Strategies in Identifying Issues Addressed in Legal Reports

Gilbert Ritschard[1], Matthias Studer[1], and Vincent Pisetta[2]

[1] Department of Econometrics, University of Geneva, Switzerland,
{gilbert.ritschard,matthias.studer}@metri.unige.ch
[2] ERIC Laboratory, University of Lyon 2, France, *vincent.pisetta@univ-lyon2.fr*

Abstract. This paper deals with the automatic retrieval of issues reported in legal texts and presents an experience with expert's reports on the application of ILO Conventions. The aim is to provide the end user, i.e. the legal expert, with a set of rules that permits her/him to find among a predefined list of issues those addressed by any new text. Since the end user is not supposed to be able to pre-process the text, we need rules that can be directly applied on raw texts. We present the strategy followed for generating the rules in this ILO legal setting and single out a few possible improvements that should significantly improve the performance of the retrieval process. Our approach consists in characterizing in a first stage a list of descriptor concepts, which are then used to get a quantitative representation of the texts. In the learning phase, using a sample of texts labeled by legal experts with the issues they actually address, we build the rules by means of induced decision trees.

Keywords: Information retrieval, content prediction, quantitative text representation, legal texts.

1 Introduction

The concern of the paper is the automatic identification of the type of issues reported by given legal texts, for example which violations are pointed out in experts' comments on the application of ILO (International Labor Office) Conventions. Such an automatic text mining process becomes necessary when we face a large number of texts for either 1) pointing out the most relevant texts when one wants to investigate a given issue, or 2) drawing synthetic analyses of the relationships between issues as well as with other factors. The objective is then essentially to provide the end user, i.e. the legal expert, with prediction rules of the issues addressed by each text. We consider the case where the issues of interest have been previously specified. We assume thus that we have a closed list of issues.

The paper describes the process followed for building such rules within a joint research project between the ILO, the University of Geneva and the University of Lyon 2 (Ritschard et al., 2007) on the Social dialogue regimes prevailing in democratic countries. We also single out the main weaknesses of the approach and propose a series of strategies for improving the process.

The approach followed consists in characterizing in a first stage a list of descriptor concepts from which we derive then a quantitative representation of the texts. In the learning phase, using a sample of texts labeled with the addressed issues by legal experts we build the rules by means of induced decision trees. A separate tree is grown for each of the issue. Each time a binary variable indicating whether the issue is present or not is used as target variable and the descriptor concepts serve as predictive attributes. The characterization of the descriptor concepts and the quantification of their importance within each text is obviously a crucial stage in our process. Once the rules were obtained, we had to provide the end user (legal expert) a simple piece of software that 1) builds in an automatic way the quantitative representation of any new text, i.e. evaluates the importance of each descriptor concept in the text, and 2) determines from that representation what the probability is that the text addresses each issue of interest.

To make our presentation less abstract, a few words are worth on the application context for which the described text mining strategy was developed. The aim of text mining was to help us identify the nature of issues raised by a Committee of experts (CEACR) regarding the application of ILO Conventions. Due to space constraints we consider here only Convention 87 on Freedom of association and protection of right to organize. What we want to know is what types of violations of this Convention does the Committee identify in its reports. Using a priori knowledge, we categorized the possible violations in the form of a list of 9 key concepts — types of violations — (Table 1) themselves derived from a more detailed list of 27 key concepts listed in Ritschard et al. (2007).

v_1	Right to life and physical integrity (not observed)
v_2	Right to liberty and security of person / Right to a fair trial (not observed)
v_3	Right to establish and join workers' organizations
v_4	Trade union pluralism
v_5	Dissolution or suspension of workers' organizations (not observed)
v_6	Election of representatives / Eligibility criteria
v_7	Organization of activities / Protection of property / Financial independence
v_8	Approval and registration of workers' organizations
v_9	Restrictions on the right to industrial action

Table 1. Retained key concepts, i.e. types of violation.

The paper is organized as follows. In Section 2 we discuss the usefulness and inconvenience of text pre-processing and explain in Section 3 the semantic preserving text representation that was retained. The learning process itself is described in Section 4, where we give also some experimentation results illustrating the efficiency of the process. Concluding remarks are given in Section 5.

2 Text Pre-Processing

Text mining (Feldman and Dagan, 1995; Fan et al., 2006) refers to the process of analysing text to extract information that is useful for particular purposes (Witten and Frank, 2005, pp 351-356). It is supposed to be more than just finding documents or pages containing a given keyword — which is what simple indexing or search engines do well. For instance, if we are looking for texts commenting on violations of the freedom to organize the election of trade union representatives, we will not be satisfied with just texts containing the keyword “election”, but we may want to consider also all terms or expressions more or less related to this notion such as for example “elected workers’ representative” or “union leader”.

