

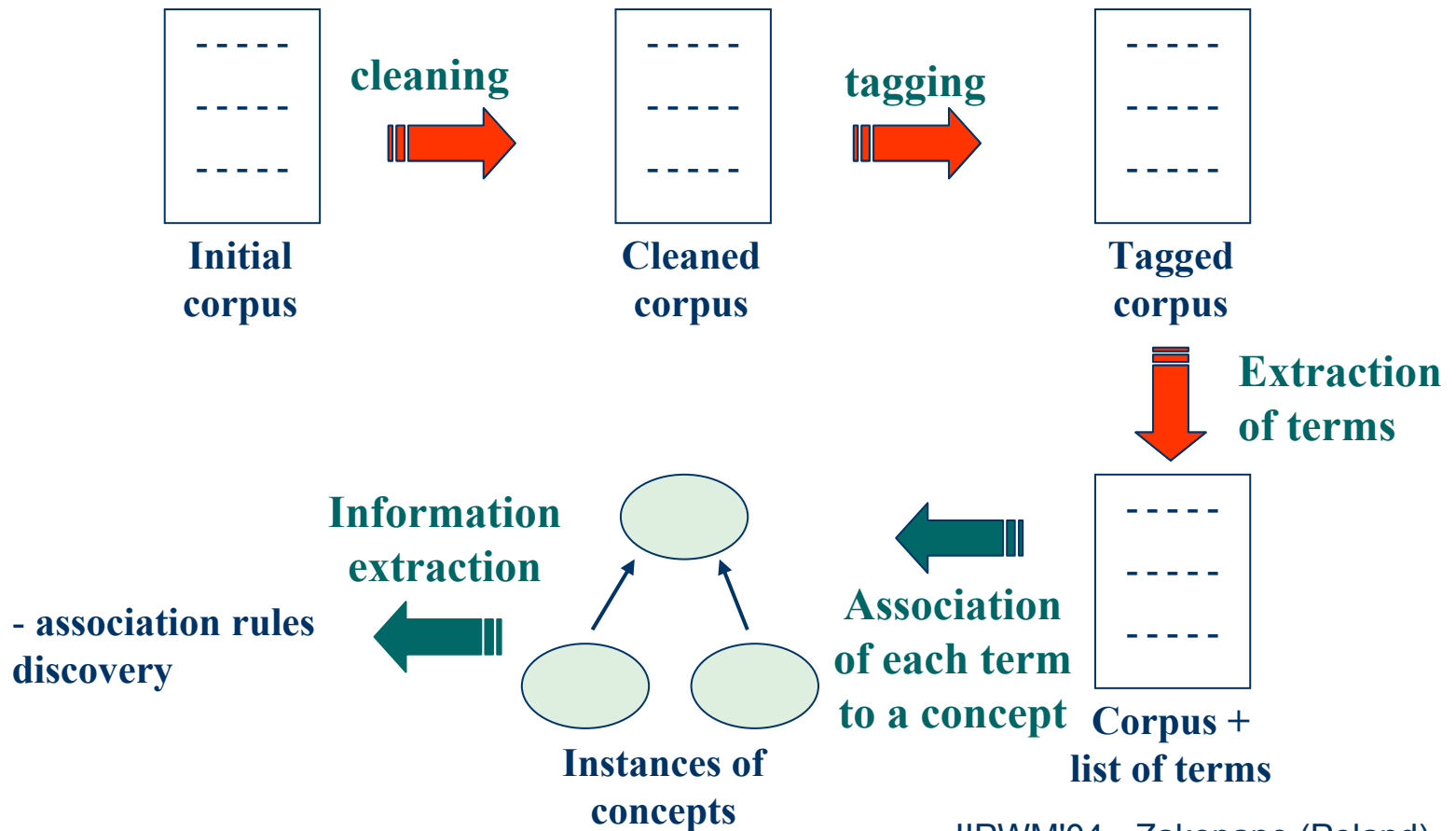
Mining texts by association rules discovery in a technical corpus

