Pruning terminology extracted from a specialized corpus for CV ontology acquisition

Mathieu Roche[1] and Yves Kodratoff[2]

[1] LIRMM - UMR 5506, Université Montpellier 2, 34392 Montpellier Cedex 5 - France
[2] LRI - UMR 8623, Université Paris-Sud, 91405 Orsay Cedex - France

Abstract. This paper presents an experimental study for extracting a terminology from a corpus made of Curriculum Vitae (CV). This terminology is to be used for ontology acquisition. The choice of the pruning rate of the terminology is crucial relative to the quality of the ontology acquired. In this paper, we investigate this pruning rate by using several evaluation measures (precision, recall, F-measure, and ROC curve).

1 Introduction

This paper presents the experimental study of an evaluation of the best rate of pruning for terminology extraction. Below, we describe our global method for terminology extraction and we define the pruning of a terminology.

The terms extracted from a specialized corpus are instances of the concepts that will become the frame of a domain ontology. In our work, the terms are extracted from a Curriculum Vitae (CV) corpus provided by the company VediorBis[3] (120000 words after various pretreatments described in [13]). This specialized corpus is written in French, it made of very short sentences and many enumerations. For example, in this field, "logiciel de gestion" (management software) is an instance of the concept called "Activité Informatique" (Computer Science Activity). The concept is defined by the expert of the field.

The first step of our terminology extraction approach is based on text normalization by using cleaning rules described in [13]. The next step provides grammatical labels for each word of the corpus, using a Part-Of-Speech tagger Etiq [1]. Etiq is an interactive system based on Brill's tagger [4] which improves the quality of the labeling of specialized corpora. Table 1 presents an example of tagged sentence from CV corpus.

We are then able to extract tagged collocations in CV corpus, such as Noun-Noun, Adjective-Noun, Noun-Adjective, Noun-Preposition-Noun collocations. For example in table 1, we can extract the Noun-Preposition-Noun collocation "logiciel de gestion" (management software).

[3] `http://www.vediorbis.com/`

Développement/SBC:sg d'/PREP un/DTN:sg **logiciel/SBC:sg de/PREP gestion/SBC:sg** du/DTC:sg parc/SBC:sg informatique/ADJ:sg ...
SBC:sg → *Noun*, *singular* PREP → *Preposition* DTC:sg, DTN :sg → *Determiners*, *singular* ADJ:sg → *Adjective*, *singular*

Table 1. Part-Of-Speech tagged corpus (in French).

The next step consists in selecting collocations more relevant according to the statistical measurements described in [13, 14]. Collocations are groups of words defined in [11, 17]. We call *terms*, collocations relevant to the field of interest.

The binary terms (or ternary for the prepositional terms) extracted at each iteration are reintroduced in the corpus with hyphens. So, they are recognized like words. We can thus carry out a new terminology extraction from the corpus taking into account of terminology acquired at the preceding steps. Our iterative method which has similarities with [8] is described in [13, 15]. This approach enables to detect very specific terms (made of several words). For example using the term "logiciel de gestion" extracted at the first iteration of our approach, after several iterations we can extract the specific term "logiciel de gestion du parc informatique" (see table 1).

The choice of the pruning rate consists in determining the minimal number of times where relevant collocations are found in the corpus.

First, this paper presents briefly the state-of-the-art of terminology extraction methods (section 2). The presentation of the application of various pruning rates is described in section 3. After the presentation of the collocations expertise in section 4, the section 5 describes various evaluation measurements of the terminology based on the problems of the choice of the pruning rate. Finally, in section 6 we discuss future work.

2 The state-of-the-art of terminology extraction approaches

In order to extract and structure the terminology, several methods are developed. Here, we will not deal with the approaches of conceptual regrouping of terminology as they are described in [16, 2].

The methods of terminology extraction are based on statistical or syntactic approaches. The TERMINO system [6] is a precursory tool that uses a syntactic analysis in order to extract the nominal terms. This tool carries out a morphological analysis containing rules, followed by an analysis of nominal collocations

using a grammar. The XTRACT system [17] is based on a statistical method. Initially XTRACT extracts binary collocations in a window of ten words which exceed a statistical significant rate. The following step consists in extracting more particular collocations (collocations of more than two words) containing the binary collocations extracted at the preceding step. ACABIT [5] carries out a linguistic analysis in order to transform nominal collocations into binary terms. They are ranked using statistical measurements. Contrary to ACABIT which is based on a statistical method, LEXTER [3] and SYNTEX [9] use syntactic analysis. This method extracts the longuest noun phrases. These phrases are transformed into "head" and "expansion" terms using grammatical rules. The terms are structured using syntactic criteria.

