Learning Interestingness Measures in Terminology Extraction
A ROC-based approach

Mathieu Roche and **Jérôme Azé** and **Yves Kodratoff** and **Michèle Sebag** [1]

Abstract. In the field of Text Mining, a key phase in data preparation is concerned with the extraction of terms, i.e. collocation of words attached to specific concepts (e.g. *Philosophy-Dissertation*). In this paper, Term Extraction is formalized as a supervised learning task, extracting a ranking hypothesis from a set of terms labeled as relevant/irrelevant by the expert. This task is tackled using the evolutionary algorithm *ROGER*, optimizing the area under the ROC curve attached to a ranking hypothesis.
Empirical validation on two real-world applications demonstrates outstanding improvements compared to state-of-art interestingness measures in Term Extraction. The approach is found robust across domains (Molecular Biology, Curriculum Vitæ) and languages (English, French).

1 INTRODUCTION

Besides the known difficulties of data mining [19], text mining presents specific difficulties due to the structure of documents and natural language [15, 20]. In particular, the construction of ontologies or terminologies [4, 22], a central task in text mining, aims at controlling the polysemy and synonymy phenomenons through structuring the words and their meanings in the application domain.

A preliminary for ontology construction is to extract the domain terms, or word collocations [4, 18, 22, 30]. Indeed, the meaning of a term (e.g. *attribute-value representation*) is not related to the meaning of its components in a simple way. Therefore, terms must be extracted to enable the conceptual analysis of the corpus documents.

Term extraction involves two tasks: detecting "interesting" collocation of words (candidate terms); classifying them according to classes predefined by the expert.

This paper focuses on the detection of interesting terms, and more precisely on defining an interestingness measure on the word collocations. The choice of an interestingness measure, mostly tackled in the literature through statistical and linguistic criteria [11, 23, 33], is currently viewed as a decision making problem, where the user has to decide which one among a number of existing criteria, is the most appropriate to her goals.

Taking its inspiration in [10], and after [32], this paper proposes instead to formalize an interestingness measure as a solution of some supervised learning problem (*Learning to Order Things*), or optimization problem. Actually, an interestingness measure, or ranking hypothesis, is assessed from its recall-precision tradeoff, measured with respect to its Receiver Operating Characteristics (ROC) curve (see [5] for ample motivations about the use of ROC curves in supervised learning). Another way to evaluate an interestingness measures is using the lift chart. The lift chart measures the variation of the precision as a function of the proportion of terms found by the system.

Accordingly, a ranking function is learned by optimizing the area under the ROC curve (AUC) [16, 21] from training examples, word collocations labelled as relevant/irrelevant by the expert.

An earlier approach, concerned with the evaluation of medical risk factors [28, 29], used a genetic algorithm termed *ROGER* (for *ROc-based GEnetic learneR*) to construct linear hypotheses maximizing the AUC criterion. In this paper, *ROGER* is extended in two ways. Firstly, exploiting the flexibility of genetic search, *ROGER* is generalized to construct non-linear hypotheses; secondly, taking advantage of the stochastic nature of genetic search, the *bagging* of the ranking hypotheses constructed along independent runs is considered.

The approach is shown to significantly outperform the state-of-the-art statistical criteria in Term Extraction. The empirical validation on two corpus, related to distinct domains (Molecular Biology and Curriculum Vitæ) and written in distinct languages (English and French), reveals a good robustness across domains and languages. Specifically, the ranking hypotheses learned from one corpus appear to outperform the statistical criteria *on the other corpus*.

The paper is organised as follows. For the sake of completeness, section 2 briefly reviews the main criteria used in Term Extraction. Section 3 presents the *ROGER* algorithm, and its extension to the construction of interestingness measures. Section 4 describes our experimental setting and the goal of the experiments, comparing three representations for term extraction respectively involving: i) statistical features; ii) statistical features plus information retrieval features; iii) the above features plus linguistic features. Section 5 reports on the experimental validation on two real-world corpora, and discusses the results obtained with respect to the state-of-the-art. The paper ends with perspectives for further research.

2 STATE OF THE ART

Without pretending at an exhaustive review, this section presents the main criteria used in terminology extraction.

The corpus $\mathcal{E}$ is composed of a set of documents, where each document D is composed of a sequence of sentences, and each sentence is a sequence of words. Furthermore, each word noted x is labelled

[1] Laboratoire de Recherche en Informatique (LRI), CNRS UMR 8623, Université Paris-Sud, 91405 Orsay Cedex, France. Email:{roche,aze,yk,sebag}@lri.fr

with its type $t(x)$ extracted using a part-of-speech tagger [8, 1].

Only collocations typed as *Noun-Noun*, *Noun-Preposition-Noun*, *Noun-Adjective* and *Adjective-Noun* are considered throughout the paper.

2.1 Mutual Information (MI, MI^3)

Mutual information [9] measures the correlation of words in a collocation. Let $P(xy)$ denote the frequency of the collocation xy in the documents[2], then the mutual information of xy is given as:

$$MI(xy) = \log_2 \frac{P(xy)}{P(x) \times P(y)} \tag{1}$$

With no loss of generality, the same ranking function is obtained by replacing $P(xy), P(x)$ and $P(y)$ in the above formula by their number of occurrences, respectively noted $nb(xy), nb(x)$ and $nb(y)$.

