Text and Web Mining Approaches in Order to Build Specialized Ontologies

Mathieu Roche [a,*], Yves Kodratoff [b]

[a] *LIRMM, CNRS, Univ. Montpellier 2, 34392 Montpellier Cedex 5 – France*
[b] *LRI, CNRS, Univ. Paris-Sud, 91405 Orsay Cedex – France*

Abstract

This paper presents a text-mining approach in order to extract candidate terms from a corpus. The relevant candidates are selected using a web-mining approach. The terms (*i.e.* relevant candidate terms) we find are the instances of specialized ontologies built during this process. The experiments are based on real data – Human Resources corpus – and they show the quality of our text and web mining approaches.

Key words: Text-Mining, Web-Mining, Terminology, Statistical Measures, Conceptual Classification

1. Introduction

Our approach extracts knowledge directly from the initial corpus. It is made up of two main phases (1) collecting a homogeneous corpus on a given topic, (2) building a categorization specific to the corpus. The first phase will not be described in this paper, since it is not performed in a completely automated way. It is important however not to forget that this phase is the basis upon which the whole text-mining process is built, since its success depends heavily on the quality and the homogeneity of the collected corpora. The various treatments described in this paper have been carried out on a French corpus on the topics of Human Resources (3,784 KB). This corpus is provided by the PerfomanSe company. [1]

The second phase of the text-mining process is based on the identification of terms in the texts. The search for terms finds the significant word groupings for the specialty field. The terms obtained are associated to concepts, i.e., those showing the same semantics are clustered. Each cluster represents a concept, meaningful to the field expert, and these basic concepts can be, and usually are, the first level of an ontology of concepts. The existence of more general concepts is asserted by the expert, when needed. For instance, our expert recognized from the Human Resources corpus that the set of French terms *véritable solidarité* (*genuine solidarity*), *travail en groupe* (*group work*), *réseau de relations* (*network of relationships*) were the linguistic observables indicating the existence of the concept *Relational* in the texts. These indications of presence of a concept are defined as "instances" of this concept. We thus obtain an ontology, the nodes of which are concepts.

In order to build the specialized ontology, we apply a text and web mining process to extract the

* Corresponding author.
Email addresses: mroche@lirmm.fr (Mathieu Roche), yk@lri.fr (Yves Kodratoff).

[1] http://www.performanse.fr/

instances of the concepts. The text-mining method extracts candidate terms (section 2) and the web-mining approach validates the extracted candidates (section 3). The results of the experiments are given in section 4. Finally the section 5 presents future work.

2. A text-mining approach in order to extract terminology

2.1. *Cleaning and PoS tagging*

In text-mining approaches all the corpora we worked upon had in common to be largely unsuitable to further linguistic treatment due to the variety of unexpected forms they contain. For instance in the Human Resouces corpus, the vocabulary has been normalized. The main point to signal is that the writing of cleaning procedures cannot be done without the help of field expert who provides sets of rules specific to the domain. For instance, in scientific domains (e.g. Biomedical domain), the complex combinations of upper and lower case letters convey a meaning the expert only can deal with.

The PoS (Part of Speech) tags each word of the cleaned texts with a grammatical label. During this step, we used Brill's tagger [2]. After tagging, we are able to extract the doublets or triplets of single words, showing a specific grammatical label (*Noun*, *Adjective*, *Preposition*, etc).

2.2. *Extracting terms*

This step automatically extracts the terms from the texts (*i.e.* relevant candidate terms). In our work we have used the Exit system to extract terminology [2] [10]. This system enables us to extract the candidate terms *Noun-Noun*, *Noun-Adjective*, *Adjective-Noun*, *Noun-Preposition-Noun* from a corpus. The next step is to select the most appropriate candidates according to a statistical measure [5,9]. The binary terms (or ternary for prepositional terms) extracted at each iteration are reintroduced into the corpus with hyphens to be recognized as words. We can start a new terminology extraction from the corpus taking into account the terminology found at the previous steps.

Before choosing the most suitable measure according the extraction of relevant terms, we perform some preliminary cleaning, driven by the expert, of the candidate terms list. We exclude the terms containing words that the expert classes as non significant. For example, the candidates containing the adjectives "other", "same", "such".

