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 function 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 Support Vector Machine (SVM).

***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)[1] [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

[1] Occ_L is defined by ranking terms according to their number of occurrences, and breaking the ties based on the term likelihoods.

i) measures (*Dice*, Occ_L) 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 $\Re$. Using the standard notations, the dataset $\varepsilon=\{(x_i,y_i),\ i=1..n,\ x_i\in\Re^d,\ y_i\in\{-1,+1\}\}$ includes n examples, where each example (collocation) is described from the values of the statistical criteria. This section describes the algorithm used to learn a term ranking hypothesis. We first briefly review the state-of-the-art related to ROC analysis in Machine Learning.

For the sake of simplicity, let us restrict the discussion to supervised binary learning. The goal of learning algorithms can actually be seen as a multi-objective optimization problem: maximizing the rate of true positive examples (percentage of positive examples correctly classified) while minimizing the rate of false positive examples (percentage of negative examples misclassified as positive).

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.

ROC curve has no sensitivities to the ratio of positives and negatives examples [15] as opposed to other accuracy measures such as Fscore [4]. One advantage of ROC curves is to naturally accommodate ill-balanced distributions and cost-sensitive learning [8].

The area under the ROC curve (AUC) is thus viewed as a global measure of the learning efficiency. As noted by [14], the area under the ROC curve is equivalent to the Wilcoxon rank statistics, the probability of ranking correctly a pair of (positive, negative) examples. Indeed the probability of ranking an interesting collocation below a non-interesting one constitutes an appropriate evaluation for an interestingness measure.

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 $\Re^d$. To each genotype $w=(w_1,...,w_d)\in\Re^d$ is associated a hypothesis h_w defined on $\Re^d$ as: $h_w(x)=\langle w,x\rangle=\Sigma\ w_j\ .\ x^j$.

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

$$F(h_w)=\Pr(h_w(x_i)>h_w(x_j)\,|\,y_i>y_j) \qquad (1)$$

B. Non linear ROGER

Thank to the flexibility of evolutionary computation a straightforward extension allows for considering (a limited kind of) non-linear hypotheses by only doubling the size of the search space. Specifically, a genetic individual $z=(w_1,...,w_d,c_1,...,c_d)\in\Re^{2d}$ is associated with the hypothesis h_z defined as:

$$h_z(x=(x^1,...,x^d))=\sum_{j=1}^{d} w_j\cdot\left|x^j-c_j\right|$$

The associated fitness $F(h_z)$ is computed as in equation (1).

In both cases, the optimization of F is achieved by a $(\mu+\lambda)$ evolution strategy, using self-adaptive mutation and uniform crossover [1].

C. BAGGED-ROGER

This paper presents a new extension of ROGER named BAGGED-ROGER which is based on the remark that independent runs of an evolutionary learning algorithms provide diverse hypotheses, namely the hypothesis reaching the best AUC value along each run.

Although these hypotheses cannot be considered truly independent as they are optimized on the same training set, it makes sense to consider their combination [3]. As shown in [10], the averaging of randomized hypotheses can exponentially amplify their advantage over the default accuracy. Formally, let $h_1,...,h_T$ denote the T normalized[2] hypotheses constructed along T independent runs of ROGER. Their agregation noted Bh, is defined as:

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

Only BAGGED-ROGER will be considered in the following. Both linear and non linear hypotheses search space.

IV. EMPIRICAL VALIDATION

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

[2] Note that these hypotheses are normalized, i.e. $\Sigma|w_j|=1$.

(20+200) Evolution Strategy[3].

Results are assessed using 10-fold stratified Cross Validation. On each training set, 21 independent ROGER runs are launched. The final ranking function is obtained by bagging (median value) the ranking functions learned over the 10 folds. This ranking function allows us to determine the rank of each example.

2) Data preparation

Two corpora were considered, respectively related to Molecular Biology (in English) and Curriculum Vitae (in French). After a first data preparation: normalisation, Part-Of-Speech Tagging (due to space limit, the interested reader is referred to [17] for more detail), we consider different collocations (e.g., *Noun-Noun*, *Adjective-Noun*, *Noun-Preposition-Noun*, etc.). For the two corpora, we only consider the more frequent type of collocations.

a) Molecular Biology

A first application was considered, within the domain of Molecular Biology. A 9,4 Mo corpus in English, composed of 6,119 abstracts, was gathered by querying Medline[4]. Only 1028 *Noun-Noun* collocations occurring at least 4 times have been considered in this application. These collocations have been labelled by a domain expert (TABLE I).

b) Curriculum Vitae (CV)

The second application aims at the automatic analysis of a Curriculum Vitae corpus, in French, provided by the VediorBis Foundation. The corpus involves 582 documents (952 Ko). All 376 *Noun-Adjective* collocations appearing at least 3 times in the documents have been manually and independently labelled by two experts (see TABLE I).