A variant of Mutual Information noted MI^3 introduced by [11] increases the score of frequent collocations; as reported by [32], MI^3 is empirically very performant.

$$MI^3(xy) \quad = \quad \log_2 \frac{nb(xy)^3}{nb(x) \times nb(y)} \tag{2}$$

2.2 Dice coefficient ($Dice$)

The Dice coefficient [31] refines the correlation estimate, accounting for the specific types of the words in a collocation. Let $nt(x)$ define the number of words with same type as x in the documents (the number of word instances with same tag as x, e.g. the number of adjectives if x is an adjective).

$$Dice(xy) \quad = \quad \frac{nb(xy)}{nt(x) \times nb(y) + nb(x) \times nt(y)} \tag{3}$$

2.3 Loglikelihood (L)

Loglikelihood differs from the above measures as it takes into account the number of cases where none of the collocation words appear [13]. This measure is widely used in extraction terminology (see e.g. [11, 23, 33]).

Let us denote $nb(xy^\star)$ (respectively $nb(x^\star y)$) as the number of occurrences of x followed by $y' \neq y$ (resp. the number of occurrences of $x'y$ with $x' \neq x$), then the loglikelihood $L(xy)$ is defined up to a constant as:

$$\begin{aligned} L(xy) = \quad & K(xy) + K(xy^\star) + K(x^\star y) + K(x^\star y^\star) \\ & -K(x) - K(y) - K(x^\star) - K(y^\star) \\ with\ K(v) = \quad & nb(v) \times \log(nb(v)) \end{aligned} \tag{4}$$

Another widely used ranking function, referred to as Occ_L, is defined by ranking terms according to their number of occurrences, and breaking the ties based on the term likelihoods.

[2] A document contains a collocation iff all collocated words contiguously appear in at least one sentence of the document.

2.4 Information Retrieval-like measures

In the neighbour field of Information Retrieval [25], another widely used measure is known as *tf-idf* (term frequency - inverse document frequency). The *tf-idf* measure aims at filtering out the terms which are present in most documents. Specifically, if $P_j(xy)$ is the frequency of term xy in document D_j, then the weight $W_j(xy)$ of xy for D_j is given as

$$W_j(xy) = -P_j(xy) \times \log_2 P(xy) \tag{5}$$

Therefore, a high *tf-idf* score is obtained when a term is frequent in at least one document ($P_j(xy)$ is high) and appears in few documents ($-\log_2(P(xy))$ is high).

Contrasting with Information Retrieval, Term Extraction is mainly interested in terms appearing in many documents. For this reason, an IR-inspired criterion referred to as *term-and-document-frequency* (*tdf*) measure, is defined as:

$$W'_j(xy) = -\frac{P_j(xy)}{\log_2 P(xy)} \tag{6}$$

2.5 Learning interestingness measures

Another approach proposed by [32] is concerned by learning a linear combination of the statistical criteria, based on a set of terms labelled as relevant/irrelevant by the expert. This approach uses AdaBoost algorithm [26].

Along the same lines, our goal is to find a ranking hypothesis, based on a propositional description of terms. Specifically, a term is represented as a vector, the components of which are i) the statistical criteria used in [32], ii) the *tf-idf* and *tdf* criterions, and iii) additional features, encapsulating shallow linguistic information.

As demonstrated by [6], such features (e.g. number of punctuation signs before or after a term) can bring significant improvements in text mining applications. In the following, two linguistic features are considered: the total number of punctuation signs *before* the word collocation (appearing between the beginning of the sentence and the occurrence of the collocation) and *after* (appearing between the end of the collocation occurrence and the end of the sentence).

The presented approach, similar in spirit to [32], presents two main differences. On one hand, the description of terms is enriched with additional features. On the other hand, this description is exploited through a ROC-based genetic algorithm described in the next section.

3 ROC-BASED INTERESTINGNESS MEASURES

This section describes the algorithm used to learn a term ranking hypothesis. This algorithm, termed *Bagged-ROGER*, extends the *ROGER* algorithm first described in [28, 29], which will presented in section 3.3 for the sake of completeness.

We first briefly review the state of the art related to ROC analysis in Machine Learning.

3.1 State of the art

The use of Receiver Operating Characteristics (ROC) curve to compare learning algorithms was first advocated in [5] to our best knowledge. Let us restrict ourselves to supervised binary learning for the sake of simplicity.

Then, 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). One advantage of this formalization is to naturally accommodate ill-balanced distributions and cost-sensitive learning [12].

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

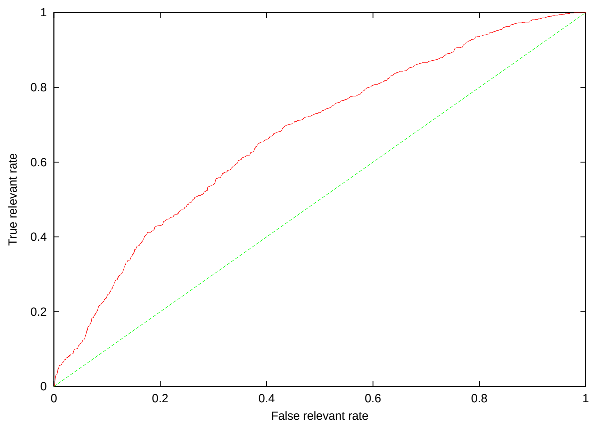

Figure 1. Example of ROC curves.