To discuss the choice of the pruning rate, we will rank collocations by using Occ_L measurement as described in [13]. This measurement which gives the best results [14] ranks collocations according to their number of occurrences (Occ). Collocations having the same number of occurrences are ranked by using the loglikelyhood (L) [7]. Thus, Occ_L is well adapted to discuss the choice of the pruning rate.

3 Pruning rate of the terminology

The principle of pruning the collocations consists in analyzing the collocations usefulness for ontology acquisition : their number has to be above a threshold of occurrences in the corpus. We can thus remove rare collocations which can appear as irrelevant for the field. Table 2 presents the various prunings we applied (first iteration of our terminology extraction approach). Table 2 shows that the elimination of collocations with one occurrence in the CV corpus allows us to remove more than 75% of the existng collocations.

	nb	**pruning 2**	**pruning 3**	**pruning 4**	**pruning 5**	**pruning 6**
Noun-Noun	1781	353 (80%)	162 (91%)	100 (94%)	69 (96%)	56 (97%)
Noun-Prep-Noun	3634	662 (82%)	307 (91%)	178 (95%)	113 (97%)	80 (98%)
Adjective-Noun	1291	259 (80%)	103 (92%)	63 (95%)	44 (97%)	34 (97%)
Noun-Adjective	3455	864 (75%)	448 (87%)	307 (91%)	222 (94%)	181 (95%)

Table 2. Pruning and proportions of pruning.

4 Terminology acquisition for conceptual classification

To build a conceptual classification, collocations evoking a concept of the field are extracted. Table 3 presents examples of French collocations associated to concepts met in the CV corpus.

Collocations	Concepts
aide comptable	Activité Gestion
gestion administrative	Activité Gestion
employé libre service	Activité Commerce
assistant marketing	Activité Commerce
chef de service	Activité Encadrement
direction générale	Activité Encadrement
BEP secrétariat	Compétence Secrétariat
BTS assistante de direction	Compétence Secrétariat
baccalauréat professionnel vente	Compétence Commerce
BTS commerce international	Compétence Commerce

Table 3. Part of conceptual classification from CV corpus (in French).

In order to validate the collocations, several categories of relevance (or irrelevance) are possible:

- **Category 1:** Collocation is relevant for conceptual classification.
- **Category 2:** Collocation is relevant but very specific and not necessarily relevant for a domain conceptual classification.
- **Category 3:** Collocation is relevant but very general and not necessarily relevant for a conceptual classification specialized.
- **Category 4:** Collocation is irrelevant.
- **Category 5:** The expert cannot judge if collocation is relevant or not.

5 Evaluation of the terminology and pruning rate

An expert evaluates Noun-Adjective collocations extracted in CV corpus using all rate pruning.

5.1 Terminology expertise

Table 4 gives the number of Noun-Adjective collocations associated with each category of expertise. Each category is described in the section 4 of this paper.

Table 4 shows the results of the expertise carried out according to various pruning rates. The most relevant collocations (category 1) are privileged by applying an large pruning rate. If all collocations are provided by the system (i.e. pruning at one), the proportion of relevant collocations is 56.3% and more than 80% with a pruning at four, five or six.

5.2 Precision, recall, and F-measure

Precision is an evaluation criterion adapted to the framework of an unsupervised approach. Precision calculates the proportion of relevant collocations extracted among extracted collocations. Using the notations of table 5, the precision is given by the formula $\frac{TP}{TP+FP}$. A 100% precision means that all the collocations extracted by the system are relevant.

pruning	**category 1**	**category 2 and 3**	**category 4**	**category 5**	Total
1	1946 (56.3%)	919 (26.6%)	395 (11.4%)	195 (5.6%)	3455
2	631 (73.0%)	151 (17.5%)	58 (6.7%)	24 (2.8%)	864
3	348 (77.7%)	73 (16.3%)	17 (3.8%)	10 (2.2%)	448
4	256 (83.4%)	36 (11.7%)	8 (2.6%)	7 (2.3%)	307
5	185 (83.3%)	29 (13.1%)	3 (1.3%)	5 (2.2%)	222
6	152 (84.0%)	23 (12.7%)	2 (1.1%)	4 (2.2%)	181

Table 4. Number of collocations in each category.

Another typical measurement of the machine learning approach is recall which computes the proportion of relevant collocations extracted among relevant collocations. The recall is given by the formula $\frac{TP}{TP+FN}$. A 100% recall means that all relevant collocations have been found. This measurement is adapted to the supervised machine learning methods where all positive examples (relevant collocations) are known.

