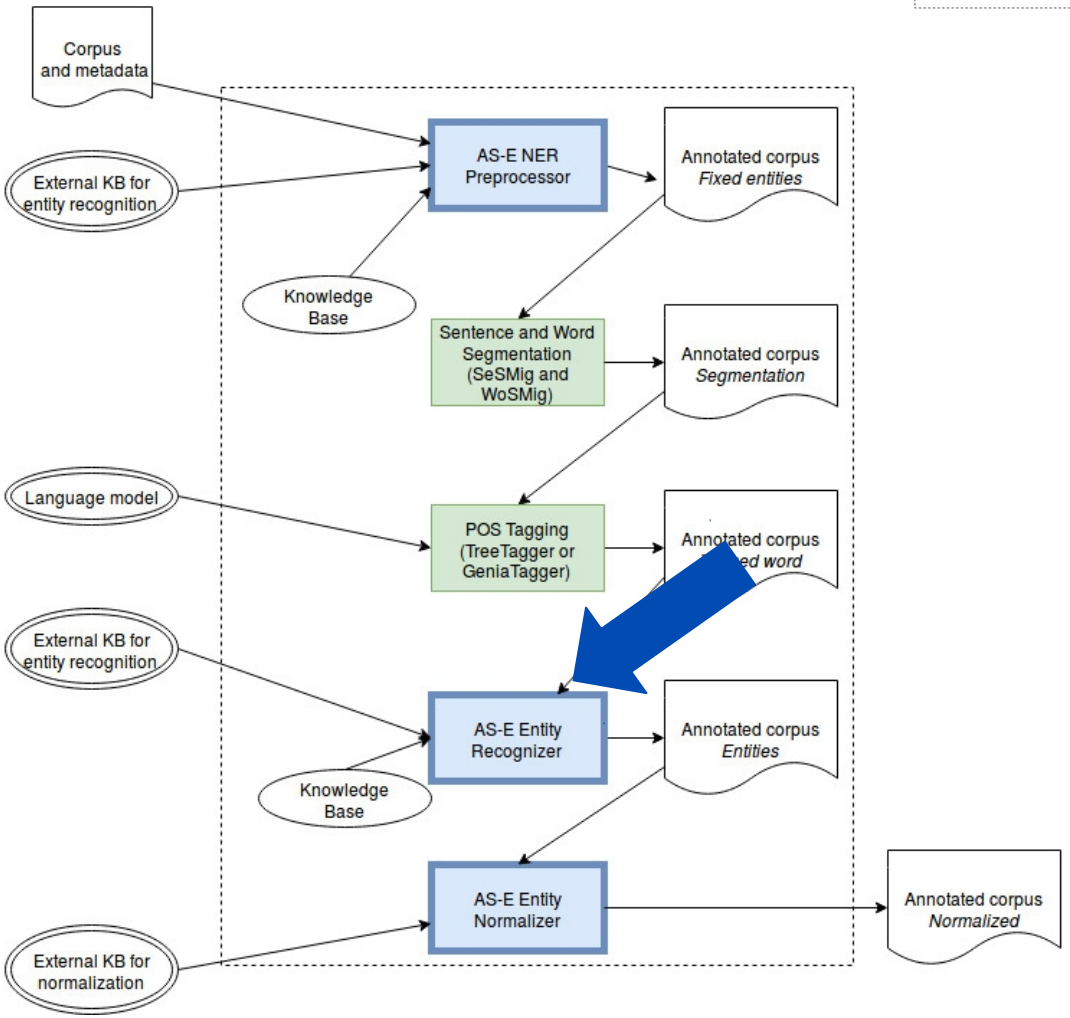
For each words, different tools could be applied to tag different features of the word such as the **form, pos, lemma, stemma** …
e.g. "...expressed during early embryogenesis "

e.g. with NLTK tools :

