Selecting a Feature Set to Summarize Texts in Brazilian Portuguese

Daniel Saraiva Leite
Undergraduate Student

Lucia Helena Machado Rino, PhD
Advisor

NILC - Núcleo Interinstitucional de Lingüística Computacional
UFSCAR - Universidade Federal de São Carlos

Overview

- Introduction: The Summarization Task
- Extractive AS based on Machine Learning
- Scenario: The SuPor System
  - Employed Methods
  - How methods are mapped into features
  - Feature selection problem
- Taking advantages of WEKA
  - Improving the Model
  - Machine Learning Techniques
- Assessments
- Final Remarks

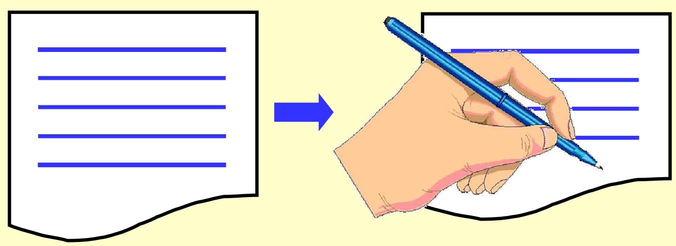
The Summarization Task

- Taking one or more texts and producing a shorter one
- The summary should convey the main content information of the original text

Two Main Approaches for Automatic Summarization

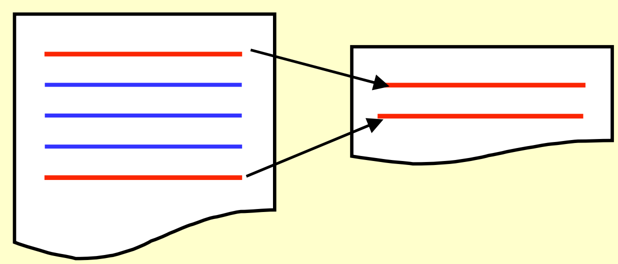
Building *abstracts*

- Rewriting the text



Building *extracts*

- *copying-and-pasting* full sentences



Extractive AS based on Machine Learning

Extractive Automatic Summarization

- How to choose sentences to include in the summary?
  - Based on the relevance of each sentence
  - Take the top relevant ones
  - Stop when desired length is achieved

Machine Learning for Extractive AS - Kupiec et al. (1995)

- Relevance ~ Likelihood of inclusion in the Extract
  - **Naïve-Bayes** is suggested
  - Shallow features of the text (E.g., location, frequency of the words, etc.) – as far back as (Luhn, 1958; Edmundson, 1969)
  - Binary representation

Using Naïve-Bayes

- **Training phase**
  - Need of a Corpus: Source Texts (**ST**) and "Ideal" Extracts (**IE**)
  - For each sentence **S** of a **ST**
    - Process its features
    - Verify if it also appears in the corresponding **IE**

If $S \in IE \rightarrow$ Class is 'Yes'

If $S \notin IE \rightarrow$ Class is 'No'

F_1	F_2	F_3	F_4	F_5	$S \in E$?
no	yes	no	yes	no	no
no	no	no	yes	yes	**yes**
no	yes	yes	yes	no	no
no	yes	no	yes	no	**yes**

We get a dataset in which each instance is the representation of a sentence of the ST

Extractive AS based on Machine Learning

Using Naïve-Bayes

- **Sentence Classifying phase**
  - Computing each sentence → features (Fi's)
  - Using Naïve-Bayes formula and the training dataset
    - Calculating its probability for class **S Є E = 'Yes'**

$$P((s \in E) \mid F_1, F_2, \ldots, F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in E) P(s \in E)}{\prod_{j=1}^{k} P(F_j)}$$

    - Is it a classification task?
      - We are always interested in probabilities for just one class

Our scenario: SuPor (Módolo, 2003)

Main aspects

- Based on Kupiec's et al. (1995) model
- An AS environment
  - User can choose features he/she wants → customization to a given AS system
  - Many different AS methods

Novelties

- Besides shallow and basic features, SuPor embeds:
  - Lexical Chains (Barzilay & Elhadad, 1999)
  - Importance of Topics (Larocca Neto et al., 2000)
  - Relationship Map (Salton et al., 1997)
- Methods mapped into binary features

SuPor Features

Name		Condition for sentence S be labeled "Yes"
F1	**Lexical Chains**	**S** must be recommended by at least one of the three heuristics of the method
F2	**Location**	**S** must appear in special positions of the text (beginning or ending)
F3	**Words Frequency**	**S** sum of its words frequency must be higher than a threshold
F4	**Relationship map**	**S** must be recommended by at least one of the three heuristics of the method
F5	**Importance of Topics**	**S** must appear in an important topic and must be very similar to such topic
F6	**Proper Nouns**	**S** must contain a number of proper nouns higher than a threshold
F7	**Sentence Length**	**S** number of words must be higher than a threshold

