OryzaGP - Japanese rice genes proteins dataset for Named Entity Recognition

Huy Do and Pierre Larmande

University of Science and Technology of Hanoi (USTH), ICT Lab
Institute of Research for the Development (iRD) Vietnam
pierre.larmande@ird.fr

Japan, …, February 2019

Outline

I. Introduction

II. Materials

III. Implementation

Introduction - Named Entity Recognition (NER)

Definition:

Kevin comes from Paris 

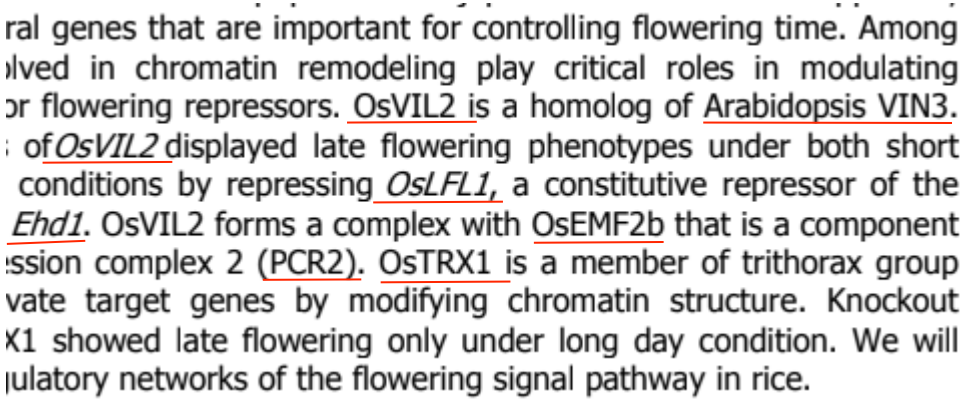
Kevin		comes	from
		Paris	
B-Per	O	O	
	B-Loc		

Why NER?

- Searching, analyzing,.. become easier
- Classifying content

Introduction - NER in this context

- Extract information of biological entities such as genes, proteins from unstructured text



ral genes that are important for controlling flowering time. Among
lved in chromatin remodeling play critical roles in modulating
or flowering repressors. OsVIL2 is a homolog of Arabidopsis VIN3.
of *OsVIL2* displayed late flowering phenotypes under both short
conditions by repressing *OsLFL1*, a constitutive repressor of the
Ehd1. OsVIL2 forms a complex with OsEMF2b that is a component
ssion complex 2 (PCR2). OsTRX1 is a member of trithorax group
vate target genes by modifying chromatin structure. Knockout
X1 showed late flowering only under long day condition. We will
ulatory networks of the flowering signal pathway in rice.

- Enrich content of database
- However, few NER study for plants, especially rice

Introduction - Objective

- Create a dataset of rice “genes” and “proteins” while there are very few benchmarks to evaluate NER for plant biology
- Create a tool to solve Named Entity Recognition for rice

Material - Dataset

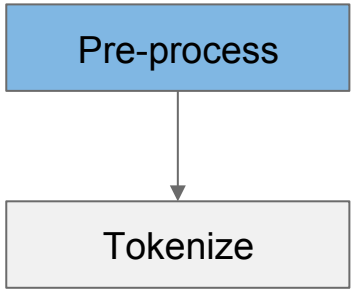
Oryzabase - a rice comprehensive database (2000)

- Lasted version contained over 21.000 genes of rice collected from 44.837 distinct articles from PubMed

Name	Oryzabase
text genre	article
text type	abstract & title
entity type	gene, protein
num of articles	10400
num of sentence	75096
num of words	2697726

Table 1: The details of dataset

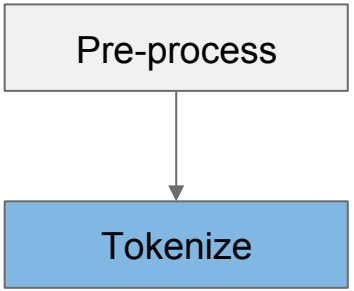
Implementation



- Focus on abstract and title of articles
- Raw data is downloaded from online resource (PubMed)

PubMedID	Title	Year	Journal	Abstract
12668768	The sucrose transporter gene family in rice.	2003	Plant Cell Physiol	In this paper we report the identification, cloning and expression analysis of four putative sucrose transporter (SUT) genes from rice, designated OsSUT2, 3, 4 and 5. Three of the four genes were identified through extensive searches of the recently published draft sequence of the rice genome. Along with the previously reported OsSUT1 we propose that these five genes comprise the rice SUT gene family. Complementary DNA clones were isolated for the four newly identified genes. The deduced proteins of all five SUT genes were predicted to contain 12 membrane-spanning helices and a domain highly conserved throughout all known plant SUTs, suggesting the four additional OsSUT genes encode functional SUTs. Reverse transcription-PCR analysis was performed in order to investigate the expression pattern of each member of the SUT family in rice. A differing but overlapping expression pattern was observed for each member of the SUT family at different stages through plant development. These results, together with the structural variations apparent from the deduced protein sequences, suggest that the five SUTs possess diverse roles in both sink and source tissues. We also discuss the classification and evolution of the rice SUT gene family, using a comparison of the gene structures and deduced amino acid sequences with other known plant SUT genes.

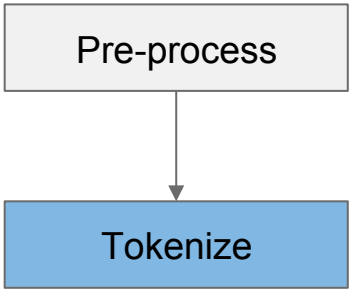
Implementation - Tokenize



- Each word is considered as a token
- A token including 3 parts:
  - the word itself
  - the part-of-speech tag (POS)
  - the entity tag (IOB format)

```
here                RB    O
we                  PRP   O
attempted           VBN   O
biochemical         JJ    O
characterization    NN    O
of                  IN    O
cyp99a2             NN    B-gene
and                 CC    O
cyp99a3             NN    B-gene
which               WDT   O
was                 VBD   O
ultimately          RB    O
achieved            VBN   O
```

Implementation - POS tag

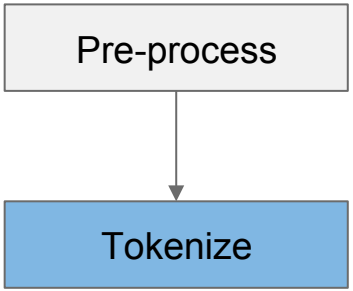


- Natural Language Toolkit (NLTK)
- Aim to exploit information from words.
- 36 kinds of tags:

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word

NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer

Implementation - Entity Tag



- In IOB (Inside, Outside, Begin) format
- Filter from Oryzabase list gene:
  - Gene family
  - Prefix
- Then matching gene to tokenize text.

Thank you for your attention!