2.3. *Limits of the text-mining approach to select the relevant terms*

Statistical measures are often adapted to select (*i.e* , to rank) the relevant terms from large corpora. For a small corpus, these statistical measures are often inadequate. In small corpora, most of the terms are present only once, then the statistics do not discriminate the terminology. To select the relevant terms, we therefore propose to use an approach that does not rely on the frequency of the terms in the corpora. Our web-mining approach is described in the following section.

3. A web-mining approach in order to select relevant terms

3.1. *Web measure*

Our web-mining method uses statistics and information from the web in order to rank the candidates. Regarding the use of ranking functions, we are close to Daille's approach [4,5]. Actually, as this one, our technique calculates the dependency between the words composing the candidate terms in order to rank them. The statistics used by B. Daille are based on the frequency of candidates in the corpus. As described in section 2.3 these statistical measures have limits. Then we propose to use statistics based on the frequency of candidates on the Web as [15] to describe the "Web popularity" of the terms. In that sense, our web-mining approach is close to Turney's approach [15].

The algorithm PMI-IR (Pointwise Mutual Information and Information Retrieval) described in [15] queries the Web via the AltaVista search engine to determine appropriate synonyms to a given query. For a given word, noted *word*, PMI-IR chooses a synonym among a given list. These selected terms, noted $choice_i$, $i \in [1, n]$, correspond to TOEFL [3] questions. The aim is to compute the $choice_i$ synonym that gives the better score.

[2] http://www.lri.fr/∼heitz/formulaire_logiciels.html

[3] Test of English as a Foreign Language

To obtain scores, PMI-IR uses several measures based on the proportion of documents where both terms are present. Turney's formula [15] is inspired by Mutual Information described in section 3.2.2. The measure calculates the proportion of documents containing both *word* and $choice_i$ (within a 10 words window), and compares with the number of documents containing the word $choice_i$. The higher this proportion is, the more *word* and $choice_i$ are seen as synonyms. This kind of web technique will be used in our work. But as argumented in the next section, we use several quality measures to order the candidates.

3.2. *Statistical Measures*

Several quality measures in the literature are based on ranking functions. They are brought out of various fields: Association rules detection [1,8], terminology extraction [5,9], and so forth. The following are the most widely used.

3.2.1. *Number of Occurrences*

The basic measure FR used is based on the number of occurrences $nb(x, y)$. This function corresponds to the number of web pages provided by the query *"x y"* with a search engine (in our work, we use the Exalead [4] search engine). Actually, the value returned by this function corresponds to the popularity of the use of the two words x and y together like a string. Note that the Turney's approach calculates the dependency of the words using the AND and NEAR operators of the search engine.

3.2.2. *Mutual Information*

One of the most commonly used measures to compute a kind of relationship between the words composing what is called a co-occurrence is Church's Mutual Information (MI). The formula is the following [3]:

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \quad (1)$$

Such a measure tends to extract rare and specific co-occurrences according to [5,9,14]. Let us notice that in this formula (1), the use of the log_2 function is not mandatory, since the latter is strictly growing. Thus, the order of the co-occurrences provided by the measure is not impacted by the application of log_2 function. In our context, $P(x, y)$ measures the probability of finding couples of words (x,y) where x et y are neighbors, and in this order. When simplified, the formula (1) could be written as follows, where nb designates the number of occurrences of words and couples of words:

$$MI(x, y) = \frac{nb(x, y)}{nb(x)nb(y)} \quad (2)$$

3.2.3. *Cubic Mutual Information*

The Cubic Mutual Information is an empirical measure based on MI, which enhances the impact of frequent co-occurrences, something which is absent in the original MI [4]. Such as measure is defined by the following formula:

$$MI^3(x, y) = \frac{nb(x, y)^3}{nb(x)nb(y)} \quad (3)$$

Vivaldi et *et al.* have estimated that the Cubic MI was the best behaving measure [16]. This measure gave good results in our work regarding the desambiguisation of the acronym definitions [12].