It must be noted that the proportion of interesting terms is very high in both datasets, which is why we could ask an expert to label them. The situation is entirely different when rare collocations are considered (see [18]).

TABLE I : MOLECULAR BIOLOGY

Frequent collocations	# collocations	relevant	irrelevant
Biology	1028	90.9%	9.1%
CV	376	85.7%	14.3%

B. Comparative validations

Table II shows the predictive accuracy of linear and non linear BAGGED-ROGER, compared to that of linear, Gaussian and quadratic Support Vector Machine using the state-of-the-art SVMtorch software[5].

Table II shows that BAGGED-ROGER using either linear or non-linear hypotheses significantly improves on all statistical measures, when compared to the state-of-the-art Machine Learning algorithm (BAGGED-SVM). Unexpectedly, it also 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.

	Occ_L	L	MI^3	*Dice*	*MI*
Biology	0.57	0.42	0.35	0.31	0.30
CV	0.58	0.43	0.40	0.39	0.31
	BAGGED-ROGER		**BAGGED-SVM**		
	Linear	*Non linear*	*Linear*	*Gaussian*	*Quadratic*
Biology	0.61 ± 0.04	**0.67** ± 0.05	0.51 ± 0.13	0.54 ± 0.12	0.32 ± 0.07
CV	0.59 ± 0.10	**0.61** ± 0.11	0.46 ± 0.13	0.42 ± 0.14	0.52 ± 0.07

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, BAGGER-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 : GENERALIY TEST, AUC: LEARNING/TESTING WITH DIFFERENT CORPORA.

	BAGGED-ROGER		**BAGGED-SVM**		
	Linear	*Non linear*	*Linear*	*Gaussian*	*Quadratic*
CV → Biology	0.63	0.71	0.59	0.42	0.48
Biology → CV	0.64	0.63	0.64	0.61	0.46

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

[3] 20 parents generate 200 offsprings using self adaptative mutation and uniform crossover with crossover rate 60%. The best 20 individuals among the parents and offsprings form the next population.

[4] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[5] http://www.idiap.ch/machine_learning.php?content=Torch/en_OldSVMTorch.txt

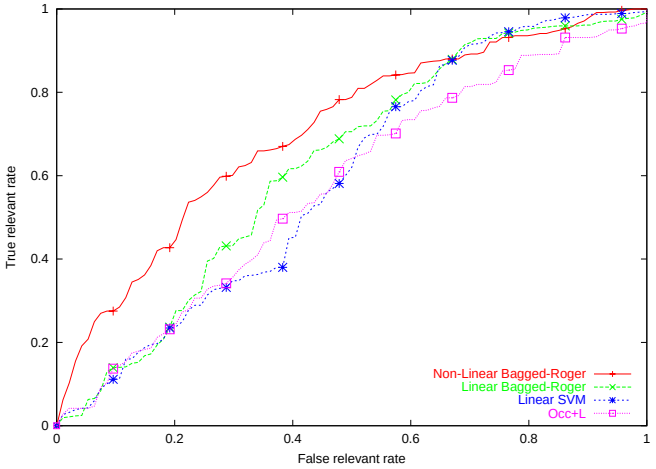
along independent runs of a genetic algorithm, optimizing the above criterion on the training set.

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 relevant, attributes: distance to the nearest typographic signs, or distance to the nearest other collocation.

FIG. I : ROC CURVES WITH RANKING FUNCTIONS LEARNED WITH THE CV APPLIED ON THE MOLECULAR BIOLOGY CORPUS.



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

[1] T. Bäck, Evolutionary Algorithms in theory and practice, 1995.

[2] D. Bourigault and C. Jacquemin, "Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology," *Proc. of EACL*, Bergen., pp. 15-22, 1999".

[3] L. Breiman, "Arcing Classifiers," *Annals of Statistics*, vol. 26, no. 3, pp. 801-845, 1998.

[4] R. Caruana and A. Niculescu-Mizil, "Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria". *Proc. of "ROC Analysis in AI" Workshop ECAI*, pp 9-18, 2004.