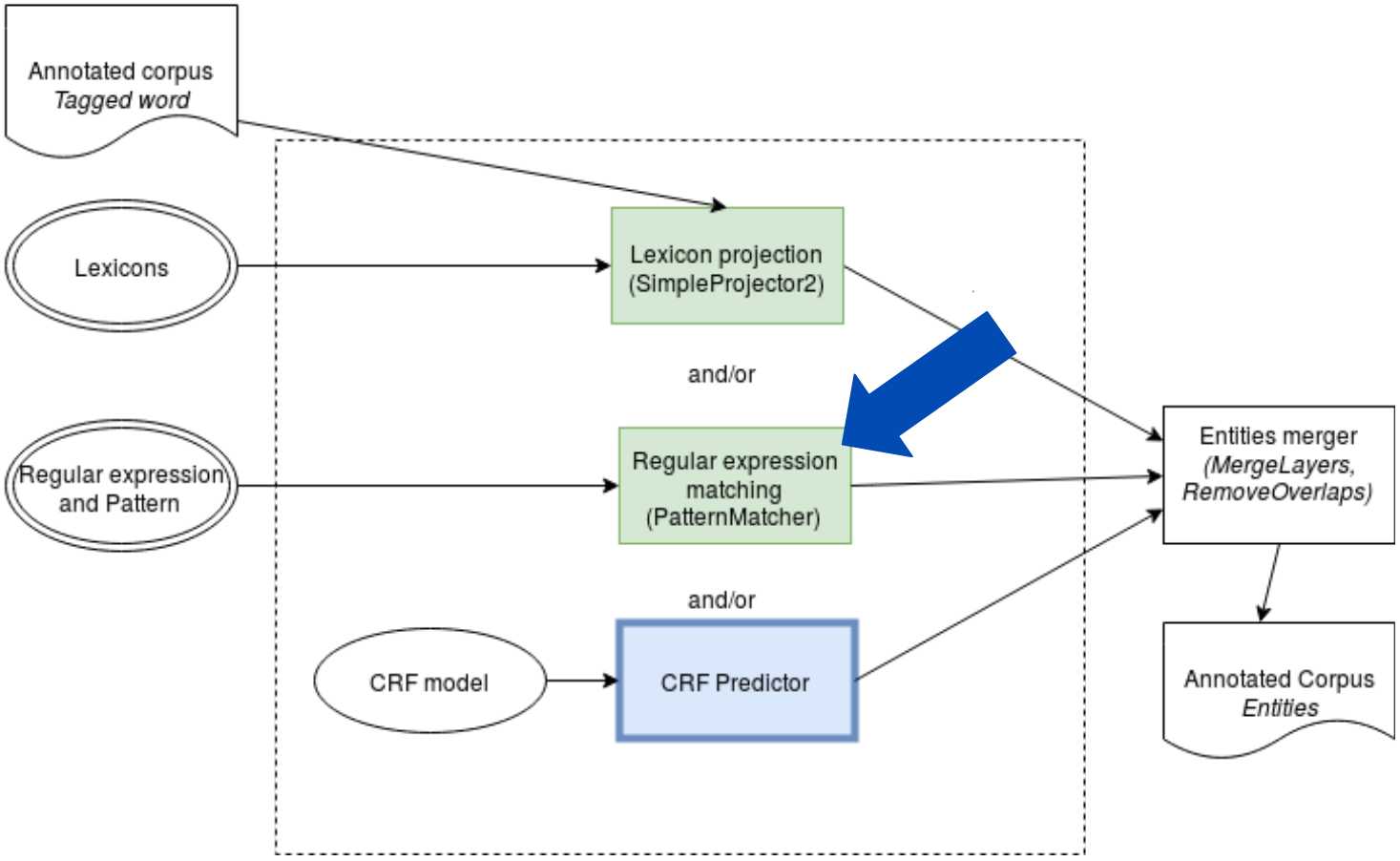
- Form : expressed during early embryogenesis
- Pos Tagging : expressed|VBN during|IN early|JJ embryogenesis|NN
- WordNet Lemma : express during early embryogenesis
- Snowball Stemmer : express dure earli embryogenesi

NER: Named Entity Recognition





Entity Recognizer



Entity Recognition with Regular Expressions

Interpretation of DNA sequences in text :
A sequence of A, C, T, G, N ≥ 3 characters, could begin with/finish by 5’ or 3’

Definition: A short DNA sequence that corresponds to a **binding site** for a protein
That could be expressed as target of genes (AFL target) or DNA sequence (AACA, (C/T)ACGTGGC , CCATTTTTGG …)

e.g. http://arabidopsis.med.ohio-state.edu/AtcisDB/bindingsites.html

Entity Recognition with Regular Expressions

Example of Boxes :
5'ACGTACGTAATG'3
AAAAAAACG
(C/G/T)ACGTG(G/T)(A/C)

Regular expression matching with these boxes
(5.{0,3})?((A|C|G|T|N|V|\(|\)){3,})+(.3{0,3}.)?