3.2.4. *Dice's Coefficient*

An interesting quality measure is Dice's coefficient [13]. It is defined by the following formula:

$$D(x, y) = \frac{2 \times P(x, y)}{P(x) + P(y)} \quad (4)$$

Similarly to the Cubic MI, Dice's coefficient weakens the impact of rare and often irrelevant co-occurrences [11]. Formula (5) leads directly to formula (4) [5], which relies on the nb occurrences of words and couple of word occurrences:

$$Dice(x, y) = \frac{2 \times nb(x, y)}{nb(x) + nb(y)} \quad (5)$$

3.3. *Using a context*

In this paper, we define the *context* as a set of significant words given by the expert to define the field. These words are added to the queries using the *AND* operator in the statistical measures. For instance, without the context, the $nb(x, y)$ function calculates the number of pages provided by the search engine by the query *"x y"*. With the context based on the

[4] http://www.exalead.fr/search/

[5] by writing $P(x) = \frac{nb(x)}{nb\_total}$, $P(y) = \frac{nb(y)}{nb\_total}$, $P(x, y) = \frac{nb(x,y)}{nb\_total}$

words c_1, ..., c_n, we apply the query *"x y" AND* c_1 *AND ... AND* c_n. The goal of this context is to restrict the searching space to the Web pages of the domain.

4. Experiments

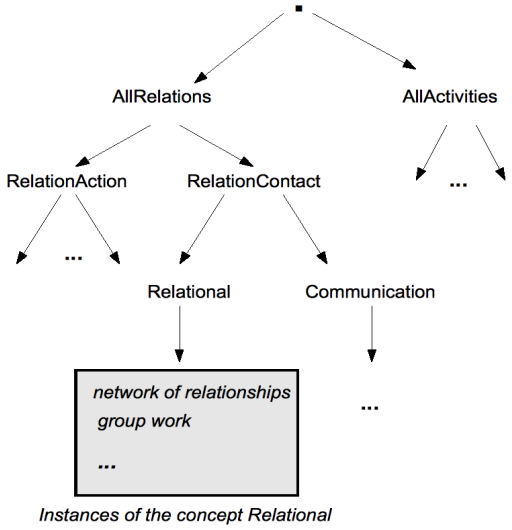
4.1. *An ontology used as a benchmark*

We have extracted terminology regarding the specialized Human Resources corpus. We can select the most frequent terminology: Candidates that appear only one or two times are not taken into account, i.e., they are pruned [11]. The Table 1 shows the results of extracted candidates in our corpus.

Patterns	number of occurrences	number of occurrences after pruning at 3	rate of pruning
Noun-Prep-Noun	4,703	1,268	73.0%
Noun-Noun	98	11	88.8%
Adjective-Noun	1,290	476	63.1%
Noun-Adjectve	5,768	1,628	71.8%

Table 1
Results of occurrence number of extracted candidate terms.

The expert manually analyzes the candidates extracted by the EXIT system in order to build a specialized ontology. Then our Human Resources ontology is composed of 3,000 terms (instances of the concepts of the first level of the built ontology). Our ontology has 3 levels: 12 concepts for the first level, [6] 7 concepts for the level 2, and 2 concepts of the level 3. The following figure shows a representation of our ontology:



The terms (i.e. instances of the concepts) manually validated will be used as benchmark in order to evaluate our web-mining approaches.

4.2. *Results of the web-mining approaches*

We have studied the 500 most frequently extracted candidates in our Human Resources corpus. 73% of these candidates were evaluated as relevant by an expert. Note that the frequent candidates in a corpus are often relevant [11].

We have evaluated the rate of relevant candidates (precision) after applying statistical measures presented in section 3.2. In Table 2, the quality measures are evaluated according to different thresholds (*i.e.* precision calculated with the n first candidates given by statistical measures). This Table shows that all the statistical measures give better results than a random classification (*e.g.*, the precision is increased by 9% up to 15% for the first 100 words).

Table 3 shows the sum of the ranks of relevant candidates using the different measures. When the value is low, the results is better. Note that minimizing the sum of the ranks of relevant candidates is

[6] French concepts: environnement, relationnel, Vous-Même, communication, stress, rôle, indépendance, influence, hiérarchie, Comportement&Attitude, activité, activité_gestion&administration

equivalent to maximizing the Area Under the ROC [7] curves (AUC - Area Under the Curve). This criterion of AUC [6] is often used to evaluate learning algorithms [7] or ranking functions [9]. We have calculated the sum of the ranks with and without the use of a context, the French word "psychologie" (*i.e.* "psychology") in our queries. This Table enables us to observe two important results:

– The use of the context improves the result for all measurements.
– Dice's measure gives better results with and without context.