Actually → 11 features (by varying preprocessing)

SuPor Drawbacks

Feature Selection Problem

- How the user can select the right feature set?
  - Difficult task → He/she must be an expert in AS and still... he/she may not be able to properly accomplish it
  - Extracts quality depends a lot on the feature set (100% in some cases)

↓

Motivation to our work

SuPor Drawbacks → Motivation to our work

- Explore means to reduce such effort of customization
  - Automatic Feature Selection!
  - **Combine** SuPor with **WEKA**
    - Free machine learning tool
    - Very comprehensive
      - Classification, Rules, Clustering
      - Data visualization and preprocessing
    - Available at www.cs.waikato.ac.nz/ml/weka/

Taking Advantage of WEKA

Two Approaches

1) Automatic Feature Selection allows judging the relevance of features subset and choosing the best!

- Methods based on Entropy measure (Shannon's Information Theory)
- Employed as a filter before classification

2) Change Features Representation

- *Hypothesis*: improving representation → Feature Selection might be not necessary
- Provide more information to the machine learning algorithm
- Try other classifiers → C4.5 (suggested by Módolo, 2003)

Approach 1: CFS (Correlation Feature Selection) – Hall, 2000

- Measure to evaluate importance of a subset of features

$$\underbrace{\sum \text{IG (feature i,classe)}}_{\text{relevance}} - \underbrace{\sum \text{IG (feature i, feature j)}}_{\text{redundancy}}$$

- Idea of low redundancy seems good for Naïve-Bayes (Independence Assumption)

- Measure employed together with a search heuristic → In WEKA, by default, Hill-Climbing

Approach 2: Improving Features Representation

Principles

- Non-binary features
- Explore numeric and multivalued features

- Sentence Length: number of words of the sentence

- Proper Nouns: number of proper nouns of the sentence

- Words Frequency: sum of the frequency of each word of the sentence

Approach 2: Improving Features Representation

- Location: according to 9 labels:

Label	Position of paragraph	Position of sentence within the paragraph
II	Initial	Initial
IM	Initial	Medial
IF	Initial	Final
MI	Medial	Initial
MM	Medial	Medial
MF	Medial	Final
FI	Final	Initial
FM	Final	Medial
FF	Final	Final

Approach 2: Improving Features Representation

- Importance of Topics: Harmonic mean between topic importance and sentence similarity to the topic
- Relationship Map and Lexical Chains: according to the heuristics that have recommended the sentence

Label	Meaning
no	No heuristics recommend the sentence
H1	Only first heuristic recommends the sentence
H2	Only second heuristic recommends the sentence
H3	Only third heuristic recommends the sentence
H1+H2	Both first and second heuristics recommend the sentence
H1+H3	Both first and third heuristics recommend the sentence
H2+H3	Both second and third heuristics recommend the sentence
H1+H2+H3	All heuristics recommend the sentence

Taking Advantage of WEKA

How to handle numeric features?

Naïve-Bayes Case

- Assume a Normal Distribution (Gaussian)
  - Not always true
- Discretize
  - Fayyad & Irani Method (1993): Discretization with low loss of information
- Estimate the probabilistic distribution (Kernel Density Estimation, John & Langley, 1995)
  - Results at least as good as assuming a normal distribution

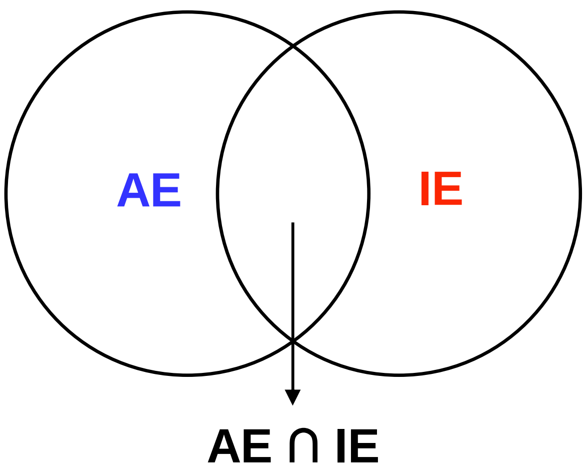
C4.5 Case

- Only choice is discretization!