As opposed to numerical data, text data are essentially unstructured. Synonymy (different expressions with same meaning) and polysemy (different meanings for a same expression), among others, make them hard to analyse in an automatic way and necessitate heavy pre-processing. The aim of the pre-processing is to transform the essentially unstructured text data into a suitable structured representation for further automatic processing. By structured representation we mean a representation where each useful notion is uniquely and unambiguously defined so that we can surely rely on the counts of its occurrences.

There are basically two main ways of representing a text: through *n*-grams and as a bag of words. The former ignores the meaning of the words and considers each subsequence of say 3 letters — 3-gram — that can be found in the words as a countable characteristic (Damashek, 1995; Mayfield and McNamee, 1998) . The second (Salton et al., 1992, 1996) retains each different observed word as a characteristic and focuses essentially on its frequency in the text and among the texts. The latter approach is best suited for our supervised classification purpose where the semantic content of the text is of primary importance.

Now, texts contain a huge number of different words. Some of them may have a same or similar meaning (synonyms), may have a context dependent meaning (polysemy), or, as in the case of function or stop words (the, to, from, or, and, ...), will clearly be useless for discrimination purposes. The general practice is then to reduce the number of descriptors by dropping useless stop words and by merging synonyms into equivalence classes.

A first step for solving ambiguities is tagging words grammatically, which can be done automatically using for instance freely available tools such as Brill (Brill, 1995) or TreeTagger (Schmid, 1994). The grammatical tag permits indeed to distinguish for example between the noun, verb or adjective usage of the word “trade”, or the conjunction, verb or adjective usage of the word “like”. This grammatical tagging will also pinpoint stop words that could be dropped from the list of descriptors.

To avoid bothering with the various inflected forms of nouns, verbs and adjectives, other often applied pre-processing operations are lemmatization and stemming (Plisson et al., 2004). The former consists in retaining just the base form — e.g. the infinite of a conjugated verb — of each encountered word, and the latter in extracting the lemma — the root — of each word. This can again be done almost automatically with freely available tools such as TreeTagger (Schmid, 1994).

In our case, since the goal is to facilitate the processing of new additional texts by legal experts with no experience in these pre-processing steps, we opted for an approach that avoids in its application phase any pre-processing operation that could not be fully automatized. Therefore, we chose to not lemmatize the texts, and resorted to grammatical tagging only during the learning phase in order to facilitate the extraction of the useful terminology.

3 The Chosen Text Representation

For the purpose of our analysis, we decided to represent the CEACR comments by means of a limited set of descriptor concepts. These concepts were defined in a partially automated process consisting in first extracting the useful terminology, then grouping the terms into concepts and eventually refining the description of the concepts. We begin by commenting the terminology extraction process.

3.1 Extracting the Useful Terminology

The terminology that could be used for predicting violations reported in the Committees's observations includes not only single words, but also composite expressions such as “trade union” or “right to organize”. It is then essential to find and list the terms useful for the analysis.

Several tools can be used for this. Some of them, such as Xtract (Smadja, 1993), ATR (Frantzi et al., 2000), Lexter (Bourigault and Jacquemin, 1999) proceed automatically either by comparison with a pre-specified lexicon or by seeking frequent sub-sequences of words. Others, such as Exit (Heitz et al., 2005), are semi-automatic and require a domain expert to guide the process. The latter are best suited when, as in our case, we do not have access to a lexicon of the considered specialized language. Since we had the possibility to interact with legal experts, we chose to extract the useful terminology with the aid of the Exit software.

The input data provided to Exit is the grammatically tagged text (the set of all comments merged into a single file). We then select the useful terms in an iterative way. First, we chose successively among single words or pairs of a given type — noun-noun, noun-adjective, adjective-noun, verb-noun, noun-verb, etc. — that satisfy a minimal frequency criterion those that the expert considers relevant for the analysis. For example, “worker organization” and “national security” are two retained pairs, the former being of the noun-noun type and the latter of the “adjective-noun” type. A grammatical tag is assigned to each new retained term according to rules that could be changed by the user. For instance, adjective-noun terms such as “national security” are automatically tagged as noun. Then by iterating the process we single out terms that include themselves previously defined terms. We get thus terms composed of more than two words such as “minimum level of service”.

3.2 Descriptor Concepts

There is a huge number of different terms — words and composite expressions — used in the CEACR comments and it is not convenient to use all of them as text descriptors. We therefore, decided to represent texts through a small number of descriptor concepts that: (i) Characterize the conceptual content of the text; (ii) Are useful for predicting the issues — violation or key concepts — reported in the observations.