Threshold n	Random	FR measure	MI measure	MI^3 measure	Dice's measure
50	0.73	0.86	0.88	**0.92**	0.88
100	0.73	0.82	**0.88**	0.83	0.85
150	0.73	0.77	0.80	0.79	**0.81**
200	0.73	0.78	0.78	**0.80**	**0.80**
250	0.73	0.77	**0.78**	**0.78**	**0.78**
300	0.73	0.77	**0.79**	0.78	0.78
350	0.73	**0.78**	0.77	**0.78**	0.77
400	0.73	**0.77**	0.76	0.76	**0.77**
450	0.73	**0.75**	0.74	0.74	0.74
500	**0.73**	**0.73**	**0.73**	**0.73**	**0.73**

Table 2
Precision of web measures taking into account the French context "psychologie".

Measures	FR	MI	MI^3	Dice
without context	87,584	88,789	87,502	**87,066**
with context	87,146	87,105	86,815	**86,597**

Table 3
Sum of the ranks of the relevant candidates.

5. Conclusion

The approach presented here extracts a terminology using a text-mining approach. The second step of our process selects relevant candidates using a web mining technique based on statistical measures. The results show that the association of statistical measures with a context improves the results. Furthermore, Dice's measure seems well suited (*e.g.*, 80% of candidate terms are relevant based on the first 200 candidates returned by this measure). These relevant candidate terms represent the instances of the built ontology. In our future work, we wish to take into account other statistical measures. We will develop a web-mining approach to automatically determine the context. This richer context could improve the results obtained with statistical measures.

[7] Receiver Operating Characteristics

Acknowledgments

We thank Serge Baquedano (PerformanSe company) for the acquisition and the expertise of the Human Resources corpus.

References

[1] J. Azé, Y. Kodratoff, A study of the effect of noisy data in rule extraction systems, in: Proceedings of EMCSR'02, vol. 2, 2002.

[2] E. Brill, Some advances in transformation-based part of speech tagging, in: AAAI, Vol. 1, 1994.

[3] K. Church, P. Hanks, Word association norms, mutual information, and lexicography, in: Computational Linguistics, vol. 16, 1990.

[4] B. Daille, Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques, Ph.D. thesis, Universié Paris 7 (1994).

[5] B. Daille, Study and Implementation of Combined Techniques for Automatic Extraction of Terminology, in: The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press, 1996.

[6] C. Ferri, P. Flach, J. Hernandez-Orallo, Learning decision trees using the area under the ROC curve, in: Proceedings of 9th International Conference on Machine Learning, ICML'02, 2002.

[7] J. Huang, C. X. Ling, Using auc and accuracy in evaluating learning algorithms, IEEE Transactions on Knowledge and Data Engineering 17 (3) (2005) 299–310.

[8] S. Lallich, O. Teytaud, Evaluation et validation des règles d'association, Numéro spécial "Mesures de qualité pour la fouille des données", Revue des Nouvelles Technologies de l'Information (RNTI) RNTI-E-1 (2004) 193–218.

[9] M. Roche, J. Azé, Y. Kodratoff, M. Sebag, Learning interestingness measures in terminology extraction. a roc-based approach, in: Proceedings of "ROC Analysis in AI" Workshop (ECAI 2004), 2004.

[10] M. Roche, T. Heitz, O. Matte-Tailliez, Y. Kodratoff, Exit: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés, in: JADT'04 (Journées internationales d'Analyse statistique des Données Textuelles), vol. 2, 2004.

[11] M. Roche, Y. Kodratoff, Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition, in: Proceedings of onToContent Workshop - OTM'06, Springer Verlag, LNCS, 2006.

[12] M. Roche, V. Prince, Managing the Acronym/Expansion Identification Process for Text-Mining Applications, International Journal of Software and Informatics 2 (2) (2008) 163–179.
[13] F. Smadja, K. R. McKeown, V. Hatzivassiloglou, Translating collocations for bilingual lexicons: A statistical approach, Computational Linguistics 22 (1) (1996) 1–38.
[14] A. Thanopoulos, N. Fakotakis, G. Kokkianakis, Comparative Evaluation of Collocation Extraction Metrics, in: Proceedings of LREC'02, 2002.
[15] P. Turney, Mining the Web for synonyms: PMI–IR versus LSA on TOEFL, in: Proceedings of the 12th European Conference on Machine Learning (ECML), 2001.
[16] J. Vivaldi, L. Màrquez, H. Rodríguez, Improving term extraction by system combination using boosting, in: Proceedings of the 12th European Conference on Machine Learning (ECML), 2001.