

Learning Interestingness Measures in Terminology Extraction.

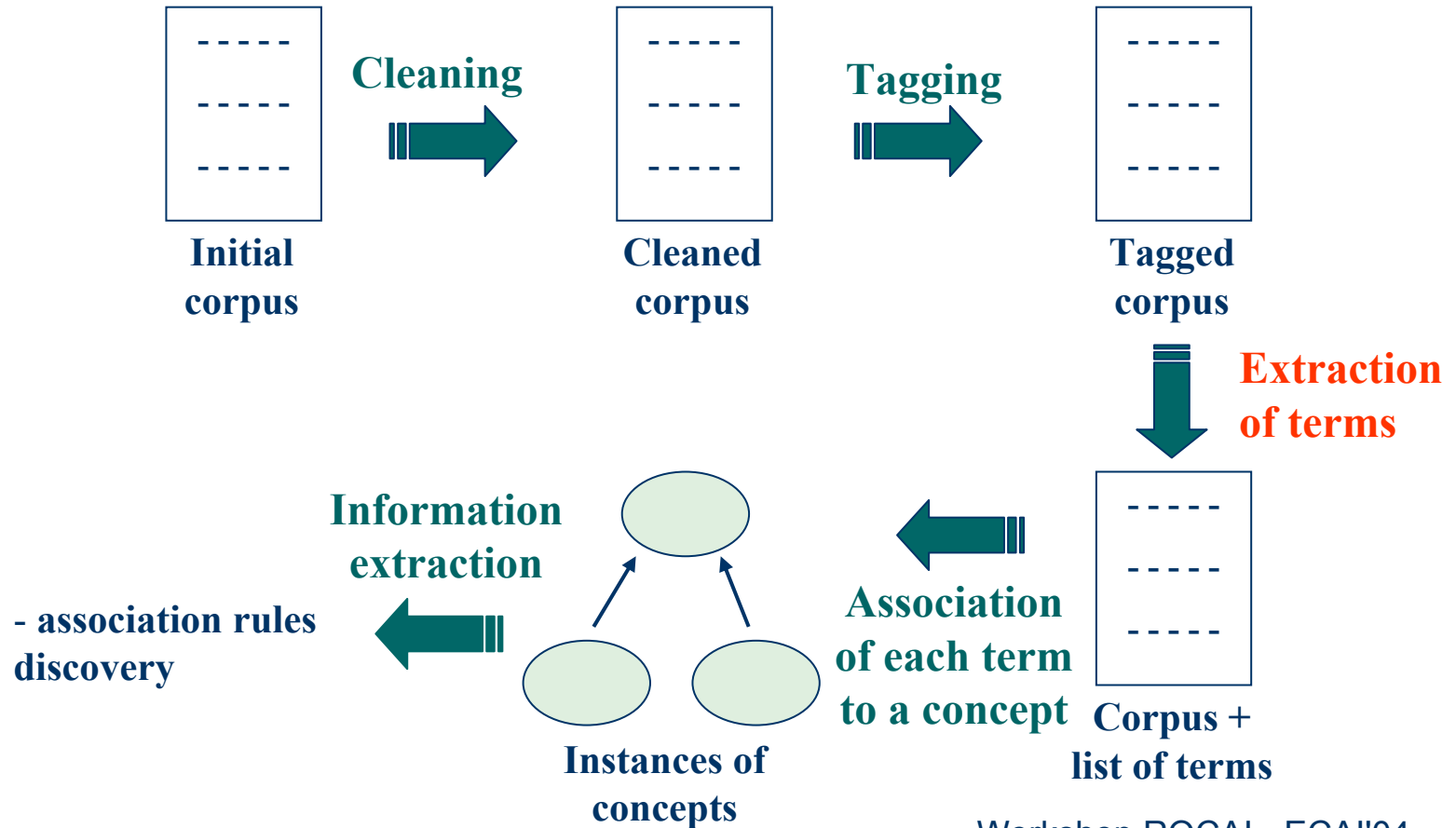
A ROC-based approach

Mathieu Roche, Jérôme Azé
Yves Kodratoff, Michèle Sebag

LRI, Université Paris-Sud - FRANCE



Global process of text mining



Terminology Extraction

Relevant collocations: *collocations are instances of a concept*

1. remote sensing
2. naive bayes
3. frequent itemset

4. real world
5. neural network
- ...