A first entirely statistical possibility of characterizing descriptor concepts (Kumps et al., 2004) would be to seek the words that best discriminate the key concepts we want to predict, and then to group them according to their co-occurrences. Lemmatization would be necessary in that case.

However, since we had the possibility to interact with legal experts, we preferred to rely on a linguistic approach. Such an approach where terms — words and expressions — are grouped according to both their statistical characteristics and the similarity of their meaning, provide concepts that are semantically better founded.

Thus, the approach followed consists in three steps carried out on the overall corpus: i) a preliminary set of concepts is built during the terminology extraction with Exit; ii) this preliminary set and the concept definitions are refined through an extensional induction process (Kodratoff, 2004) with the legal experts; and iii) the experts’ amended list is once again compared with the text content for a final coherence check.

The preliminary concept set is obtained in a semi-automatic way by starting the term extraction process with a high threshold, which provides a relatively short list of terms. Those terms may be considered as initial representatives of the main conceptual axes that can be found inside the texts. We obtain a starting set of concepts after possibly grouping terms with similar semantic meaning. Then, we repeat the process by lowering successively the minimal frequency threshold. At every iteration, we get additional terms and then assign each one of them to the most appropriate preexisting concept. In case there is no reasonable preexisting concept with which the new term could be associated, a new concept is created. At the end of the terminology extraction we get our preliminary list of concepts, where each concept is characterized by its list of associated terms.

This preliminary list of descriptor concepts serves then as a starting list for the experts who may either confirm the relevance of the concepts or change them to fit their overall knowledge of the domain. The preliminary list is thus transformed into an expert’s amended list of concepts.

In order to increase even further the coherence of the amended descriptor concepts, we carried out some additional checking. Indeed, we observed that the overall corpus of CEACR comments contains some infrequent terms that clearly belong to one of the retained descriptor concepts. Ignoring them would undoubtedly be a source of errors. The goal of the additional checking is to browse the corpus for such relevant but infrequent terms. More specifically, for each term already associated to a concept, we look for the presence in the corpus of synonyms and alternative inflection forms as well as for the presence of extended terms obtained by inserting one or more words in the term. For example, the term "call a strike" is frequent in the corpus and was detected as representing the strike action descriptor concept. Less frequent expressions such as "calling a strike" or "calling of a strike", were not detected however. The search of such alternative forms is easily done by browsing the terms found with regular expressions. For example, using the two strong words "call" and "strike", all three aforementioned terms were found with the PCRE regular search expression:[1]

```
"/[^;\.]*call[^;\.\,]{0,45}strike[^;\.]*/i"   .
```

As for synonyms, a lexicon such as the online WordNet may be useful for usual terms. For a specialized corpus such as the one formed by our legal texts, it is more helpful to ask experts in the domain. This is what was done in our analysis. Good sense may also prove useful. For example, we noticed in the reports that experts used independently and equivalently the terms "trade union" and "workers organization". Hence, each time a concept definition list included a term such as "registration of a trade union", we augmented, when it made sense, the list with "registration of a workers organization", even when this new expression was infrequent in the corpus.

The final list of descriptor concepts is given in Table 2 and examples of their list of associated terms can be found in Ritschard et al. (2007).

The designing of the descriptor concepts is clearly a crucial stage of our text mining process. It is also time-consuming and requires clever tuning through individual interventions from both the domain experts and the text mining experts. Furthermore, because of these multiple personal interventions, the resulting descriptor concepts remain somewhat subjective. Improvement and systematization of the process is possible and would here be necessary. It requires, however, an access to a detailed ontology of the concerned legal domain which does not yet exist. The designing of such an ontology that puts together the characteristic terminology of the domain, organizes it in terms of concepts and sub-concepts, and also describes the interrelation between concepts would then be our next development priority.

[1] The regular expression searches the text for expressions in which the word "call" is preceded by any sequence of characters other than a semi-column or a dot, the word "strike" is followed by any sequence of characters other than a semi-column or a dot, and the two words are separated by any sequence of at most 45 characters other than a semi-column, a dot or a comma.

c_1	Life and physical integrity	c_{10}	Industrial action
c_2	Liberty and security of persons	c_{11}	Essential service
c_3	Property and financial independence	c_{12}	Arbitration
c_4	Service	c_{13}	Strike action
c_5	Pluralism	c_{14}	Union establishment limitations
c_6	Election	c_{15}	Specific workers
c_7	Opinion and expression freedom	c_{16}	Number of workers
c_8	Restrictions on trade union activities	c_{17}	Supervision
c_9	Trade union approval		