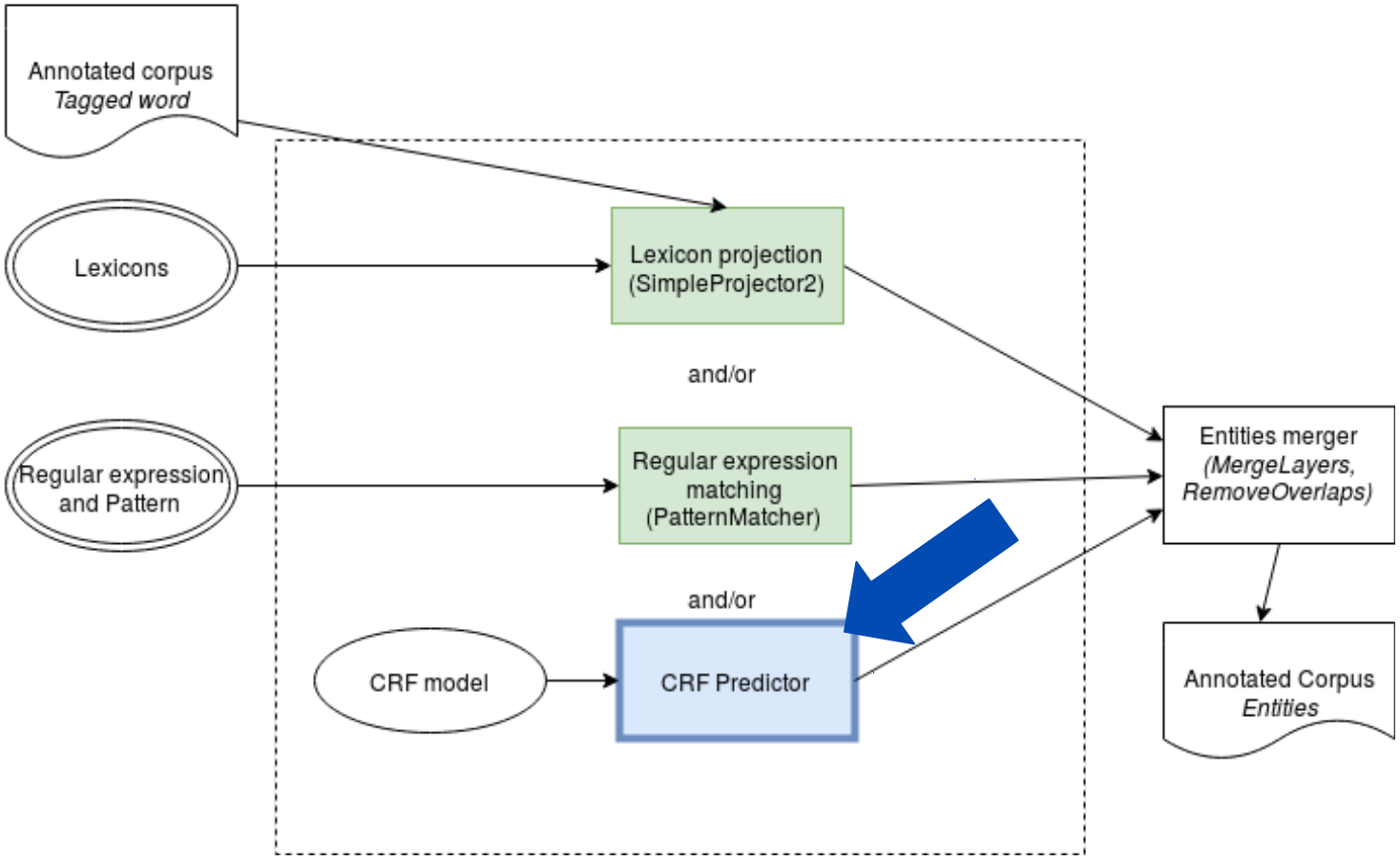
Explanation of this Regular Expression :
https://regex101.com/r/nhgHxb/1

Entity Recognition with Pattern Matching

Prerequisite : Detection of some entities with lexicon, regular expressions
Aim to predict entities that follows patterns

e.g. : [Gene] transcript
[Gene] level
[Gene] mRNA
} [RNA]

Entity Recognizer



Entity Recognition with Machine Learning

Prerequisite : Examples from manual annotations
from the challenge BioNLP-ST SeeDev

Aim to predict entities by learning features from manual annotation examples and a mathematical algorithm

NER: Named Entity Recognition