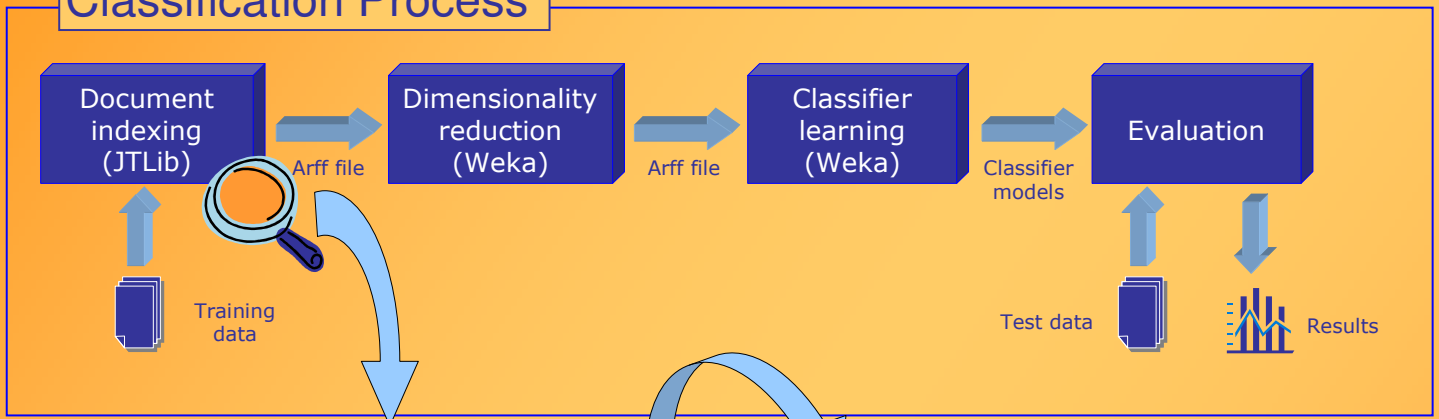


Feature Engineering and Quick Prototyping of Gene Mention Classifiers*

Manuel J. Maña López
Jacinto Mata Vázquez
Universidad de Huelva
{manuel.mana, mata}@diesia.uhu.es

Francisco Carrero García
José M^a Gómez Hidalgo
Universidad Europea de Madrid
{francisco.carrero, jmgomez}@uem.es

Classification Process



Taxonomy of Attribute Types

Intrinsic

Information used to compute the attribute comes only from the same example

Contextual extrinsic

The information is obtained from the processed example, but also from other examples that have a strong relation with it

Global extrinsic

The information comes from all the examples in the set

JTLib

attrs

- Common pre-defined attributes

jtFormatter

- Pre-processing of attributes
- Processing of intrinsic attributes
- Processing of contextual and global extrinsic attributes
- Generation of data set

tasks

- Sub-packages that implement several classifiers

Selected Attributes

Attribute type	Attribute Name
Intrinsic	hyphen
	punctuation
	initCaps
	lettersAndDigits
Global extrinsic	number
	frequentWords
	frequentWordsInEntity
	prev1Unigrams
	prev2Unigrams
Contextual extrinsic	startingWords
	endingWords
	endOfSentence
	punctuation ± <i>n</i>
	initCaps ± <i>n</i>
frequentWords ± <i>n</i>	
frequentWordsInEntity ± <i>n</i>	

Ranking of Attributes (IG)

	Attribute		Attribute
1	frequentWords	8	frequentWordsInEntity + 1
2	frequentWordsInEntity	9	lettersAndDigits
3	frequentWords - 1	10	startingWords
4	frequentWords + 1	11	frequentWords - 2
5	endingWords	12	frequentWords + 2
6	frequentWordsInEntity - 1	13	frequentWordsInEntity - 2
7	prev1Unigrams	14	prev2Unigrams

Results

	C4.5 unpruned	C4.5 pruned
Precision	50.09	53.37
Recall	46.12	42.46
F-measure	48.02	47.29