Assessment

Characteristics

- Corpus TeMário (Rino & Pardo, 2003) – 100 news texts
- Same methodology of a former experiment (Rino et al., SBIA'04)
  - Compression Rate = 30% (extract length / source text length)
  - 10-fold cross validation
  - Compare automatic extracts (AE) with their corresponding ideal extracts (IE)
- Measures
  - Precision
  - Recall
  - F-measure



$$P = \frac{| AE \cap IE |}{| AE |}$$

$$R = \frac{| AE \cap IE |}{| IE |}$$

$$F = 2 \frac{P \times C}{P + C}$$

Assessment

Results

Model	Classifier	Numeric Handling	Feature Selection	Recall (%)	Precision (%)	F-measure (%)
M1	Naïve-Bayes	KDE	No	43,9	47,4	**45,6**
M2			CFS	42,8	46,6	44,6
M3		Discretization	No	42,2	45,8	43,8
M4			CFS	42,0	45,9	43,9
M5	C4.5	Discretization	No	37,7	40,6	39,1
M6			CFS	40,2	43,8	41,9

Best model = M1 → SuPor-2 !

Assessment

Comparing with former results (Rino et al., SBIA'04)

System	Precision (%)	Recall (%)	F-measure (%)	% above Random
SuPor-2	47,4	43,9	45,6	47
SuPor	44.9	40.8	42.8	38
ClassSumm	45.6	39.7	42.4	37
From-Top (B)	42.9	32.6	37.0	19
TF-ISF-Summ	39.6	34.3	36.8	19
GistSumm	49.9	25.6	33.8	9
NeuralSumm	36.0	29.5	32.4	5
Random order (B)	34.0	28.5	31.0	0

B = Baseline

Final Remarks

Some issues

- Why did Naïve-Bayes outperform C4.5?
  - Related to the way C4.5 calculates probabilities
  - NB performs well for ranking (Zhang & Su, 2004)
- Why didn't CFS bring better results overall?
  - Features got more informative → Feature Selection not needed anymore

Final Remarks

Overall results

• SuPor-2 → significant improvements over SuPor

• Expert user may not be necessary anymore → Using all features yields good results

Future work

• Explore new features

• New classifiers → especially probabilistic ones (e.g., Bayesian Networks)

•Improve even more features informativeness

Thank you!

Questions?

daniel_leite@dc.ufscar.br

References

Barzilay, R.; Elhadad, M. (1997). *Using Lexical Chains for Text Summarization*. In the Proc. of the Intelligent Scalable Text Summarization Workshop, Madri, Spain. Also In I. Mani and M.T. Maybury (eds.), Advances in Automatic Text Summarization. MIT Press, pp. 111-121, 1999.

Fayyad, Usama ; Irani, Keki. (1993). *Multi-interval discretization of continuous-valued attributes for classification learning*. In Proceedings of IJCAI'93.

Hall, M. (2000). *Correlation-based feature selection of discrete and numeric class machine learning*. In Proceedings of the International Conference on Machine Learning, pp. 359-366, San Francisco, CA. Morgan Kaufmann Publishers.

Hearst, M. (1997). TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics , 23 (1), pp. 33-64

John, G. ; Langley, P. (1995). *Estimating continuous distributions in Bayesian classifiers*. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (pp. 338-345)

Kupiec, Julian ; Pedersen, Jan ; Chen, Francine (1995). *A trainable document summarizer*. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp 68-73.

Larocca Neto, J.; Santos, A. D.; Kaestner, C. A. A.; Freitas, A. A. (2000). Generating Text Summaries through the Relative Importance of Topics. In M. C. Monard and J. S. Sichman (Eds.), *Iberamia-Sbia 2000*, pp. 300-309. Springer-Verlag, Berlin, Heidelberg.

Leite, D. S. ; Rino, L. H. M. (2006a). *A migração do SuPor para o ambiente WEKA: potencial e abordagens.* Série de Relatórios do NILC. NILC-TR-06-03. São Carlos-SP, Janeiro, 35p.

References

Leite, D. S.; Rino, L.H.M. (2006b). *SuPor: extensões e acoplamento a um ambiente para mineração de dados*. Série de Relatórios do NILC. NILC-TR-06-07. São Carlos – SP, Agosto, 22 p.

Módolo, M. (2003). *SuPor: an Environment for Exploration of Extractive Methods for Automatic Text Summarization for Portuguese* [in Portuguese]. MSc. Dissertation. Departamento de Computação, UFSCar.

Pardo, T.A.S. e Rino, L.H.M. (2004). *Descrição do GEI - Gerador de Extratos Ideais para o Português do Brasil*. Série de Relatórios do NILC. NILC-TR-04-07. São Carlos-SP, Agosto, 10p.