The area under the ROC curve (AUC) is thus viewed as a global measure of the learning efficiency. As noted by [21], 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 swapping two irrelevant/relevant instances constitutes an appropriate quality criterion for an interestingness measure.

The bias and variance of the AUC criterion have been studied by [24] and compared to that of the misclassification error. An analytical and empirical study suggests that, though the AUC bias might be higher, 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 tackled in the literature in a number of ways, ranging from evolutionary programming of neural nets [17], to greedy optimization of decision trees [16]. Recently, this problem was turned into a differentiable optimization problem by encapsulating the comparison of any two examples into a sigmoid function [21], and tackled by a gradient-based approach.

The *ROGER* algorithm [28, 29] tackles the AUC optimization using evolution strategies, among the most efficient evolutionary algorithms for numerical optimization [3, 27].

3.2 *ROGER*

ROGER investigates the space of continuous hypotheses, mapping the example space onto the real-valued space $\mathbb{R}$. After the standard notations, the dataset $\mathcal{E}$ is composed of n examples $(\mathbf{x}_i, y_i), i = 1..n$ where $\mathbf{x}_i \in X$ denotes the i-th example description ($X \subset \mathbb{R}^d$, d being the number of features) and y_i denotes the associated label ($y_i = \pm 1$).

In a first version [29], *ROGER* was exploring the space of linear hypotheses (hyperplanes) on $\mathbb{R}^d$. To each genotype $\mathbf{w} = (w_1, \ldots, w_d) \in \mathbb{R}^d$ is associated a (phenotype) hypothesis h_w defined on X as:

$$h_w(\mathbf{x}) = < \mathbf{w}, \mathbf{x} >$$

The fitness $\mathcal{F}(\mathbf{w})$ associated to genotype $\mathbf{w}$ is defined as the fraction of pairs of (positive, negative) examples that are ranked correctly according to h_w :

$$\mathcal{F}(w) = Pr(h_w(\mathbf{x}_i) > h_w(\mathbf{x}_j) \mid y_i > y_j) \tag{7}$$

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

$$h_z(\mathbf{x} = (x^1, \ldots, x^d)) = \sum_{j=1}^{d} w_j \times |x^j - c_j|$$

The associated fitness is computed as in equation (7).

In both cases, the optimization of $\mathcal{F}$ is achieved by an evolution strategy, using self-adaptive mutation and crossover with crossover rate .6; same parameter values were used in both cases.

3.3 *Bagged-ROGER*

Another extension of *ROGER* named *Bagged-ROGER* 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 [7]. As noted in [14], the averaging of randomized hypotheses can exponentially amplify their advantage over the default accuracy.

Formally, let $h_1, \ldots, h_T$ denote the T hypotheses constructed along T independent runs of *ROGER*, and normalized. Their agregation noted Bh, is defined as:

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

Only *Bagged-ROGER* will be considered in the following, as it significantly outperforms *ROGER* on the considered applications.

4 EXPERIMENTAL SETTING AND GOAL

The first goal of the experiments is to investigate the robustness of the approach comparatively with the existing interestingness measures (section 2).

A second goal is to study the impact on the performances of the representation of examples, including respectively:

i) the statistical features MI, MI^3, *Dice, Loglikelihood*, Occ_L;
ii) the above plus IR-like features (section 2.4);
iii) the above plus shallow linguistic features (section 2.5).

Along the same lines, we investigate the impact of the hypothesis space (linear, $\mathbb{R}^d$, or non-linear, $\mathbb{R}^{2d}$).

Last and overall, the generalization properties of the interestingness measures obtained are examined, considering:

i) a test set with different distribution as the training set;
ii) a test set extracted from another corpus: in another application domain, and in another language.

Experimental setting.

In all experiments, *Bagged-ROGER* (with linear and non-linear hypotheses) is implemented within a (20+200) Evolution Strategy, with 20 parents selected deterministically from 20 parents plus 200 offspring, using self-adaptive mutation, uniform crossover with crossover rate 60% [3].

Results are assessed using 10-fold stratified CV. On each training set, 20 independent *ROGER* runs are launched. The final ranking function is obtained by bagging the ranking functions learned over the 10 folds. This ranking function allows us to determine the rank of each example.

5 EXPERIMENTAL VALIDATION

Two corpora were considered, respectively related to Curriculum Vitæ(in French) and Medline abstracts (in English).

5.1 Curriculum Vitæ

The first application aims at the automatic analysis of Curriculum Vitæ, corpus kindly provided by the VediorBis Fundation (in French). After a first data preparation step detailed in [23], the corpus involves 582 documents (952 Ko). The study focuses on collocations typed as *Noun-Adjective*, which represent 44% of all collocations.

Two sets of collocations are defined: the frequent collocations (376 collocations appearing at least three times in the documents), and the rare collocations (all other 2822 collocations).

Table 1. Learning and Validation Datasets.