Sort by one measure of interest

1. real world
2. neural network
3. frequent itemset

4. remote sensing
5. naive bayes
- ...

Sort by an other measure of interest

Corpora (1/2)

- **CV corpus**: Set of Curriculum Vitae (VediorBis company, 2470 Ko) - French corpus

Relations	# collocations	# collocations after pruning of 3
Noun-Noun	1781	162
Noun-Preposition-Noun	3634	307
Adjective-Noun	1291	103
Noun-Adjective	3455	448

- **Molecular Biology corpus**: Abstracts of papers on the topics of Molecular Biology (9424 Ko) - English corpus.

Relations	# collocations	# collocations after pruning of 4
Noun-Noun	22241	3332
Noun-Preposition-Noun	4363	251
Adjective-Noun	23284	2547

Corpora (2/2)

Data sets	# collocations	% positives collocations	% negatives collocations
CV frequent	376	85.7	14.3
CV rare	2822	56.6	43.4
Biology frequent	1028	90.9	9.1

Statistical criteria in terminology extraction and AUC

Statistical criteria	AUC <i>Frequent collocations</i> CV corpus	AUC <i>Frequent collocations</i> Biology corpus
Occ_L - <i>number of occurrences + Loglikelihood</i>	0.58	0.57
L - <i>Loglikelihood</i> [Dunning 1993]	0.43	0.42
MI³ - <i>Mutual Information with cube</i> [Daille et al. 1998]	0.40	0.35
Dice - <i>Dice coefficient</i> [Smadja et al. 1996]	0.39	0.31
MI - <i>Mutual Information</i> [Church and Hanks 1990]	0.31	0.30

Roger (ROc-based GEnetic learneR)

[Sebag *et al.* ICDM'03, EA'03]

- Goal: learning a ranking function by optimization of the area under the ROC curve.
- Input: set of annotated collocations described by values of measures of interest
- Output: a rank for each collocation indicating the relevance (first: more relevant, last: more irrelevant)

Algorithm, 1

Linear approach

$$h(\text{Ex}) = \sum w_i \times \text{att}_i(\text{Ex}) \quad \text{with } (\text{Ex}, +/-)$$

Non linear approach

$$h(\text{Ex}) = \sum w_i \times | \text{att}_i(\text{Ex}) - c_i | \quad \text{with } (\text{Ex}, +/-)$$