Pardo, T.A.S.; Rino, L.H.M. (2003). *TeMário: Um Corpus para a Sumarização Automática de Textos*. NILC Tech. Report. NILC-TR-03-09. São Carlos, Outubro, 12p.

Quinlan, J.R. (1993). *C4.5 Programs for machine learning*. San Mateo, Morgan-Kaufman, 1993.

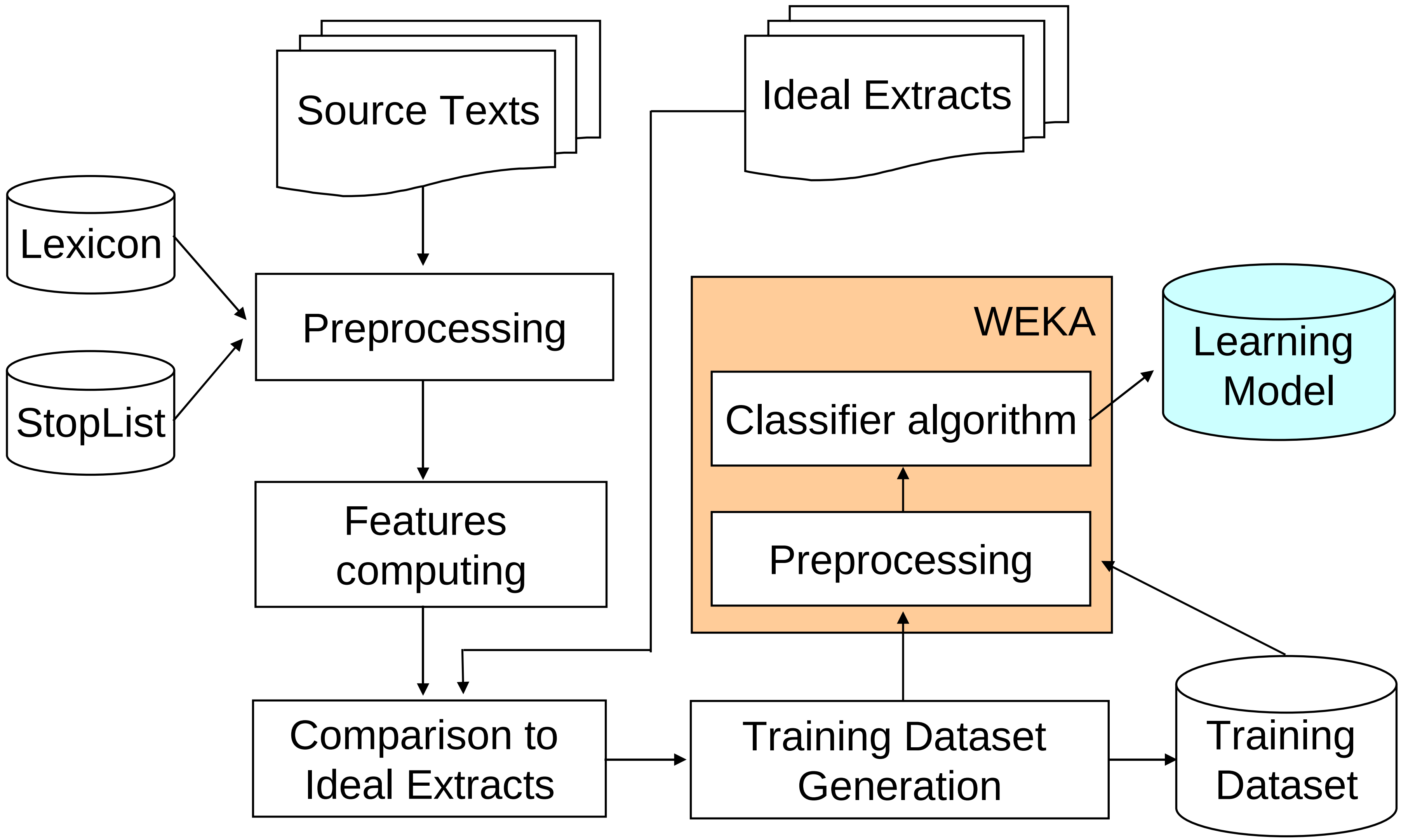
Rino, L.H.M.; Pardo, T.A.S.; Silla Jr., C.N.; Kaestner, C.A.; Pombo, M. (2004). *A Comparison of Automatic Summarization Systems for Brazilian Portuguese Texts*. In the Proceedings of the XVII Brazilian Symposium on Artificial Intelligence - SBIA2004. São Luís, Maranhão, Brazil.