Data	# collocations	relevant *(class 2)*	irrelevant *(classes 0 and 1)*
Frequent Collocations (Learning dataset)	376	85.7%	14.3%
Rare Collocations (Validation dataset)	2822	56.6%	43.4%

The manual analysis of these documents leads to define several relevant topics (e.g. *Linguistic Competence, Commercial Competence, Management Activity*). All 376 frequent collocations are manually and independently labelled by two experts (2 hours each), classified in 4 classes :

-1 The expert cannot evaluate the collocation (*"MBA CLS"*)
0 Irrelevant collocation (*"management year"*)
1 Semantically relevant collocation, but not adapted to the expert focus (too general or too specific wrt the concepts defined) (*"summer holidays"*)
2 Semantically relevant collocations, relevant to the expert concepts (*"part time jobs"*)

Term extraction on frequent collocations

Tables 2 and 3 display the AUC performance obtained for linear and non-linear ranking hypotheses constructed with *Bagged-ROGER*, compared to that of the standard statistical measures. The corresponding ROC curves are displayed on Fig. 2.

Table 2. Frequent Collocations: Performance of ranking hypotheses based on statistical criteria.

Statistical Criteria	*AUC*
Occ_L	0.58
L	0.43
MI^3	0.40
$Dice$	0.39
MI	0.31

Table 3. Frequent Collocations: Performance of ranking hypotheses based on *Bagged-ROGER* with linear and non-linear hypotheses

Bagged-ROGER		
representation of the domain	*Linear AUC*	*Non-Linear AUC*
(i)	0.69	0.73
(ii)	0.69	0.74
(iii)	0.68	0.74

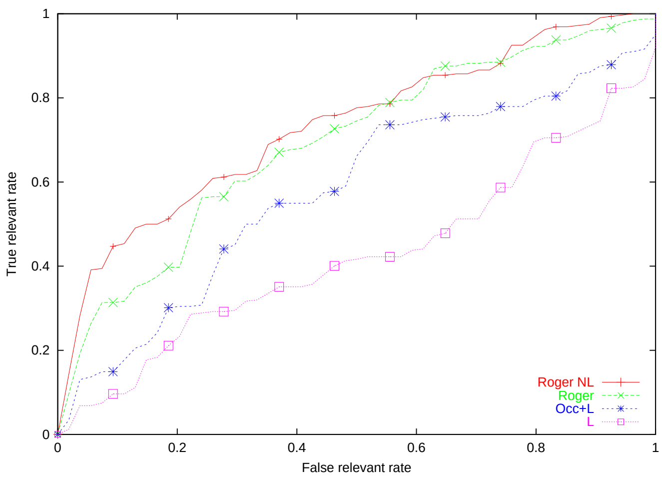

Figure 2. ROC Curves on Frequent Collocations (Linear and Non-linear *Bagged-ROGER*, Occ_L and L).

These experiments show that the best statistical criterion is the Occ_L measure. However, this criterion only slightly improves on the default hypothesis (which would correspond to the diagonal line $TP = FP$).

In opposition, *Bagged-ROGER* significantly improves on all statistical measures, using either linear or non-linear hypotheses. Interestingly, non-linear hypotheses appear slightly but significantly more accurate than linear hypotheses. The computational runtime is one hour on PC Pentium IV for each representation considered (i.e. a total of 200 *ROGER* runs).

The beginning of the ROC curve interestingly illustrates the trade-off between true relevant and false relevant terms in the top ranked terms. A more detailed picture is given by the lift chart (Fig. 3), plotting the precision (percentage of relevant terms) *versus* the fraction of selected terms. Fig. 3 shows that the term ranking hypothesis constructed by non-linear *Bagged-ROGER* is significantly better than the standard statistical criteria (e.g. out of the half top ranked collocations, 94.6% are relevant for *Bagged-ROGER* against 90% for Occ_L and 82% for L).

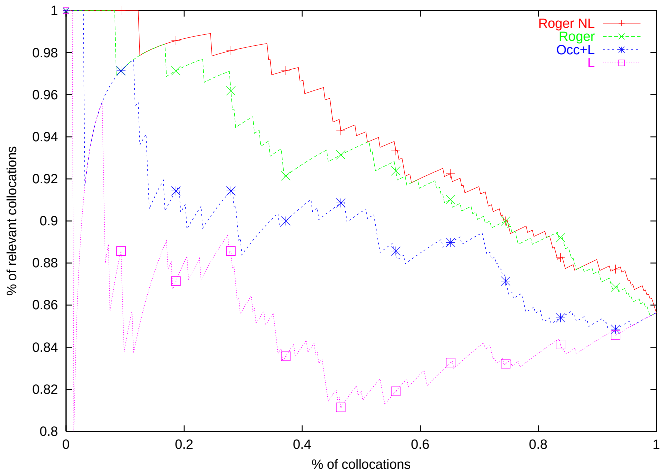

Figure 3. Lift Curves on Frequent Collocations (Linear and Non-linear *Bagged-ROGER*, Occ_L and L).