Hypothesis quality : Area under the ROC curve

$h \rightarrow (\text{rank}(\text{Ex}), \text{Label}(\text{Ex}))$

sort the examples by increasing rank



+ : relevant collocation

- : irrelevant collocation

Algorithm, 2

Optimisation tool

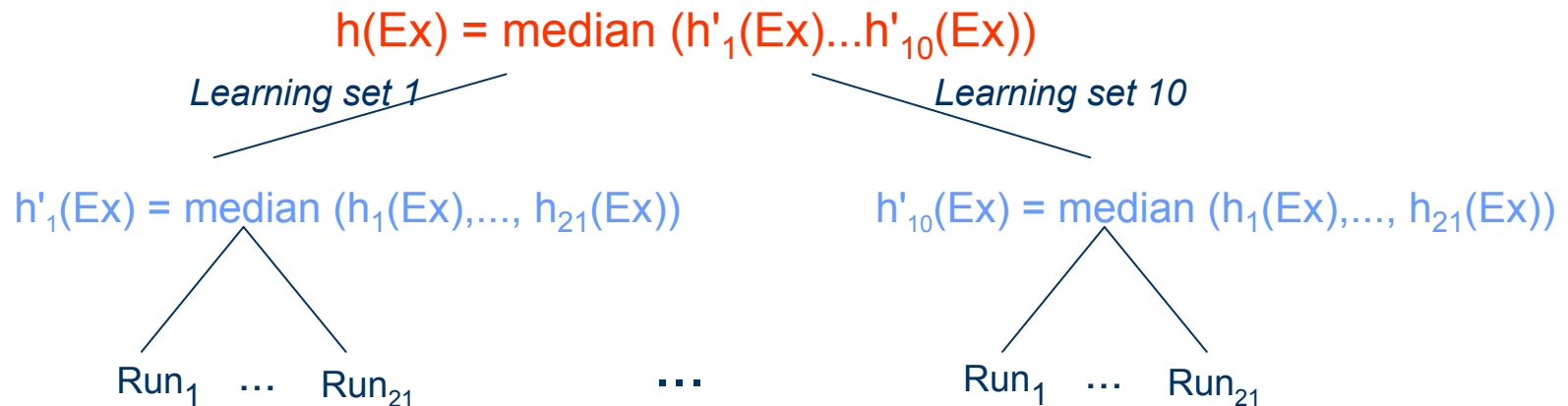
- Evolutions strategies (use of EvolC)

Experimental Validation

- 90% Learn , 10% Test
- 21 independant runs
- Bagged result over all the runs using a median

Bagged-Roger

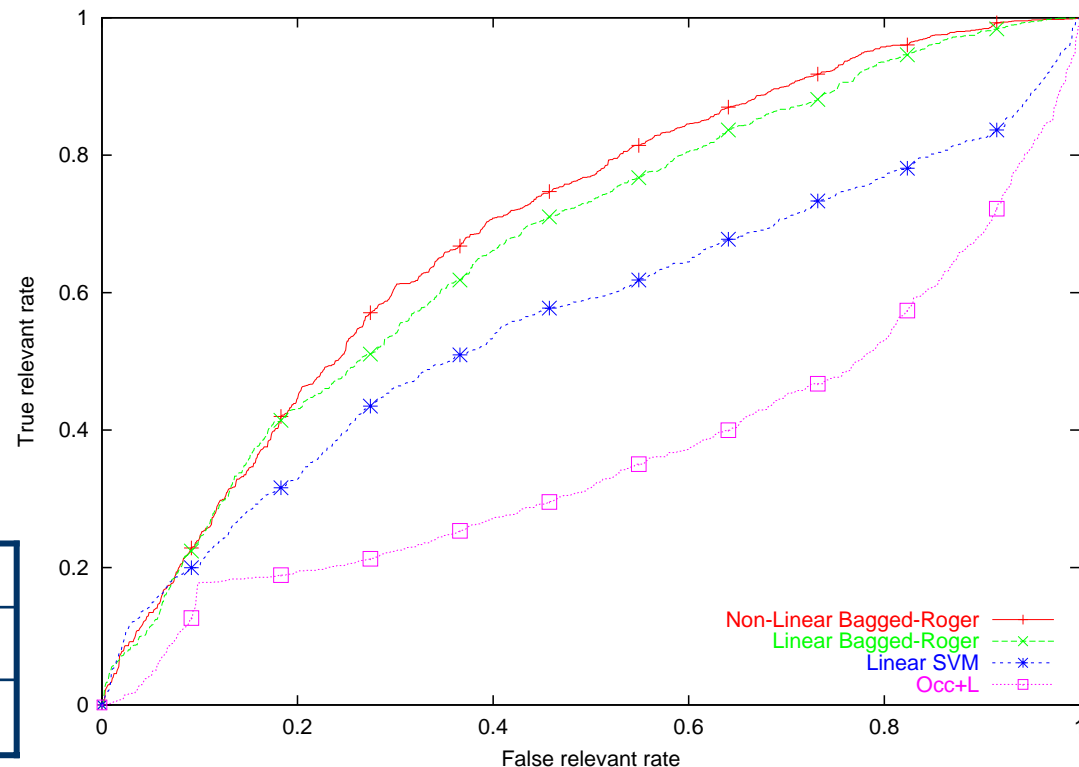
- Different independent hypotheses are bagged in order to increase the quality of the ROC curve
- All the hypotheses learned by Roger are bagged using a median of median