	Relevant collocations	Irrelevant collocations
Collocations evaluated as relevant by the system	TP (True Positive)	FP (False Positive)
Collocations evaluated as irrelevant by the system	FN (False Negative)	TN (True Negative)

Table 5. Contingency table at the base of evaluation measurements.

It is often important to determine a compromise between recall and precision. We can use a measurement taking into account these two evaluation criteria by calculating the F-measure [19] :

$$F - measure(\beta) = \frac{(\beta^2 + 1) \times Precision \times Recall}{(\beta^2 \times Precision) + Recall} \quad (1)$$

The parameter β of the formula (1) regulates the respective influence of precision and recall. It is often fixed at 1 to give the same weight to these two evaluation measurements.

The table 6 shows a large pruning and gives the highest precision. In this case, the recall is often small, i.e. few relevant collocations extracted. With $\beta = 1$, we can see in table 6 that the F-measure is highest without applying pruning. This is due to the high result of the recall without pruning. Indeed, as specified in table 2, a pruning at two prevents the extraction of 75% of Noun-Adjective collocations.

Table 7 shows varying β in order to give a more important weight to the precision ($\beta < 1$) gives a F-measure logically higher in the case of a large pruning. This underlines the limits of this evaluation criterion because the results of the F-measure can largely differ according to the value of β. Thus, the following section presents another evaluation criterion based on ROC curves.

Pruning	**Precision**	**Recall**	**F-measure**
1	59.7%	100%	74.8%
2	75.1%	32.4%	45.3%
3	79.4%	17.9%	29.2%
4	85.3%	13.1%	22.8%
5	85.2%	9.5%	17.1%
6	85.9%	7.8%	14.3%

Table 6. Precision, recall, and F-measure with $\beta = 1$.

β	1	1/2	1/3	1/4	1/5	1/6	1/7	1/8	1/9	1/10
1	74.8%	64.9%	62.2%	61.1%	60.6%	60.4%	60.2%	60.1%	60.0%	59.9%
2	45.3%	59.5%	66.4%	69.7%	71.5%	72.5%	73.2%	73.6%	73.9%	74.1%
3	29.2%	47.0%	59.1%	66.1%	70.2%	72.7%	74.3%	75.4%	76.2%	76.8%
4	22.8%	40.7%	55.1%	64.5%	70.5%	74.3%	76.9%	78.7%	80.0%	80.9%
5	17.1%	32.9%	47.4%	58.0%	65.2%	70.1%	73.5%	75.9%	77.7%	79.0%
6	14.3%	28.6%	42.9%	54.1%	62.0%	67.6%	71.6%	74.4%	76.5%	78.1%

Table 7. F-measure according to various values of β (1, ..., 1/10) and various rates of pruning (1, ..., 6).

5.3 The ROC curves

In this section ROC curves (Receiver Operating Characteristics) are presented (see also work of [10]). Initially the ROC curves come from the field of signal treatment. ROC curves are often used in the field of medicine to evaluate the validity of diagnostic tests. The ROC curves show in X-coordinate the rate of false positive (in our case, rate of irrelevant collocations) and in Y-coordinate the rate of true positive (rate of relevant collocations). The surface under the ROC curve (*AUC - Area Under the Curve*), can be seen as the effectiveness of a measurement of interest. The criterion relating to the surface under the curve is equivalent to the statistical test of Wilcoxon-Mann-Whitney (see work of [20]).

In the case of the collocations ranking in using statistical measurements, an perfect ROC curve corresponds to obtaining all relevant collocations at the

beginning of the list and all irrelevant collocations at the end of the list. This situation corresponds to $AUC = 1$. The diagonal corresponds to the performance of a random system, progress of the rate of true positive being accompanied by an equivalent degradation of the rate of false positive. This situation corresponds to $AUC = 0.5$. If the collocations are ranked by decreasing interest (i.e. all relevant collocations are after the irrelevant ones) then $AUC = 0$. An effective measurement of interest to order collocations consists in obtain a AUC the highest possible value. This is strictly equivalent to minimizing the sum of the rank of the positive examples.

The advantage of the ROC curves comes from its resistance to imbalance (for example, an imbalance in number of positive and negative examples). We can illustrate this fact with the following example. Let us suppose that we have 100 examples (collocations). In the first case, we have an imbalance between the positive and negative examples with only 1 positive and 99 negative examples. In the second case, we have 50 positive and 50 negative examples. Let us suppose that for these two cases, the positive examples (relevant collocations) are presented at the top of the list ranked with statistical measurements.

