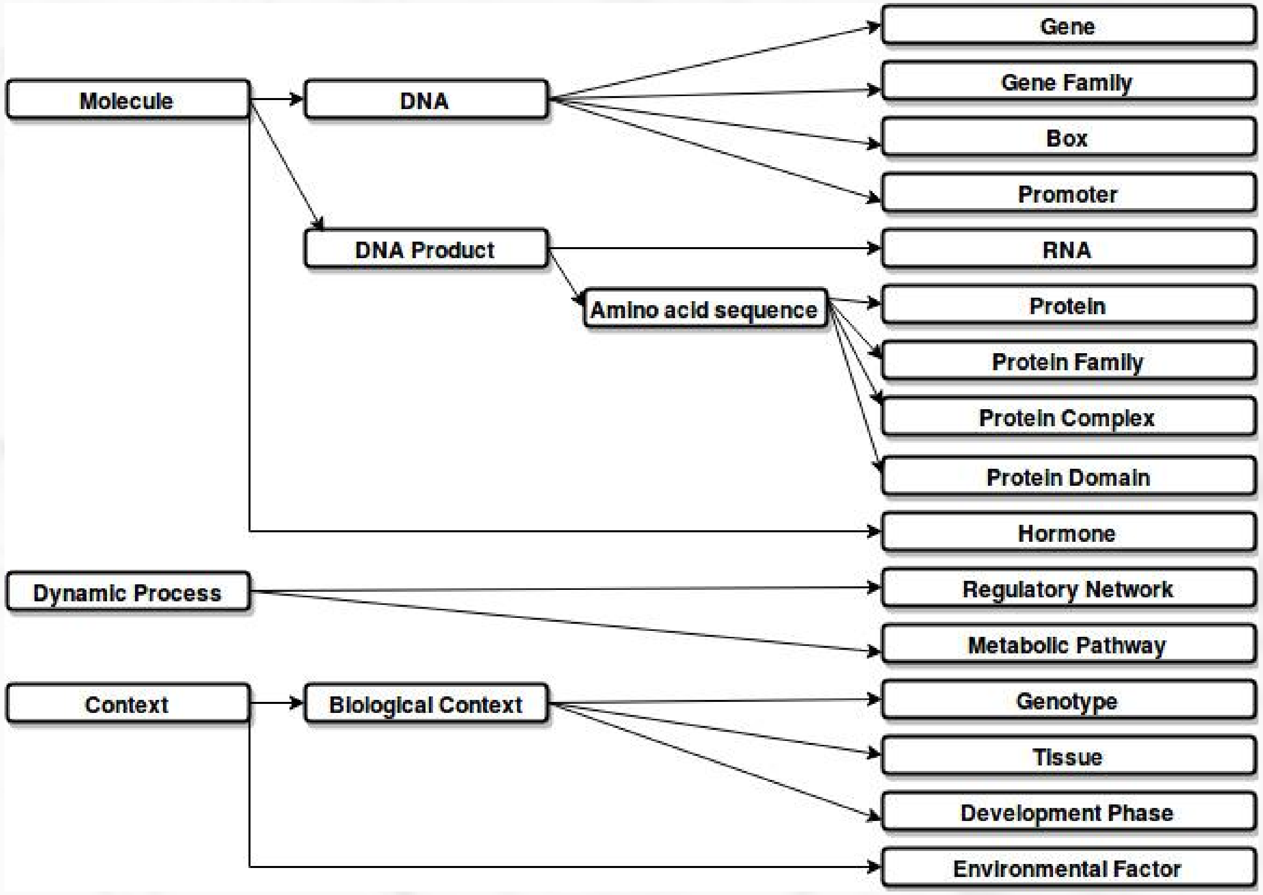
Text-mining methods used for information extraction in plant scientific papers

2. From text to words

Knowledge model of entities

16 entities



CC-BY

NER & Text segmentation

The processing order is important :

To keep structures of words and sentences that do not respect the "classical" structure and could be segmented by these processes.