First validation: *rare collocations* - CV Corpus

	AUC <i>Rare collocations</i> CV corpus
<i>Occ_L</i>	0.37
<i>Dice</i>	0.32
<i>MI³</i>	0.30
<i>L</i>	0.30
<i>MI</i>	0.29

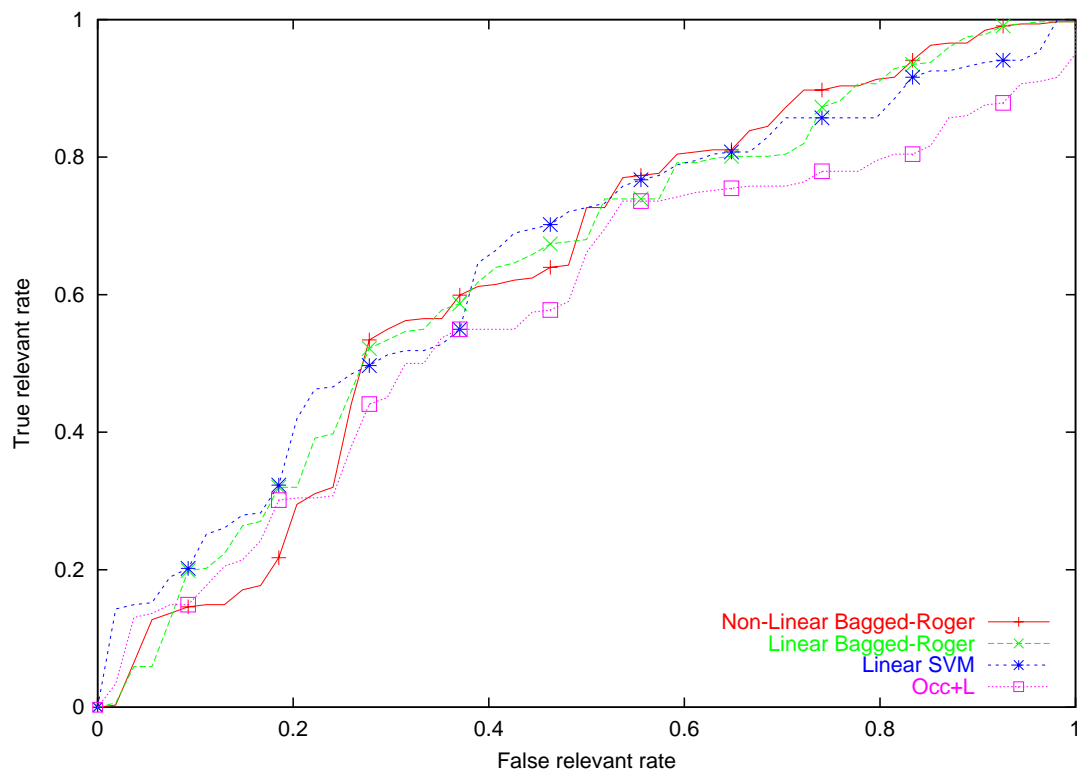
SVM	Bagged-Roger	
<i>Linear</i>	<i>Linear</i>	<i>Non-Linear</i>
0.56	0.67	0.70



Second validation: *learning with Biology and application with CV*

	AUC <i>Frequent collocations</i> CV corpus
Occ_L	0.58
L	0.43
MI^3	0.40
$Dice$	0.39
MI	0.31