In both cases, the ROC curves are strictly similar with $AUC = 1$ (see figures 1(a) and 1(c)). Thus, getting relevant collocations in the top of the list is emphasized by evaluating the ROC curves and the AUC. With calculation of F-measure (with $\beta = 1$), we obtain two extremely different curves (see figures 1(b) and 1(d)). Thus, imbalances between positive and negative examples strongly influence F-measure contrary to the ROC curves.

From one pruning to another, the rate of relevant and irrelevant collocations can appear extremely different. It means that we are in presence of an imbalance between the classes. For example, applying a pruning at six, 84% of collocations are relevant against 56% whithout pruning (see table 4). The table 8 calculates the various AUC by choosing various pruning rates. Then, in this case, using ROC curves and AUC is particularly well adapted.

Pruning	**AUC**	**Pruning**	**AUC**
1	0.4538	4	0.5012
2	0.5324	5	0.5432
3	**0.5905**	6	0.5447

Table 8. AUC with several prunings.

Figure 2 shows example of AUC and ROC curve with several prunings. Table 8 shows that pruning is better adapted since AUC corresponds to a pruning at three for Noun-Adjective collocations of the CV corpus. Figure 3 shows the

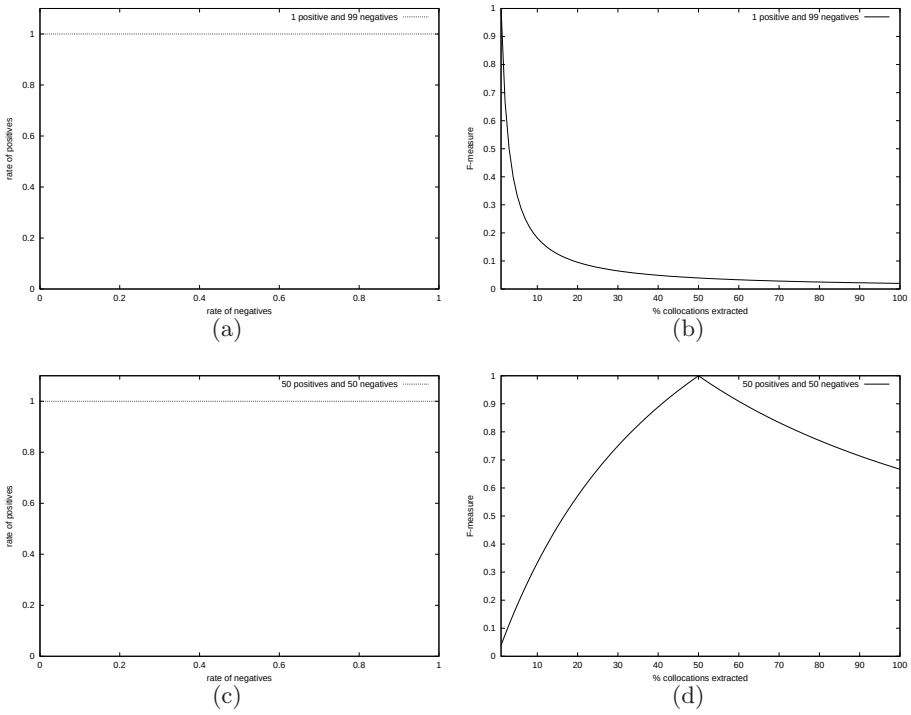

Fig. 1. ROC Curve (a) and F-measure (b) with 1 positive example placed at the top of the list and 99 negative examples placed at the end of the list. ROC Curve (c) and F-measure (d) with 50 positive examples at the top of the list and 50 negative examples at the end of the list. For the calculation of the F-measure, $\beta = 1$.

	Pruning	*Relevance*
Collocation 1	**4**	**-**
Collocation 2	**2**	**+**
Collocation 3	**2**	**-**
Collocation 4	**1**	**-**
Collocation 5	**1**	**+**

pruning	AUC	ROC curve
3	0	
2	0,5	
1	0,33	

Fig. 2. Example of several prunings.