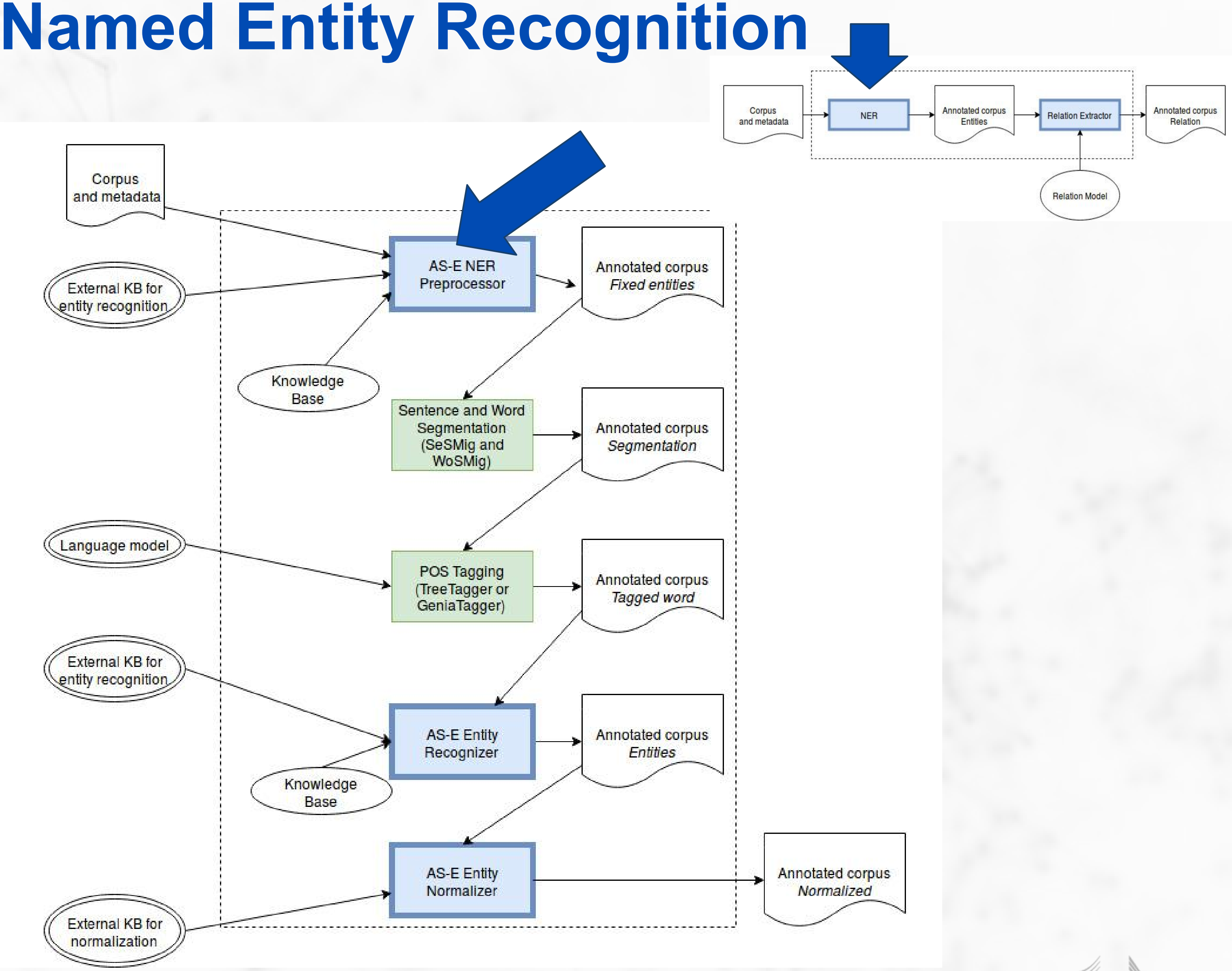
NER & Text segmentation

e.g. In the plant use-case, this is the case for gene names which may have number, punctuation inside the term:
CRP810_1.3
1,2-DIOXYGENASE

In our case, we detect some entities using lexicons before the text segmentation

CC-BY

NER: Named Entity Recognition


Corpus and metadata
NER
Annotated corpus Entities
Relation Extractor
Annotated corpus Relation
Relation Model
Corpus and metadata
External KB for entity recognition
Knowledge Base
AS-E NER Preprocessor
Annotated corpus *Fixed entities*
Sentence and Word Segmentation (SeSMig and WoSMig)
Annotated corpus *Segmentation*
Language model
POS Tagging (TreeTagger or GeniaTagger)
Annotated corpus *Tagged word*
External KB for entity recognition
Knowledge Base
AS-E Entity Recognizer
Annotated corpus *Entities*
External KB for normalization
AS-E Entity Normalizer
Annotated corpus *Normalized*

