Learning to Order Terms: 
Supervised Interestingness Measures in Terminology Extraction

Jérôme Azé, Mathieu Roche, Yves Kodratoff, and Michèle Sebag

Abstract—Term Extraction, a key data preparation step in Text Mining, extracts the terms, i.e. relevant collocation of words, attached to specific concepts (e.g. genetic-algorithms and decision-trees are terms associated to the concept “Machine Learning”). In this paper, the task of extracting interesting collocations is achieved through a supervised learning algorithm, exploiting a few collocations manually labelled as interesting/not interesting. From these examples, the ROGER algorithm learns a numerical function, inducing some ranking on the collocations. This ranking is optimized using genetic algorithms, maximizing the trade-off between the false positive and true positive rates (Area Under the ROC curve). This approach uses a particular representation for the word collocations, namely the vector of values corresponding to the standard statistical interestingness measures attached to this collocation. As this representation is general (over corpora and natural languages), generality tests were performed by experimenting the ranking inducement learned from an English corpus in Biology, onto a French corpus of Curriculum Vitae, and vice versa, showing a good robustness of the approaches compared to the state-of-the-art. As the representation considered is domain and language independent, generality tests were performed by experimenting the ranking function learned from one corpus to another. The paper ends with perspectives for further research.

Keywords—Text-mining, Terminology Extraction, Evolutionary algorithm, ROC Curve.

I. INTRODUCTION
Besides the known difficulties of data mining, text mining presents specific difficulties due to the structure of documents and natural language. In particular, the construction of ontologies or terminologies [2,16] which is a central task in text mining, aims at controlling the polysemy and synonymy of words by structuring the words and their meanings in the application domain.

A preliminary step for ontology construction is to extract the domain terms, or words collocations [2,16,23]. Terms extraction involves two tasks: detecting “interesting” collocation of words (terms) and classifying them according to classes predefined by an expert.

This paper focuses on the detection of interesting terms, and more precisely on defining a ranking criteria on the words collocations. Based on [13], this paper formalizes an interestingness measure as a solution of some supervised learning problem (Learning to Order Things, [6]), or optimization problem. Actually, an interestingness measure, or ranking hypothesis, is assessed from its recall-precision trade-off, measured with respect to its Receiver Operating Characteristics (ROC) curve. Accordingly, a ranking function is learned by optimizing the area under the ROC curve (AUC) [11,14] from a few words collocations labelled as relevant/irrelevant by an expert.

The paper is organised as follows. Section II briefly reviews the main criteria used in terms extraction. Section III presents the ROGER (ROC-based GEnetic learner) algorithm, and its extension to the construction of interestingness measures are presented. Section IV reports on the experimental validation on two real-world corpora, and discusses the results obtained with respect to the state-of-the-art. As the representation considered is domain and language independent, generality tests were performed by experimenting the ranking function learned from one corpus to another. The paper ends with perspectives for further research.

II. TERMS EXTRACTION MEASURES
Different statistical criteria are used in systems of terminology extraction, for instance ACABIT [8] uses loglikelihood measure [9] and KEA [27] uses TF x IDF measure. The statistical criteria (value of the measures and the rank of each collocation) used in our approach are:

- Mutual Information ($MI$) [5]
- Mutual Information with cube ($MI^3$) [7]
- Dice Coefficient ($Dice$) [24]
- Loglikelihood ($L$) [9]
- Number of occurrences + Loglikelihood ($Occ_L$) [18]

The choice of an interestingness measure, mostly tackled in the literature through statistical and linguistic criteria [7,17,28] is currently viewed as a decision making problem.

Another approach based on learning an interestingness measure, is proposed by Vivaldi et al. [26]. They represent collocations from the values of the statistical criteria and use Adaboost [20] to automatically construct a discriminant hypothesis.

The presented work follows [26] with two main differences

LRI, Univ. Paris XI, 91405 Orsay Cedex, France. {aze,roche,yk,sebag}@lri.fr

\[ Occ_L = \text{defined by ranking terms according to their number of occurrences, and breaking the ties based on the term likelihoods.} \]
i) measures (Dice, Occ) are added to the description of collocations; ii) the learning problem is one of preference learning [13] instead of discriminant learning.