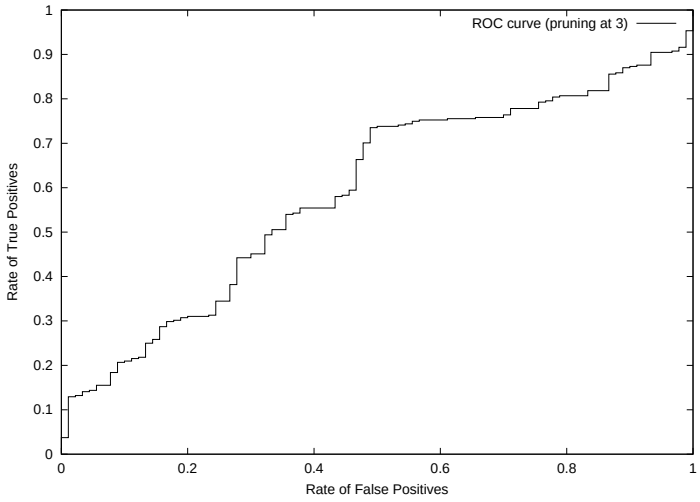

Fig. 3. ROC curve with pruning at 3.

ROC curve related to a pruning at three. This objective criterion based on AUC corresponds to the empirical choice of pruning at three applied in work of [12, 18].

6 Conclusion and perspectives

The experimental study conducted in this paper enables the discussion of the choice of the pruning rate for terminology extraction in view of ontology acquisition. Various criteria of evaluation exist such as precision, recall, and F-measure which takes into account these two criteria. A defect of the F-measure is the choice not always obvious of a parameter best adapted to privilege precision or recall in the calculation. Thus, in this paper, we propose to use ROC curves and AUC to evaluate the choice of pruning. This criterion is not sensitive to imbalance between the classes (such as classes of relevant and irrelevant collocations).

In a future work, we will improve quality of normalization and we will add new CV to increase the number of collocations extracted. Our experiments on the CV corpus show that a pruning at three seems well adapted. In our future work, we propose to compare this result with the one for other specialized corpora. So, we will carry out a complete expertise of collocations of other fields. This will require non negligible expert work.

References

1. A. Amrani, Y. Kodratoff, and O. Matte-Tailliez. A semi-automatic system for tagging specialized corpora. In *Proceedings of PAKDD'04*, pages 670–681, 2004.

2. N. Aussenac-Gilles and D. Bourigault. Construction d'ontologies à partir de textes. In *Actes de TALN03*, volume 2, pages 27–47, 2003.
3. D. Bourigault and C. Jacquemin. Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Proceedings of EACL'99, Bergen.*, pages 15–22, 1999.
4. E. Brill. Some advances in transformation-based part of speech tagging. In *AAAI, Vol. 1*, pages 722–727, 1994.
5. B. Daille. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *P. Resnik and J. Klavans (eds). The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press*, pages 49–66, 1996.
6. S. David and P. Plante. De la nécessité d'une approche morpho syntaxique dans l'analyse de textes. In *Intelligence Artificielle et Sciences Cognitives au Quebec*, volume 3, pages 140–154, 1990.
7. Ted E. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
8. D.A. Evans and C. Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of ACL*, pages 17–24, Santa Cruz, US, 1996.
9. C. Fabre and D. Bourigault. Linguistic clues for corpus-based acquisition of lexical dependencies. In *Corpus Linguistics, Lancaster*, pages 176–184, 2001.
10. C. Ferri, P. Flach, and J. Hernandez-Orallo. Learning decision trees using the area under the ROC curve. In *Proceedings of ICML'02*, pages 139–146, 2002.
11. M. A. K. Halliday. *System and Function in Language*. Oxford University Press, London, 1976.
12. C. Jacquemin. *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. PhD thesis, Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes, 1997.
13. M. Roche. *Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes*. PhD thesis, Université de Paris 11, Décembre 2004.
14. M. Roche, J. Azé, Y. Kodratoff, and M. Sebag. Learning interestingness measures in terminology extraction. A ROC-based approach. In *Proceedings of "ROC Analysis in AI" Workshop (ECAI 2004), Valencia, Spain*, pages 81–88, 2004.
15. M. Roche, T. Heitz, O. Matte-Tailliez, and Y. Kodratoff. Exit: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In *Proceedings of JADT'04*, volume 2, pages 946–956, 2004.
16. M. Shamsfard and A. A. Barforoush. The state of the art in ontology learning: a framework for comparison. *The Knowledge Engineering Review, Volume 18 , Issue 4 (December 2003)*, pages 293–316, 2003.
17. F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.
18. A. Thanopoulos, N. Fakotakis, and G. Kokkianakis. Comparative Evaluation of Collocation Extraction Metrics. In *Proceedings of LREC'02*, volume 2, pages 620–625, 2002.
19. C.J. Van-Risbergen. *Information Retrieval*. 2nd edition, London, Butterworths, 1979.
20. L. Yan, R.H. Dodier, M. Mozer, and R.H. Wolniewicz. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In *Proceedings of ICML'03*, pages 848–855, 2003.