[5] K.W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, vol. 16, pp. 22-29, 1990.

[6] W. Cohen, R. Schapire, and Y. Singer, "Learning to Order Things," *Journal of Artificial Intelligence Research*, vol. 10, 243-270, 1999.

[7] B. Daille, E. Gaussier, and J.M. Langé, "An Evaluation of Statistical Scores for Word Association," *The Tbilisi Symposium on Logic, Language and Computation, CSLI Publications*, pp. 177-188, 1998.

[8] P. Domingos, "Meta-Cost: A general method for making Classifiers Cost Sensitive," *Knowledge Discovery from Databases*, pp. 155-164, 1999.

[9] T.E. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, vol. 19, n°1, pp. 61-74, 1993.

[10] R. Esposito and L. Saitta, "Monte Carlo Theory as an Explanation of Bagging and Boosting," *Proc. of International Joint Conference on Artificial Intelligence,* pp. 499-504, Morgan Kaufman Publishers, 2003.

[11] C. Ferri, P. Flach, and J. Hernandez-Orallo, "Learning decision trees using the area under the ROC curve," *Proc. of International Conference on Machine Learning (ICML)*, pp. 139-146, 2002.

[12] D.B. Fogel, E.C. Wasson, and E.M. Boughton, "Evolving Neural Networks for Detecting Breast Cancer," *Cancer Letters*, vol. 96, pp. 49-53, 1995.

[13] *Y. Freund, R. Iyer,* R. E. Schapire, Y. Singer, "An Efficient Boosting Algorithm for Combining Preferences", *Journal of Machine Learning Research*, 4(Nov):933-969, 2003.

[14] R. Jin, Y. Liu, L. Si, J. Carbonell, and A. Hauptmann, "A New Boosting Algorithm Using Input-Dependent Regularizer," *Proc. of International Conference on Machine Learning (ICML)*, *AAAI Press*, 2003.

[15] A. Kolcz, A. Chowdhury, J. Alspector, "Data duplication: An Imbalance Problem?" *Workshop on Learning from Imbalanced Data Sets II (ICML)*, 2003

[16] G. Nenadic, H. Mima, I. Spasic, S. Ananiadou, and J. Tsujii, "Terminology-based Literature Mining and Knowledge Acquisition in Biomedicine", *International Journal of Medical Informatics*, vol. 67, pp 33-48, 2002.

[17] M. Roche, J. Azé, O. Matte-Tailliez, and Y. Kodratoff, "Mining texts by association rules discovery in a technical corpus," *Proc. of IIPWM'04, Springer Verlag*, pp. 89-98, 2004.

[18] M. Roche, J. Azé, Y. Kodratoff and M. Sebag, "Learning Interestingness Measures in Terminology Extraction. A ROC-based approach," *Proc. of "ROC Analysis in AI" Workshop ECAI*, pp 81-88, 2004.

[19] S. Rosset, "Model Selection via the AUC," *Proc. of International Conference on Machine Learning (ICML)*, 2004.

[20] R.E. Schapire, "Theoretical views of boosting," *Proc. of European Conference on Computational Learning Theory*, pp. 1-10, 1999.

[21] M. Sebag, N. Lucas, and J. Azé, "ROC-based Evolutionary Learning: Application to Medical Data Mining," *Proc. of International Conference on Artificial Evolution (EA), Springer Verlag*, pp. 384-396, 2004.

[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 International Conference on Data Mining (ICDM)*, pp. 637-640, 2003.

[23] F. Smadja, "Retrieving collocations from text: Xtract," *Computational Linguistics*, vol. 19, no. 1, pp. 143-177, 1993

[24] F. Smadja, K. R. McKeown, and V. Hatzivassiloglou, "Translating collocations for bilingual lexicons: A statistical approach," *Computational Linguistics*, vol. 22, n°1, pp. 1-38, 1996.

[25] V.N. Vapnik, "The Nature of Statistical Learning," Springer Verlag, 1995.

[26] J. Vivaldi and L. Marquez and H. Rodriguez, "Improving Term Extraction by System Combination Using Boosting," *Lecture Notes in Computer Science, vol 2167*, pp. 515-526, 2001.

[27] I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin, and C.G. Nevill-Manning. Kea: Practical automatic keyphrase extraction. *Proc. of DL '99*, pp. 254-256, 1999.

[28] F. Xu, D. Kurz, J. Piskorski, and S. Schmeier, "A Domain Adaptive 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.