Validation on rare collocations

Although the number of infrequent collocations forbids their manual labelling in the general case, it is widely acknowledged in the Text Mining domain that rare cases provide valuable information [2]. For this reason, the ranking hypothesis learned from the frequent collocations was finally tested on the rest of the collocations, the rare ones (2822 collocations appearing at most twice in the corpus).

In order to evaluate the results, the rare collocations were labeled manually by the first author (2 days).

The results obtained (Tab. 4, Fig. 4) confirm that statistical criteria are not appropriate to deal with rare information (the ROC curves attached to Occ_L and L are below the default curve). Unexpectedly, it appears on this problem that statistical criteria such as MI should better be used in reverse order. It must be emphasized that the relevance of *a priori* criteria strongly depends on the corpus and task at hand. In opposition, the relevance of the ranking hypothesis extracted by *Bagged-ROGER* appears to hold beyond the specificities of the training set, which is confirmed by the lift chart (Fig. 5).

It is no wonder that the best generalization performances are obtained for the non linear hypotheses based on representation (i) (Tab. 5, Fig. 4).

Table 4. Rare Collocations: Performance of ranking hypotheses based on statistical criteria

Statistical Criteria	*AUC*
Occ_L	0.37
$Dice$	0.32
MI^3	0.30
L	0.30
MI	0.29

Table 5. Rare Collocations: Performance of ranking hypotheses based on *Bagged-ROGER* with linear and non-linear hypotheses learned from Frequent Collocations.

Bagged-ROGER		
representation of the domain	*Linear AUC*	*Non-Linear AUC*
(i)	0.67	0.70
(ii)	0.65	0.69
(iii)	0.62	0.69

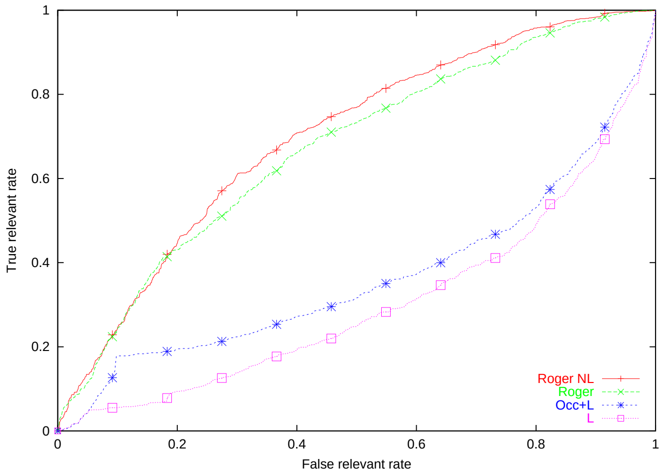

Figure 4. ROC Curves on Rare Collocations (Linear and Non-linear *Bagged-ROGER*, Occ_L and L).

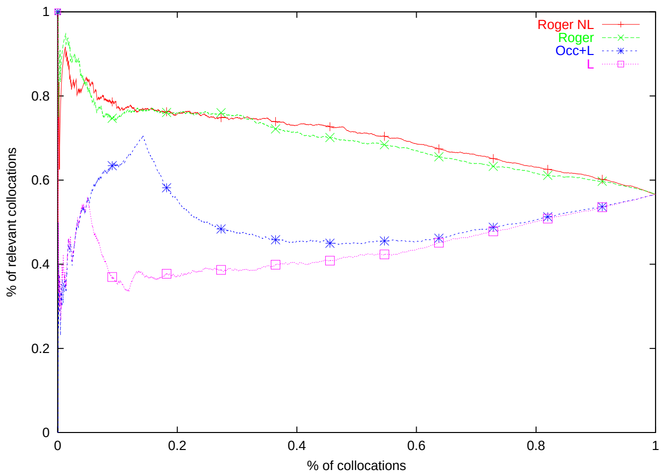

Figure 5. Lift Curves on Rare Collocations (Linear and Non-linear *Bagged-ROGER*, Occ_L and L).

5.2 Molecular Biology

A second application was considered, within the domain of Molecular Biology. A corpus in English, composed of 6,119 abstracts (9,4 Mo), was gathered by querying Medline[3].

Only *Noun-Noun* collocations have been considered in this application; all 1,028 frequent collocations (occurring at least 4 times) have been labelled by a domain expert O. Matte-Tailliez (Table 6).

Table 6. Noun-Noun collocations with the Molecular Biology corpus.

Data	# collocations	relevant	irrelevant
Frequent Collocations	1028	90.9%	9.1%

Bagged-ROGER was applied on this corpus with same experimental setting as for the former corpus (Tab. 7 and 8, Fig. 6 and 7). As observed with the previous application, the best results are obtained for non-linear ranking hypotheses (over .75) against over .65 for the linear case. However *Bagged-ROGER* significantly outperforms the statistical criteria, ranging from .30 to .57. As for the previous corpus, the best statistical criterion is Occ_L and the worst one is MI.

Table 7. Performance of ranking hypotheses based on statistical criteria with the Molecular Biology corpus.

Statistical Criteria	*AUC*
Occ_L	0.57
L	0.42
MI^3	0.35
$Dice$	0.31
MI	0.30

Table 8. Performance of ranking hypotheses based on *Bagged-ROGER* with linear and non-linear hypotheses learned from Molecular Biology corpus and applied on the same corpus.