Entity Recognition

There are two types of entities to recognize

- Named Entities (e.g. author names, genes, geographical locations …) denoted by rigid designators : NER Preprocessor

- Complex entities (e.g. development phase, pathway , tissue…) expressed in natural language : Entity Recognizer

NER Preprocessor

Aim to annotate entities which are defined by rigid designators

Example: Person Name, Bibliographical quote , Gene , Protein and their families, RNA

Tools : Projection with Lexicon
Regular Expressions

NER Preprocessor : Name detection

Named Entities Recognition :
Very useful for tagging person, authors, organisations, geographical localisation…
Tools : Stanford Named Entity Tagger

e.g: Stanford Potential tags: Organization Location Person

"American Society of Plant Biologists MUCILAGE-MODIFIED4 Encodes a Putative Pectin Biosynthetic Enzyme Developmentally Regulated by APETALA2, TRANSPARENT TESTA GLABRA1, and GLABRA2 in the Arabidopsis Seed Coat1 Tamara L. Western2, Diana S. Young, Gillian H."

online : http://nlp.stanford.edu:8080/ner/process

NER Preprocessor with Regular Expressions

It may be useful to detect bibliographical references in text, avoiding some errors in the detection of other entities.

Bibliographical references are generally of the form : (Authora A., et al 2000)

NER Preprocessor with Regular Expressions

An example of pattern that could match with similar bibliographical reference is:

```
\(([\p{L}-\s\.,]+\s\d{4}[a-zA-Z]?[\s;]*)+?\)
```

that matches with

(Meinke et al., 1994)
(Baumlein et al., 1994; Parcy et al., 1997)
(Leung and Giraudat, 1998)

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

NER Preprocessor with Lexicon

Using a lexicon in "learning by rote" for Detection of the **sequence of characters** as entities

e.g. : **words to be excluded** from future predictions : stopwords

Lexicon

a
about
above
after
again
against
all
am
an
and
any
are
aren't
as
at
be
because
been...

CC-BY

NER Preprocessor with Lexicon

Using a lexicon in "learning by rote" for Gene / Protein detection :
Lexicon provided by TAIR annotation

Lexicon

AP2
CRP810_1.3
1,2-Dioxygenase
LEC 2

NER Preprocessor with Lexicon

Morphosyntactic changes of lexicon :

- *e.g.* adding space between letters and numbers
- changes in hyphen and spaces ...

Lexicon → New Lexicon

Lexicon	New Lexicon
AP2	AP 2
CRP810_1.3	AP_2
1,2-Dioxygenase	AP-2
LEC 2	CRP810 1.3
	CRP 810 1.3
	...