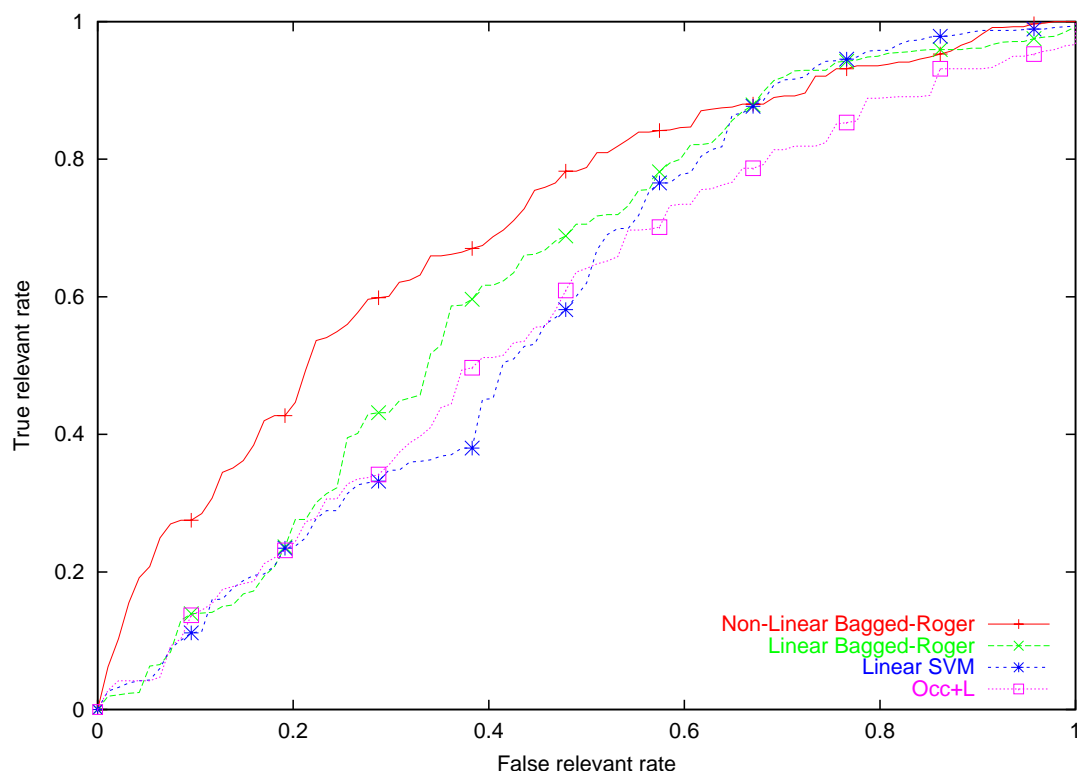
SVM	Bagged-Roger	
<i>Linear</i>	<i>Linear</i>	<i>Non-Linear</i>
0.65	0.64	0.63



Third validation: *learning with CV and application with Biology*

	AUC <i>Frequent collocations Biology corpus</i>
Occ_L	0.57
L	0.42
MI^3	0.35
$Dice$	0.31
MI	0.30

SVM	Bagged-Roger	
<i>Linear</i>	<i>Linear</i>	<i>Non-Linear</i>
0.59	0.63	0.71



Conclusion

- good behavior over different domains and languages
- learning of a set of optimal functions bagged using the median
- model learned from one corpus can be successfully applied to an other one

Perspectives

- use more information to describe the data (contextual, linguistic, ...)
- introduce an expert in the learning loop
 - by using in-line learning
 - study the minimal set of annotated examples in order to obtain a good generalisation (validation by an expert)
- use other functions (polynomial, gaussian, ...)

Annexe



Second step : Tagging

The modulation of the
biological activities
of mitochondrial
histone Abf2-protein
...