Bagged-ROGER		
representation of the domain	*Linear AUC*	*Non-Linear AUC*
(i)	0.68	0.76
(ii)	0.63	0.76
(iii)	0.66	0.76

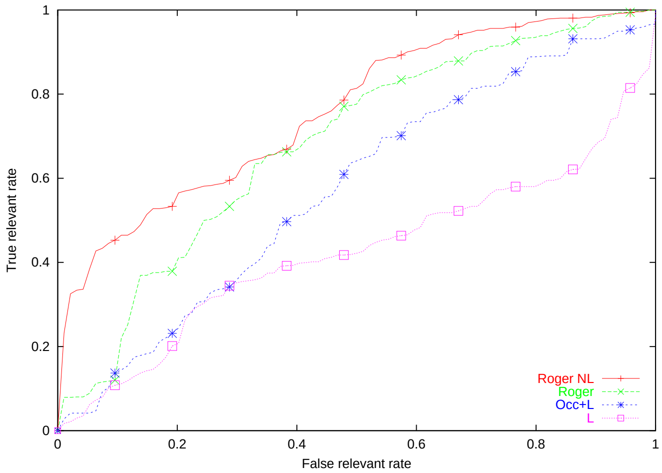

Figure 6. ROC Curves with the Molecular Biology corpus (Linear and Non-Linear *Bagged-ROGER*, Occ_L and L).

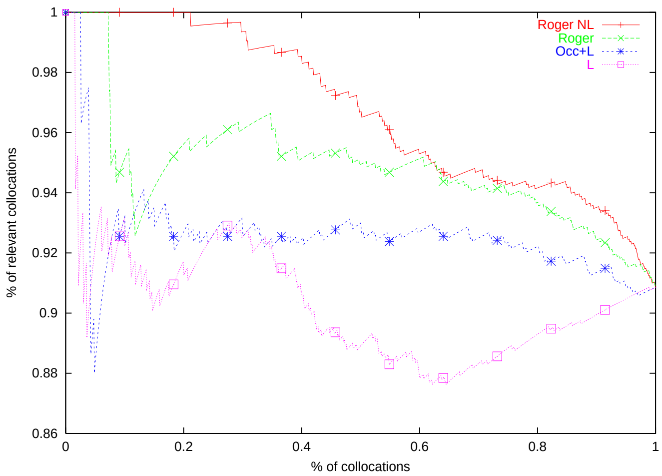

Figure 7. Lift Curves with the Molecular Biology corpus (Linear and Non-Linear *Bagged-ROGER*, Occ_L and L).

5.3 Generality across domains and languages

Finally, we decided to apply the agregated hypothesis constructed by *Bagged-ROGER* on one corpus, to the other corpus. Surprisingly, the results obtained are good; though the interestingness measure learned by *Bagged-ROGER* on the same corpus outperforms that learned on the other corpus, still the latter significantly outperforms the statistical criteria (with confidence .95 using a t-test).

Table 9 shows the AUC criterion measured on the Molecular Biology corpus, of the interestingness measure learned from the CV corpus.

Table 9. Performance of ranking hypotheses based on *Bagged-ROGER* with linear and non-linear hypotheses learned from CV corpus applied to Molecular Biology.

Bagged-ROGER		
representation of the domain	*Linear AUC*	*Non-Linear AUC*
(i)	0.63	0.71
(ii)	0.64	0.69
(iii)	0.64	0.69

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

Table 10. Performance of ranking hypotheses based on *Bagged-ROGER* with linear and non-linear hypotheses learned from the Molecular Biology corpus applied to CV (Frequent Collocations).

Bagged-ROGER		
representation of the domain	*Linear AUC*	*Non-Linear AUC*
(i)	0.64	0.63
(ii)	0.54	0.65
(iii)	0.53	0.65

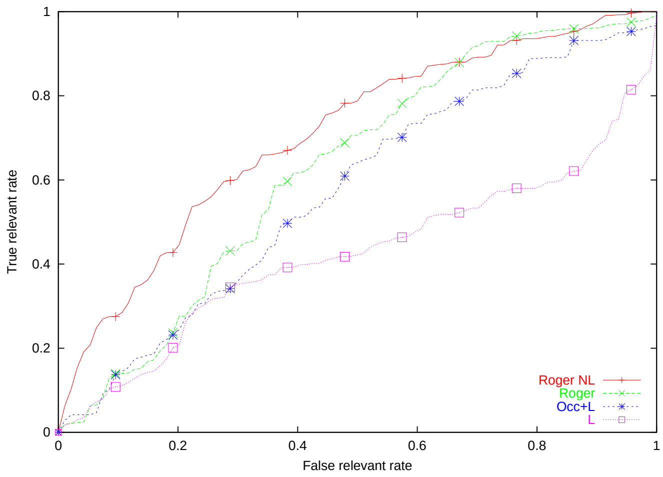

Figure 8. ROC Curves with ranking function learned with the CV applied on the Molecular Biology corpus (Linear and Non-Linear *Bagged-ROGER*, Occ_L and L).