CC-BY

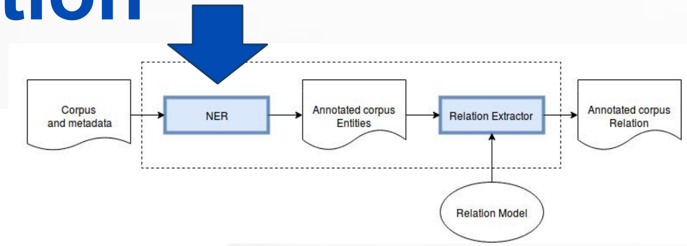
NER Preprocessor with Lexicon

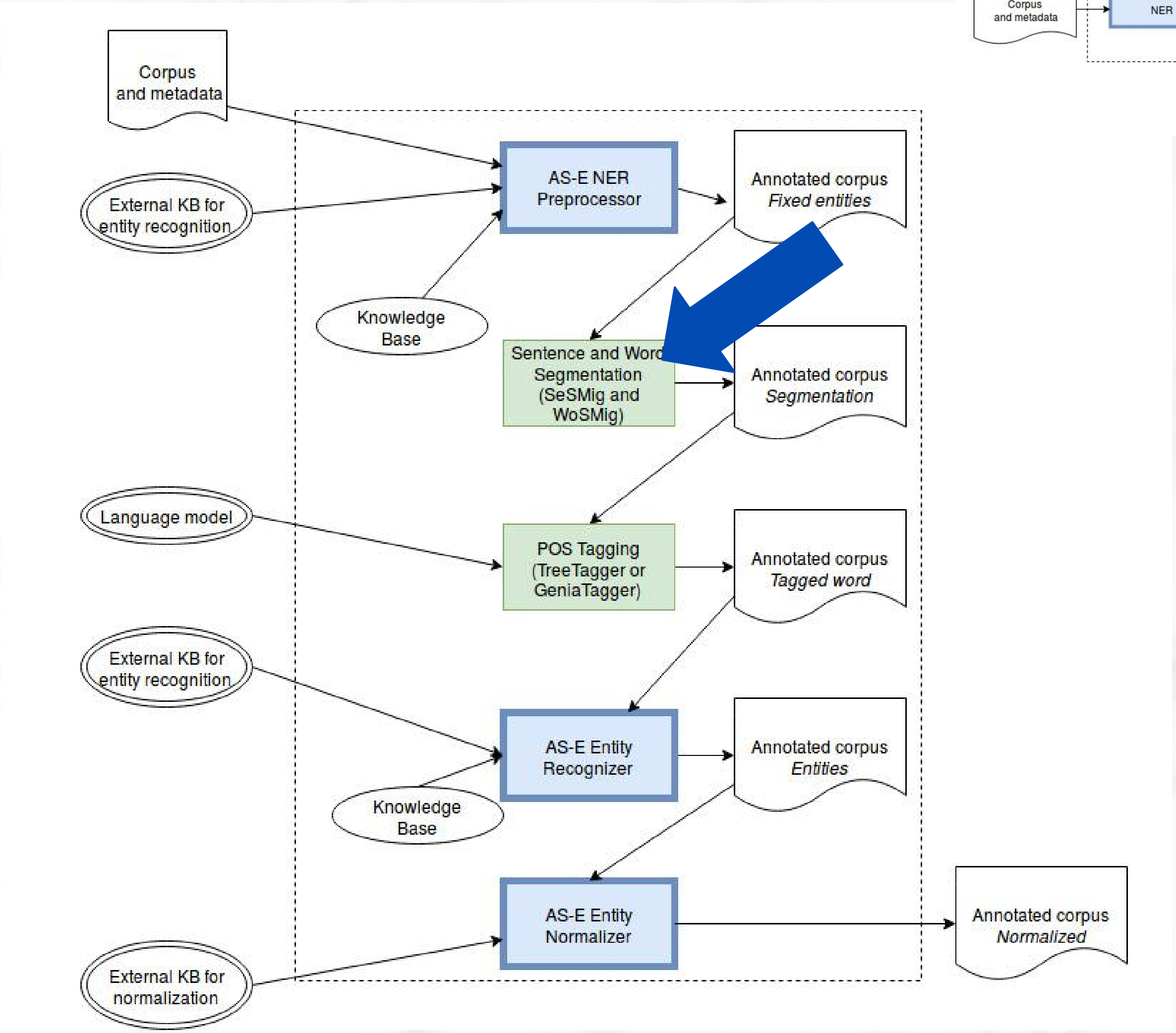
Parameterization of lexicon projection
e.g. CaseInsensitive : the match allows case substitutions on all characters

Lexicon	*e.g.* could match with
AP2	ap2
CRP810_1.3	crp810_1.3
1,2-Dioxygenase	1,2-DIOXYGENASE
LEC 2	Lec 2

CC-BY

NER: Named Entity Recognition





Corpus and metadata
External KB for entity recognition
AS-E NER Preprocessor
Annotated corpus *Fixed entities*
Knowledge Base
Sentence and Word Segmentation (SeSMig and WoSMig)
Annotated corpus *Segmentation*
Language model
POS Tagging (TreeTagger or GeniaTagger)
Annotated corpus *Tagged word*
External KB for entity recognition
AS-E Entity Recognizer
Annotated corpus *Entities*
Knowledge Base
AS-E Entity Normalizer
Annotated corpus *Normalized*
External KB for normalization