Mathieu Roche, Jérôme Azé
Oriane Matte-Tailliez, Yves Kodratoff

LRI, Université Paris-Sud - FRANCE



Global process of text mining



Corpora

- **Data mining corpus**: Introductions of papers relative to the field of Data Mining (369 Ko) - English corpus.
- **Molecula Biology Corpus**: Abstracts of papers on the topics of Molecular Biology (9424 Ko) - English corpus.
- **CV corpus**: Set of Curriculum Vitae (VediorBis company, 2470 Ko) - French corpus
- **Human Resources corpus**: Set of texts commenting in natural language the results of ability tests (PerfomanSe company, 3784 Ko) - French corpus

First step : cleaning

~~1: Biochim Biophys Acta 2001 Dec 30;1522(3):175-86~~

The modulation of the biological activities of mitochondrial histone Abf2p by yeast PKA and its possible role in the regulation of mitochondrial DNA content during glucose repression.

~~Cho JH, Lee YK, Chae CB.~~

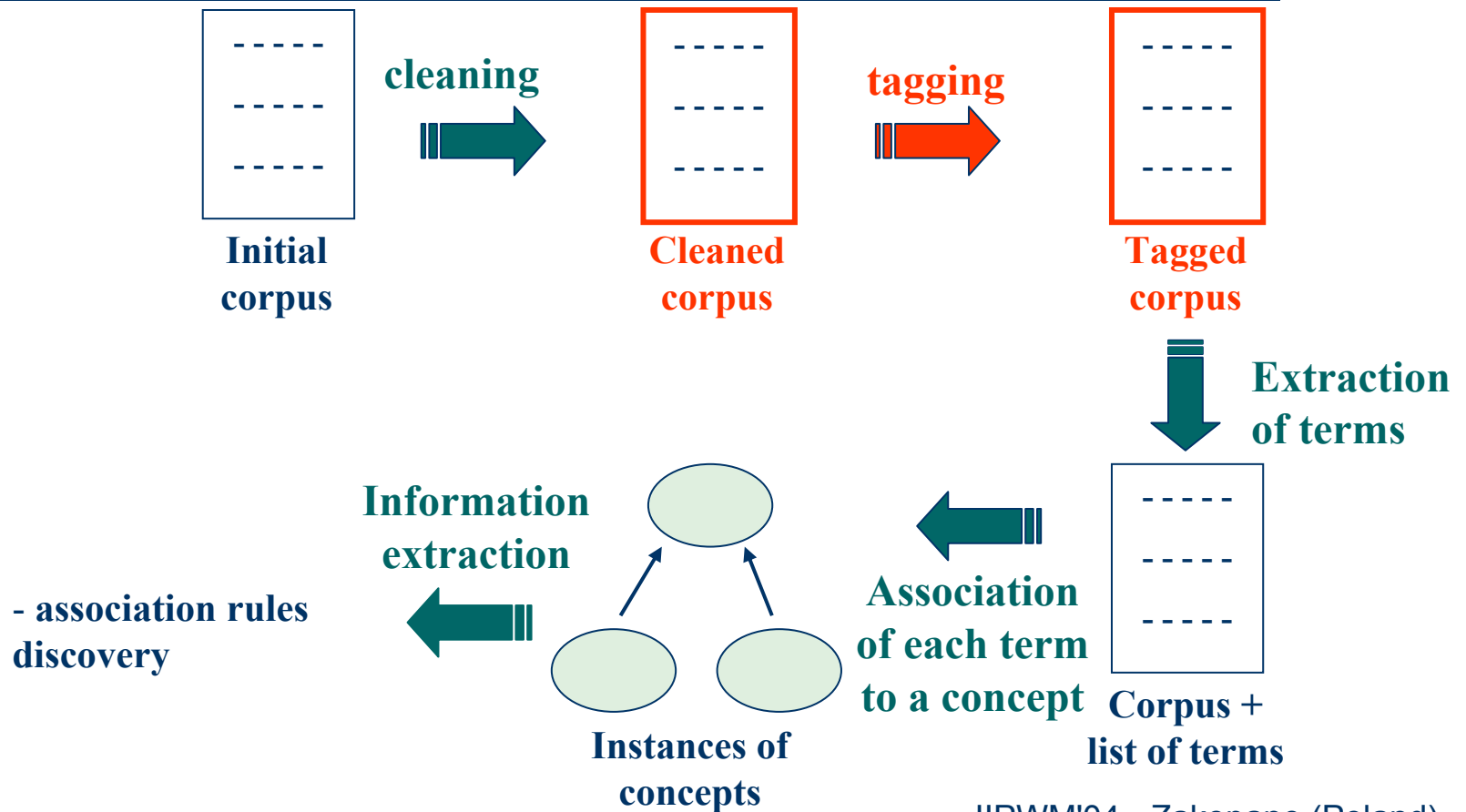
~~Department of Life Science and Division of Molecular and Life Science, Pohang University of Science and Technology, 790-784, Pohang, South Korea~~