III. OVERVIEW

A. Linear ranking function

The ROGER algorithm [21,22] tackling the AUC optimization using evolution strategies, is among the most efficient evolutionary algorithms for numerical optimization [1]. ROGER investigates the space of continuous hypotheses, mapping the example space onto the real-valued space $\mathbb{R}$. Using the standard notations, the dataset $\mathcal{D} = \{(x_i, y_i), i=1..n\}$ includes $n$ examples, where each example (collocation) is described from the values of the statistical measures such as $\text{Fscore}$ [4].

The ROC curve depicts the trade-off between both objectives achieved by a learning algorithm and represented in the False Positive / True Positive Ratios plane. The ideal hypothesis corresponds to point $(0,1)$, with no false positive and 100% true positive examples.

The area under the ROC curve (AUC) is thus viewed as a global measure of the learning efficiency. As noted by [14], the averaging of randomized hypotheses can exponentially amplify their advantage over the default statistics.

The bias and variance of the AUC criterion have been studied by [19] and compared to the criteria of the misclassification error. An analytical and empirical study suggests that though the AUC bias might be higher than for the misclassification cost, its variance is lower; this can be explained as AUC is an order $n^2$ statistics, $n$ being the number of examples, whereas the misclassification cost is an order $n$ statistics.

The optimization of AUC constitutes a NP-complete problem, which has been undertaken in the literature in a number of ways, from evolutionary programming of neural nets [12] to greedy optimization of decision trees [11]. Recently, this problem was turned into a differentiable optimization problem by encapsulating the comparison of any two examples into a sigmoid function [14], and resolved by a gradient-based approach.

In earlier works [21,22], ROGER was exploring the space of linear hypotheses on $\mathbb{R}^d$. To each genotype $w=(w_1,\ldots,w_d) \in \mathbb{R}^d$ is associated a hypothesis $h_w$ defined on $\mathbb{R}^d$ as:

$$h_w(x) = \langle w, x \rangle = \sum_{j=1}^{d} w_j x_j$$

The fitness $F(h_w)$ is defined as the fraction of pairs of (positive, negative) examples that are ranked correctly according to $h_w$:

$$F(h_w) = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{T} \mathbb{1}(h_w(x_i) > h_w(x_j))$$

This section presents experimental setting, and discusses the results obtained.

A. Experimental setting

1) Optimization

In all experiments, BAGGED-ROGER involves the bagging of hypotheses extracted along 21 independent runs, using a

$$Bh(x) = \text{Median}\{h(x), t=1..T\}$$

Only BAGGED-ROGER will be considered in the following. Both linear and non linear hypotheses search space.
improves on the representation of SVMs using either linear, Gaussian and quadratic (using default options), the average AUC of ranking hypotheses.

### TABLE II: AVERAGE AUC OF RANKING HYPOTHESES BASED ON STATISTICAL CRITERIA.

<table>
<thead>
<tr>
<th></th>
<th>Occ</th>
<th>L</th>
<th>MI</th>
<th>Dice</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>0.57</td>
<td>0.42</td>
<td>0.35</td>
<td>0.31</td>
<td>0.30</td>
</tr>
<tr>
<td>CV</td>
<td>0.58</td>
<td>0.43</td>
<td>0.40</td>
<td>0.39</td>
<td>0.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Linear</th>
<th>Non linear</th>
<th>Linear</th>
<th>Gaussian</th>
<th>Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td>0.61 ± 0.04</td>
<td>0.67 ± 0.05</td>
<td>0.51 ± 0.13</td>
<td>0.54 ± 0.12</td>
<td>0.32 ± 0.07</td>
</tr>
<tr>
<td>CV</td>
<td>0.59 ± 0.10</td>
<td>0.61 ± 0.11</td>
<td>0.46 ± 0.13</td>
<td>0.42 ± 0.14</td>
<td>0.32 ± 0.07</td>
</tr>
</tbody>
</table>

### C. Generality tests

Finally, we take advantage of the fact that the representation of collocations is domain independent. This allows us to use a model learned from one corpus onto another one (different domains and/or languages).

Table III and Figure I demonstrate the good accuracy of the ranking functions, respectively learned by BAGGED-ROGER and SVM, when learned from a dataset and applied on another dataset. The unexpected robustness of the approach suggests that the representation of collocations provided by statistical measures is sufficiently precise to allow for discrimination. Further research (see next section) is concerned with investigation of this representation in more depth.

These results surprisingly show that ranking functions extracted from Biology behave well on the CV corpus, for both BAGGED-SVM and BAGGED-ROGER.