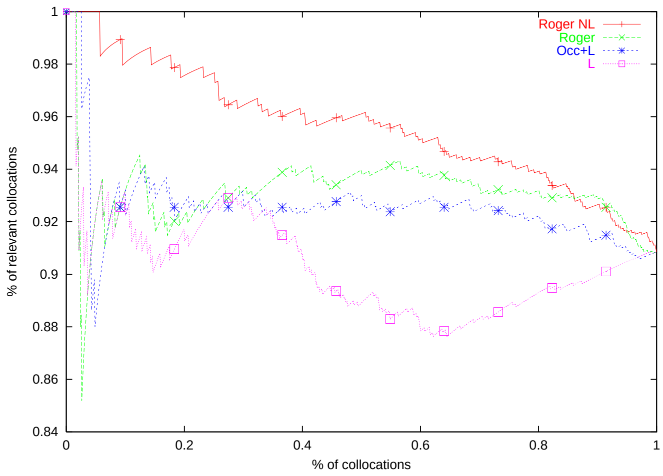

Figure 9. Lift Curves with ranking function learned with the CV applied on the Molecular Biology corpus (Linear and Non-Linear *Bagged-ROGER*, Occ_L and L).

These results show that *Bagged-ROGER* improves in all considered experiments on the standard statistical criteria; in opposition, *ROGER* standalone shows a behavior similar to that of Occ_L and L, in particular in the beginning of the ROC curve (Fig. 8 and 9).

Table 10 symmetrically shows the performance of the interestingness measure learned from the Molecuar Biology corpus, on the CV one.

Similarly, the differences between the interestingness measure learned from the CV corpus and the Medline corpus on the latter corpus are not that much (Tab. 8 and 9, Fig. 10).

Such good robustness was not expected, even more so as both applications regard distinct domains and languages. Ongoing research is examining in more depth the ranking hypotheses obtained and validating them on other corpora.

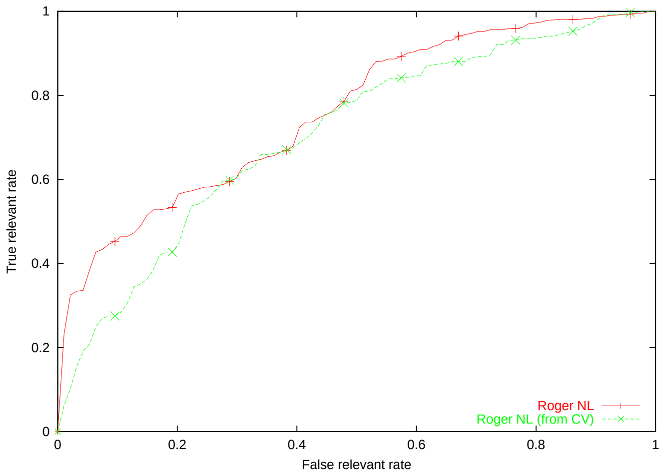

Figure 10. ROC Curves with the Molecular Biology corpus (Non-Linear *Bagged-ROGER*) learned from CV and Biology.

6 CONCLUSION AND PERSPECTIVES

This paper proposes the use of techniques from Supervised Machine Learning for the Term Extraction task in Text Mining. Specifically, from a small set of terms, manually labelled as relevant/irrelevant by the expert, a ranking hypothesis is extracted using the ROC-based evolutionary optimization algorithm *Bagged-ROGER*.

The experimental validation on a real-world domain application demonstrates that the approach significantly improves on the standard statistical and IR-related criteria on the frequent terms. Interestingly, the ranking hypotheses learned from the frequent terms appear to be relevant on the infrequent terms too, which might significantly reduce the expert effort during the intensively time-consuming phase of text preparation.

Further research will investigate in more depth the strength and weaknesses of the presented approach, considering more corpora, various types of target concepts, and involving additional linguistic features.

In parallel, an incremental extension of the approach will be considered, iteratively proposing the expert the top ranked term wrt the current ranking hypothesis, and refining the hypothesis according to the expert answer.

Acknowledgements

We would like to thank Oriane Matte-Tailliez for the expertise of the terms in Molecular Biology.

REFERENCES

[1] A. Amrani, Y. Kodratoff, and O. Matte-Tailliez, 'A semi-automatic system for tagging specialized corpora', in *Proceedings of the Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04), to appear*, (2004).

[2] N. Aussenac-Gilles and D. Bourigault, 'The Th(IC)2 initiative: Corpus-Based Thesaurus Construction for Indexing WWW Documents', in *Proceedings of the EKAW'2000 Workshop on Ontologies and Texts, Vol-51*, (2000).

[3] T. Bäck, *Evolutionary Algorithms in theory and practice*, New-York:Oxford University Press, 1995.

[4] D. Bourigault and C. Jacquemin, 'Term extraction + term clustering: An integrated platform for computer-aided terminology', in *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL '99), Bergen.*, pp. 15–22, (1999).

[5] A.P. Bradley, 'The use of the area under the roc curve in the evaluation of machine learning algorithms', *Pattern Recognition*, **30**(7), 1145–1159, (1997).

[6] J. Brank and J. Leskovec, 'Download Estimation on KDD cup 2003', in *KDD Cup 2003*, eds., Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg, (2003).

[7] L. Breiman, 'Arcing classifiers', *Annals of Statistics*, **26**(3), 801–845, (1998).