The mitochondrial histone, Abf2p, of *Saccharomyces cerevisiae* is essential for the maintenance of mitochondrial DNA (mtDNA) and appears to play an important role in the recombination and copy number determination of mtDNA.

~~PMID: 11779639 [Epub ahead of print - in process]~~

- Rules to standardize the corpus. For example, in Molecular Biology, we replaced by "C-term" all occurrences of "carboxy-terminal," "carboxy termini," "carboxyl terminal," "COOH-terminal," etc.

Global process of text mining



Second step : Tagging

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

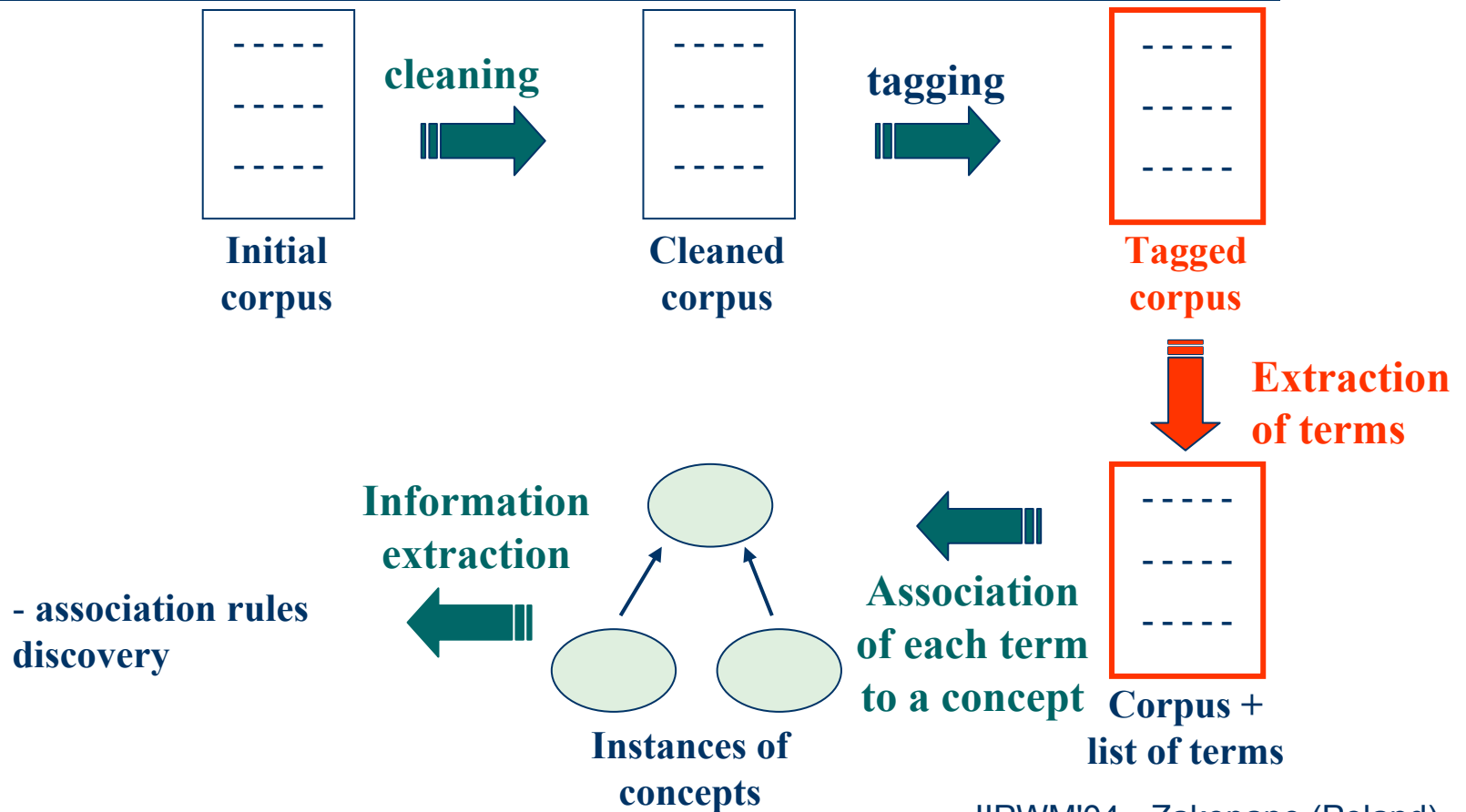


Brill's tagger

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

Global process of text mining



Using EXIT (Iterative Extraction of Terminology)