The tentative interpretation offered for this finding is related to the fact that the biology dataset is much better represented than the CV dataset. However, BAGGED-ROGER also features a good generality of the ranking function extracted from the CV when applied on Biology (compared to SVM). This better robustness might be explained from the stability of the model involving the vote of ten (extracted along the 10-fold Cross Validation) hypotheses involving the bagging of 21 hypotheses each.

Other results are presented in the web page: http://www.lri.fr/ia/fdt/Roger.

### TABLE III: GENERALITY TEST, AUC: LEARNING/TESTING WITH DIFFERENT CORPORA.

<table>
<thead>
<tr>
<th></th>
<th>BAGGED-ROGER</th>
<th>BAGGED-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear</td>
<td>Non linear</td>
</tr>
<tr>
<td>Biology</td>
<td>0.63</td>
<td>0.71</td>
</tr>
<tr>
<td>CV</td>
<td>0.64</td>
<td>0.63</td>
</tr>
</tbody>
</table>

### V. CONCLUSION AND PERSPECTIVES

This paper claims that supervised learning can significantly improve the task of term extraction, by learning an estimated relevance function from a few terms manually labelled as interesting / not interesting by the expert.

The approach combines three main features: i) the numerical representation of the examples (collocations) described from the values of a set of standard statistical interestingness measures; ii) a learning optimization criterion, based on the Wilcoxon statistics (area under the ROC curve); iii) the bagging of the various relevance functions learned...
along independent runs of a genetic algorithm, optimizing the
above criterion on the training set.

Experimental validation compared to state-of-the-art
machine learning algorithms, shows the robustness of the
above approach. Interestingly, the set of interestingness
measures provides a domain- and language-independent
description of the collocations, which allows for exploiting the
relevance function learned from one corpus onto another
corpus. Generality tests performed across two corpora show
that the performances of the relevance function are gracefully
degraded as it applies on a corpus in another domain, and,
which was even more unexpected, in another language.

The key question opened by this work is whether the set of
current interestingness measures provides enough information
to discriminate the interesting collocation, and accurately
learn the (subjective) interestingness measure of the expert.
This question must be answered considering more corpora;
however, such experimental validations are limited as they
require that the expert manually labels all collocations which
is hardly feasible when the fraction of interesting collocations
is low, which is the usual case.

Further work is concerned with enriching the representation
of collocations using non directly discriminant, but possibly
elementary, attributes: distance to the nearest typographic signs,
π distance to the nearest other collocation.

ACKNOWLEDGMENT

We thank Mary Felkin for her English review, Oriane
Matte-Tailliez for the expertise of the terms in Molecular
Biology, and PASCAL (Pattern Analysis, Statistical
Modelling and Computational Learning) Network of
Excellence for its support.

REFERENCES
[2] D. Bourguilt and C. Jacquierin, Term Extraction + Term Clustering:
of EACL, Bergen, pp. 15-22, 1999".
Scores for Word Association," The Thilisi Symposium on Logic,
[10] R. Esposito and L. Saitta, "Monte Carlo Theory as an Explanation of
Bagging and Boosting," Proc. of International Joint Conference on
using the area under the ROC curve," Proc. of International Conference
on Machine Learning (ICML), pp. 139-146, 2002.
Networks for Detecting Breast Cancer," Cancer Letters, vol. 96, pp. 49-
Algorithm for Combining Preferences", Journal of Machine Learning
Imbalance Problem?" Workshop on Learning from Imbalanced Data Sets II (ICML), 2003
[16] G. Nenadic, H. Mima, I. Spasic, S. Ananiadou, and J. Tasiuji,
"Terminology-based Literature Mining and Knowledge Acquisition in
association rules discovery in a technical corpus," Proc. of HIWSM'04,
[22] M. Sebag, N. Lucas, and J. Azé, "Impact studies and sensitivity analysis
in medical data mining with ROC-based genetic learning," Proc. of IEEE
[23] F. Smadja, "Retrieving collocations from text: Xtract," Computational
Linguistics, vol. 19, no. 1, pp. 143-177, 1993
[26] J. Vivaldi and L. Marquez and H. Rodriguez, "Improving Term
Extraction by System Combination Using Boosting," Lecture Notes in
[27] L.H. Witten, G.W. Paynter, E. Frank, C. Gutwin, and C.G. Nevill-
Manning, "Practical automatic keyphrase extraction. Proc. of DL
Approach to Automatic Acquisition of Domain Relevant Terms and their
Relations with Bootstrapping." Proc. of LREC 2002, the third international conference on language resources and evaluation, 2002.