[8] E. Brill, 'Some advances in transformation-based part of speech tagging', in *AAAI, Vol. 1*, pp. 722–727, (1994).

[9] K.W. Church and P. Hanks, 'Word association norms, mutual information, and lexicography', *Computational Linguistics*, **16**, 22–29, (1990).

[10] W. Cohen, R. Schapire, and Y. Singer, 'Learning to order things', *Journal of Artificial Intelligence Research*, **10**, 243–270, (1999).

[11] B. Daille, E. Gaussier, and J.-M. Langé, 'An evaluation of statistical scores for word association', in *J.Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Levy, and E. Vallduvi (eds) The Tbilisi Symposium on Logic, Language and Computation: Selected Papers, CSLI Publications*, pp. 177–188, (1998).

[12] P. Domingos, 'Meta-cost: A general method for making classifiers cost sensitive', in *Knowledge Discovery from Databases*, pp. 155–164. Morgan Kaufmann, (1999).

[13] T.E. Dunning, 'Accurate methods for the statistics of surprise and coincidence', *Computational Linguistics*, **19**(1), 61–74, (1993).

[14] R. Esposito and L. Saitta, 'Monte Carlo Theory as an Explanation of Bagging and Boosting', in *Proceeding of the Eighteenth International Joint Conference on Artificial Intelligence*, eds., Georg Gottlob and Toby Walsh, pp. 499–504. Morgan Kaufman Publishers, (2003).

[15] D. Faure and C. Nédellec, 'ASIUM: Learning subcategorization frames and restrictions of selection', in *10th European Conference on Machine Learning (ECML 98) – Workshop on Text Mining*, eds., C. Nédellec and C. Rouveirol, Chemnitz Allemagne, (Avril 1998).

[16] C. Ferri, P. A. Flach, and J. Hernndez-Orallo, 'Learning decision trees using the area under the ROC curve', in *Proceedings of the 19th International Conference on Machine Learning*, ed., Morgan Kaufmann, pp. 179–186, (2002).

[17] D.B. Fogel, E.C. Wasson, and E.M. Boughton, 'Evolving neural networks for detecting breast cancer', *Cancer Letters*, **96**, 49–53, (1995).

[18] M. A. K. Halliday, *System and Function in Language*, Oxford University Press, 1976.

[19] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2000.

[20] C. Jacquemin, 'A symbolic and surgical acquisition of terms through variation', in *Statistical and Symbolic Approaches to Learning for Natural Language Processing*, ed., Springer Verlag, pp. 425–438, (1996).

[21] R. Jin, Y. Liu, L. Si, J. Carbonell, and A. Hauptmann, 'A New Boosting Algorithm Using Input-Dependent Regularizer', in *ICML 2003*, eds., Tom Fawcett and Nina Mishra. AAAI Press, (2003).

[22] G. Nenadić, H. Mima, I. Spasić, S. Ananiadou, and J. Tsujii, 'Terminology-based Literature Mining and Knowledge Acquisition in Biomedicine', *International Journal of Medical Informatics*, **67**, 33–48, (2002).

[23] M. Roche, J. Azé, O. Matte-Tailliez, and Y. Kodratoff, 'Mining texts by association rules discovery in a technical corpus', in *Proceedings of IIPWM'04 (Intelligent Information Processing and Web Mining), Springer Verlag series "Advances in Soft Computing"*, pp. 89–98, (2004).

[24] S. Rosset, 'Model Selection via the AUC', in *Proceedings of the Twenty-First International Conference on Machine Learning (ICML'04), to appear*, (2004).

[25] G. Salton, 'Developments in automatic text retrieval', *Science*, **253**, 974–979, (1991).

[26] R.E. Schapire, 'Theoretical views of boosting', in *Proceedings of EuroCOLT-99, European Conference on Computational Learning Theory*, pp. 1–10, (1999).

[27] H.-P. Schwefel, *Numerical Optimization of Computer Models [translation of dissertation plus extra paragraph on correlated mutations along a research report of 1975]*, 1981.

[28] M. Sebag, J. Azé, and N. Lucas, 'Impact studies and sensitivity analysis in medical data mining with ROC-based genetic learning', in *Proceedings of the Third IEEE International Conference on Data Mining, ICDM 2003*, pp. 637–640, (2003).

[29] M. Sebag, N. Lucas, and J. Azé, 'ROC-based Evolutionary Learning: Application to Medical Data Mining', in *Proceedings of the 6th International Conference on Artificial Evolution, EA 2003*, ed., Springer Verlag, (to appear in 2004).

[30] F. Smadja, 'Retrieving collocations from text: Xtract', *Computational Linguistics*, **19**(1), 143–177, (1993).

[31] F. Smadja, K. R. McKeown, and V. Hatzivassiloglou, 'Translating collocations for bilingual lexicons: A statistical approach', *Computational Linguistics*, **22**(1), 1–38, (1996).

[32] J. Vivaldi, L. Màrquez, and H. Rodríguez, 'Improving term extraction by system combination using boosting', *Lecture Notes in Computer Science*, **2167**, 515–526, (2001).

[33] 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', in *LREC 2002, Third international conference on language resources and evaluation*, (2002).