Witten, Ian H. ; Frank, Eibe (2005). *Data Mining: Practical machine learning tools and techniques*, 2ª Ed., Morgan Kaufmann, San Francisco.

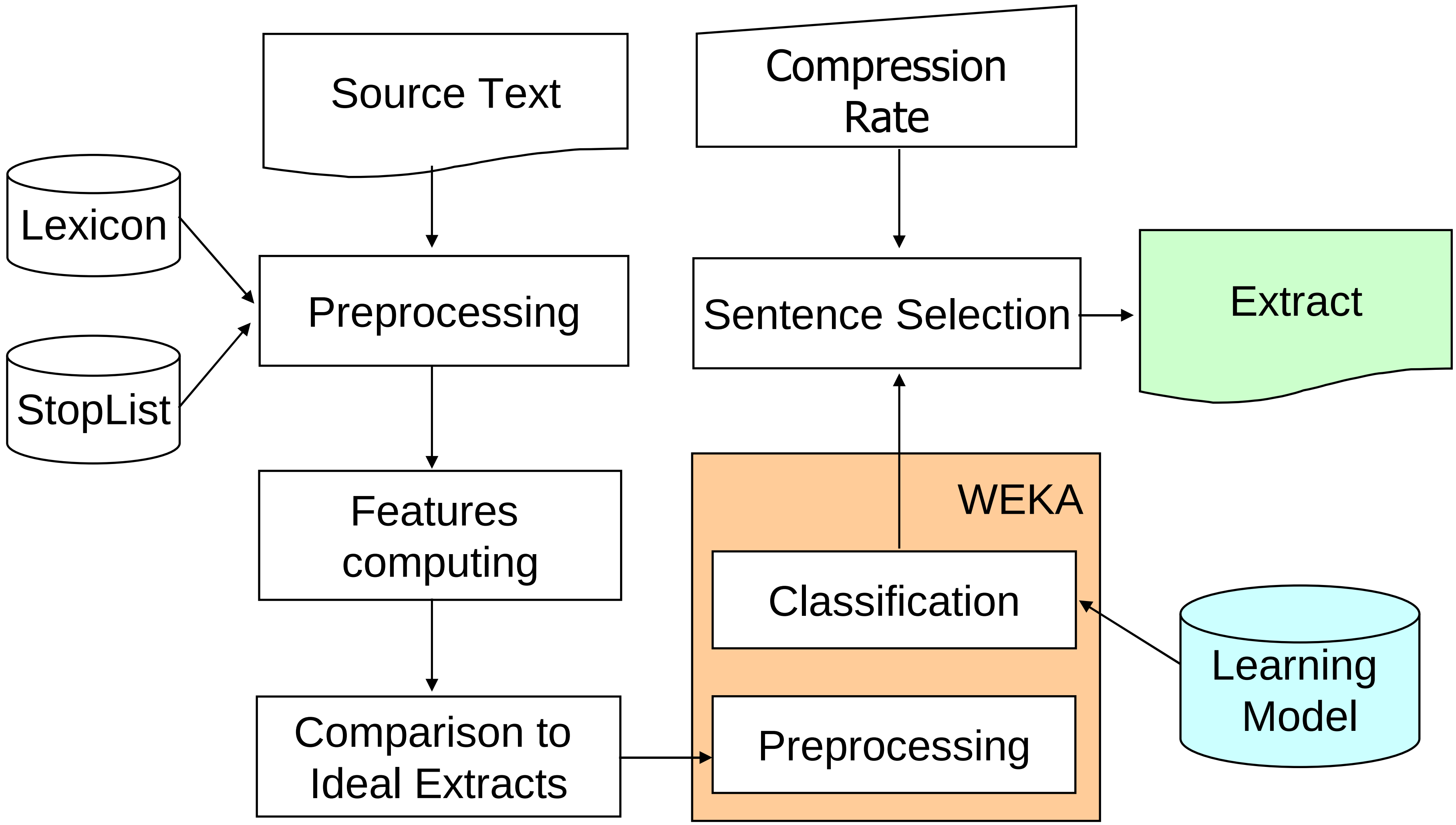
Zhang, H. ; Su, J. (2004). *Naive Bayesian classifiers for ranking*. Proceedings of the 15th European Conference on Machine Learning (ECML2004), Springer.

Salton, G.; Singhal, A.; Mitra, M.; Buckley, C. (1997). Automatic Text Structuring and Summarization. *Information Processing & Management*, 33(2), pp. 193-207.

SuPor-2 Architecture: Training Phase



SuPor-2 Architecture: Sentence Selection Phase



χ^2 Analysis