

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

1. Context

Information Extraction and Biology

Use-case : Regulatory Network in plant



How do genes regulate the development of my plant?

What is a text?

A sequence of characters?

lorem ipsum dolor sit amet
consectetur adipiscing elit
quisque ante tellus pulvina
vitaesollicitudin necposu
erequismassanulla justo
augue aliquam in euismod
in mollis id justo nullam
ssamassa pharetra eget
venenatis

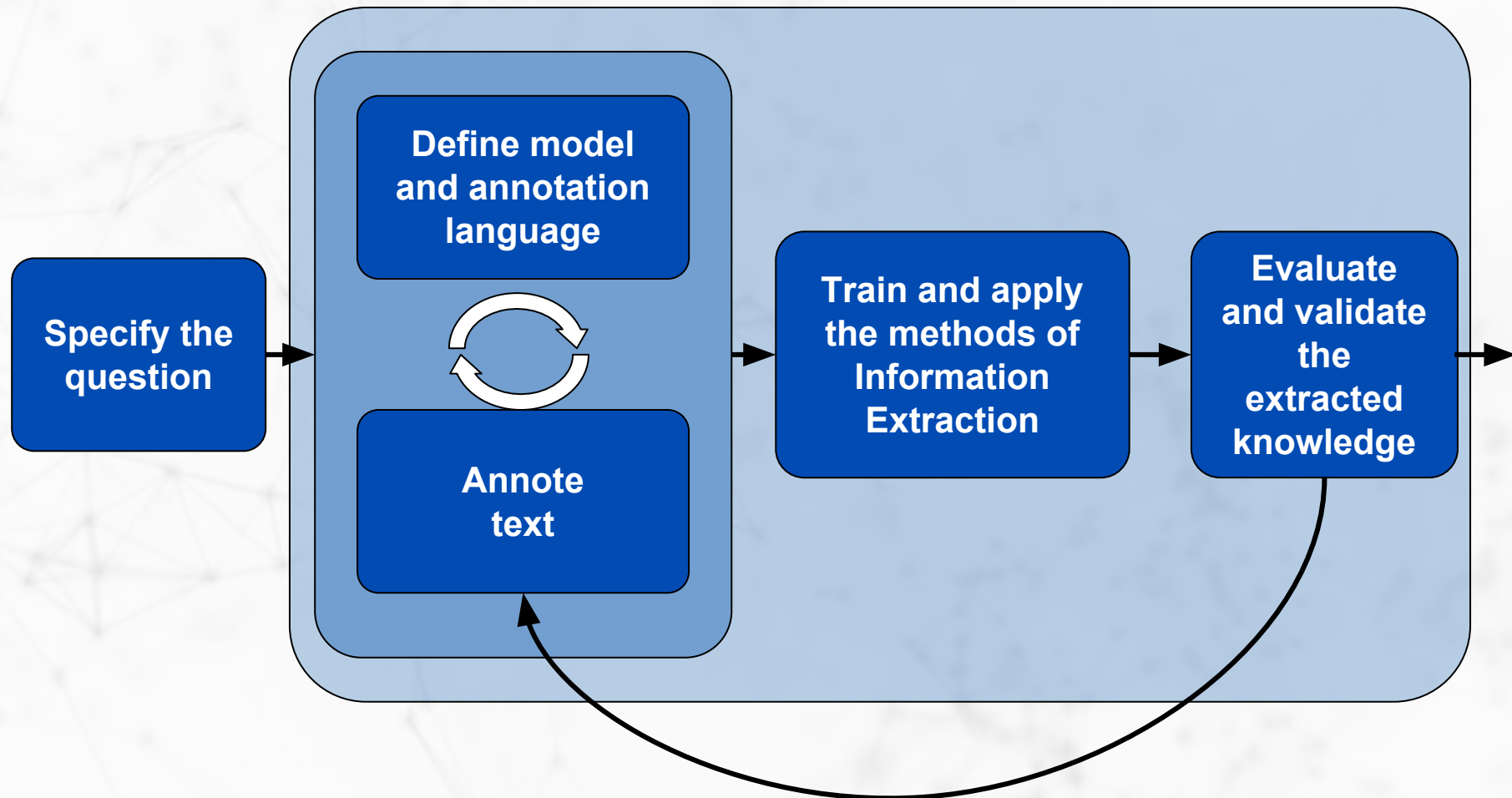
A sequence
of words?

lorem ipsum dolor sit
amet, consectetur
adipiscing elit quisque
ante tellus, pulvina
vitae sollicitudin nec,
posuere quis massa.
nulla justo augue,
aliquam in euismod in,
mollis id justo nullam
massa massa, pharetra
eget venenatis.

Words and
sentences

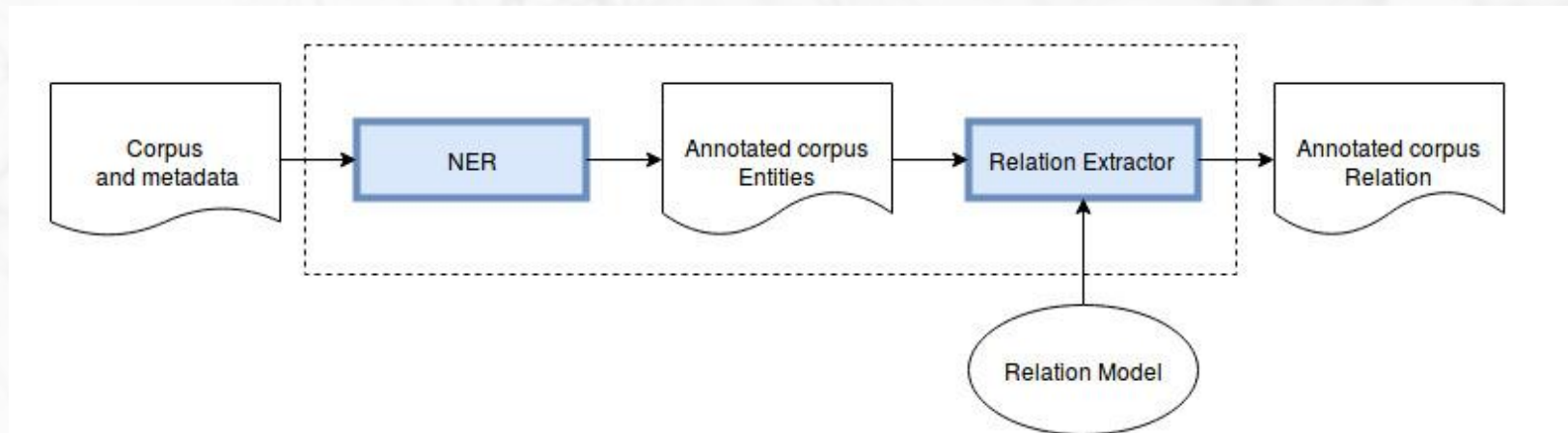
Lorem ipsum dolor sit
amet, consectetur
adipiscing elit. Quisque
ante tellus, pulvina
vitae sollicitudin nec,
posuere quis massa.
Nulla justo augue,
aliquam in euismod in,
mollis id justo. Nullam
massa massa, pharetra
eget venenatis.

TDM Application Design Approach



How to process the text?

Multiple steps are needed to go from a text corpus to relevant extracted information (entities and relations)



How to process the text?

The first steps usually involve tools to detect sentences and words, depending, for example, on spaces, punctuation, etc.

This is a text segmentation.