Brill's tagger

The/**DT** modulation/**NN** of/**IN**
the/**DT** biological/**JJ**
activities/**NNS** of/**IN**
mitochondrial/**JJ** histone/**NNP**
Abf2-protein/**NNP** ...

- Using ETIQ [Amrani *et al.*, 04], a tagger based on Brill's tagger for specialized corpus.
 - Lexical rules to improve quality of tagging.
 - Adding new tags. For example, *formula tag* in Molecular Biology corpus

Measures (1/2)

- Mutual information [Church and Hanks, 90]

$$IM(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

$$\Rightarrow I(x, y) = \log_2 \frac{nb(x, y)}{nb(x)nb(y)}$$

Measures (2/2)

- Loglikelihood [Dunning, 93 ; Daille 98 ; Xu 02]

	<i>y</i>	<i>y' with y' ≠ y</i>
<i>x</i>	a	b
<i>x' with x' ≠ x</i>	c	d

$$L = a \log(a) + b \log(b) + c \log(c) + d \log(d) - (a+b) \log(a+b) - (a+c) \log(a+c) - (b+d) \log(b+d) - (c+d) \log(c+d) + (a+b+c+d) \log(a+b+c+d)$$

We can use other measures: Mutual Information with cube, Dice coefficient, etc.

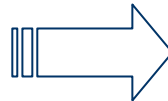
Evaluation of the measures

Precision

$$precision = \frac{\text{number of relevant collocations extracted}}{\text{number of collocations extracted}}$$

1. real world
2. neural network
3. frequent itemset
4. remote sensing
5. naive bayes
...

Collocations extracted



1. **real world**
2. **neural network**
3. **frequent itemset**
4. remote sensing
5. **naive bayes**
...

Relevant collocations: collocations are instances of a concept

- *Lift chart* measures the variation of the precision as a function of the proportion of terms found by the system.

Algorithm, 2

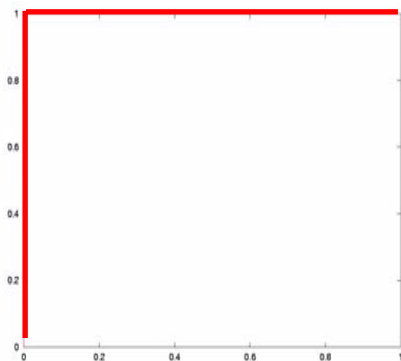
Optimization step

maximise the area under the ROC curve

⇔ minimise the rank's sum of positives examples

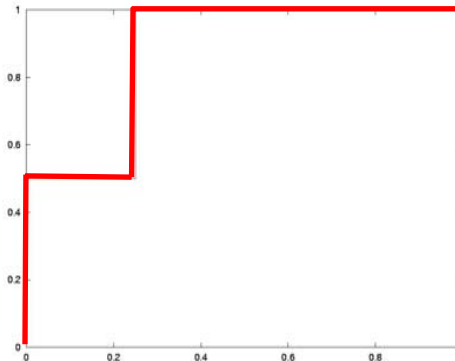
$\Sigma\text{rank} = 21$

$h_1: ++++++-----$



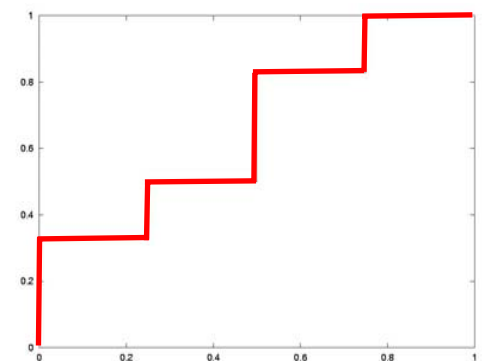
$\Sigma\text{rank} = 25$

$h_2: +++-++++---$



$\Sigma\text{rank} = 26$

$h_3: +-+-+--+--+$



Evolutionary algorithm approach