- Iterative process to extract terminology : *to extract terminology at the n -th iteration, we can use terms found at the $(n-1)$ th one to build more complex terms.*
- Using measures to extract terminology.

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



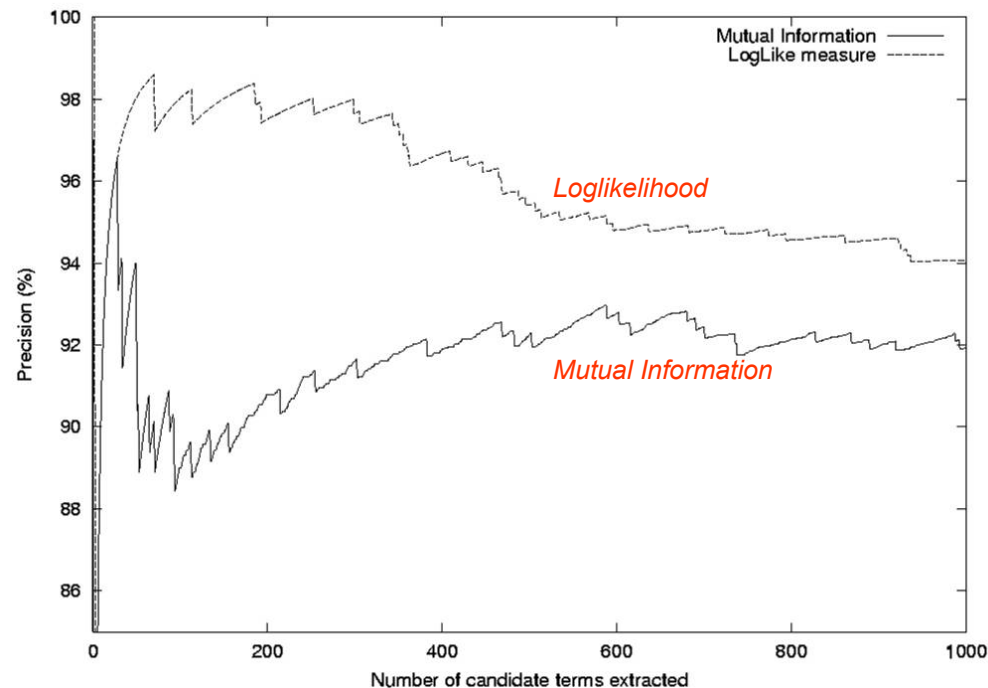
1. **real world**
2. **neural network**
3. **frequent itemset**
4. remote sensing
5. **naive bayes**
...

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.

Evaluation of the measures

Lift chart with the Molecular Biology corpus (adjective-noun relation).



Conclusion and Perspectives

- The expert have an essential role in text-mining process.
- Complete chain of treatment: an error appearing at the beginning of the process can completely spoil the results of the following steps.
- Combination of measures to extract terminology [submitted paper at ECML'04]