Table 2. Retained descriptor concepts

3.3 The Quantitative Text Representation

Having now defined our descriptor concepts, we get a quantitative representation of the texts by assigning for each document (comment) a load on each concept. A classical way is to use the *tf* × *idf*, which is the term frequency (*tf*) — indeed the term count — in the document weighted by the inverse of the document frequency (*idf*), the document frequency being the number of documents in which the concept has been observed (Salton and Buckley, 1988). The general idea of this *tf* × *idf* is that a term — a concept in our case — is characteristic of a text when it is frequently mentioned in it (high *tf*) and only few other documents mention it (high *idf*). Let tf_{ij} be the term frequency of concept j in document i, and idf_j be the inverse term frequency of concept j. Formally, the inverse document frequency is defined as $\log(d/d_j)$, where d is the total number of documents and d_j the number of documents mentioning concept j. The *tf* × *idf* weight of concept j in a document i, is then

$$w_{ij} = tf_{ij}\, idf_j = tf_{ij}\, \log\left(\frac{d}{d_j}\right) .$$

With this formulation, the lengthier a document i the greater chance it has to have large tf_{ij}'s and hence important weights. To avoid this size effect, Salton et al. (1992) propose the length normalized form $\tilde{w}_{ij} = w_{ij}/||\boldsymbol{w}_i||$, with $\boldsymbol{w}_i$ the vector of the $tf_{ij} \times idf_j$'s of the document i.

For our objectives, what matters is the absolute place devoted to a given concept in a comment whatever other issues the comment addresses. In that sense, the normalized *tf* × *idf* is not useful in our setting. In other words, we consider that the importance of a concept in a text is reflected by its number of occurrences independently of the document's length.

Using the *tf* × *idf*'s of the retained descriptor concepts, our text data set can be put in the form of a classical quantitative data table as illustrated in Table 3, which exhibits an extract of the data for comments on the application of Convention 87.

CEACR Comment	Descriptor Concepts							
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	$\cdots$
Algeria 1991	0	0	0	0	2.75	0	0.8	$\cdots$
Argentina 1991	0	0	0	0	20.59	2.39	0.8	$\cdots$
Bangladesh 1991	1.0	0.77	2.35	1.24	0	1.59	5.59	$\cdots$
$\cdots$								

Table 3. Extract of data representing comments in terms of descriptor concepts

4 Learning Process

Through the previous steps, i.e. extracting useful terms, organizing them into a limited number of relevant descriptor concepts and finally measuring the importance devoted to each descriptor concept by each CEACR comment with the $tf \times idf$ weight, we were able to code the comments numerically. What remains now is to learn the prediction rules.

This learning phase requires a learning sample of texts — comments — previously labeled in accordance with the type of violation they report. The labeling was done by a legal expert for 78 out of 671 CEACR texts concerning Convention 87. The labels are represented by a set of ℓ 0-1 indicator variables v_k, $k = 1, \ldots, \ell$ that take value 1 when the text mentions violation k, and zero otherwise. Remember that the violations we are interested in correspond to the key concepts listed in Table 1.

Using this learning sample the aim is to find rules for predicting each key concept (violation) from the quantified descriptor concepts. We then consider successively each key concept in turn, and build the prediction rule for it. Letting c_j denote the $tf \times idf$ of the jth descriptor concept, we look for each k for a prediction rule $\hat{v}_k = f_k(c_1, \ldots, c_c)$.

Since our texts are numerically coded, classical supervised statistical or machine learning techniques may be considered. We used induced classification trees, which produce usually good classification results and have the advantage of being easily applicable, of detecting automatically interaction effects of the predictors and of providing easily interpretable rules.

Classification trees are grown by seeking, through recursive splits of the learning data set, some optimal partition of the predictor space for predicting the outcome class, i.e. whether the comment does or does not report a violation of type k. Each split is done according to the values of one predictor — descriptor concept —. The process is greedy. At first step, it tries all predictors to find the "best" split using, for quantitative predictors as those we face here (the concept $tf \times idf$'s), an automatic local optimal discretisation. Then, the process is repeated at each new node until some stopping rule is reached. This requires a local criterion to determine the "best" split at each node. The choice of the criterion is the main difference between the various tree growing methods that have been proposed in the literature.

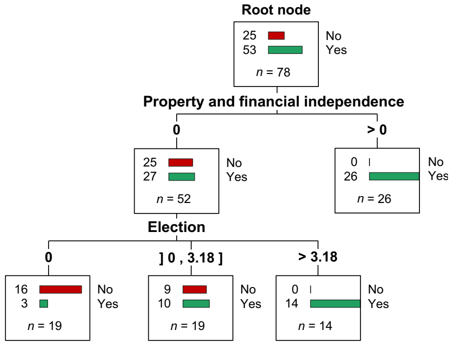

Fig. 1. Induced tree for v_7, Restrictions on organization of trade union activities

Figure 1 shows the tree grown for violation 7 — restrictions on the organization of trade union activities — using Exhaustive CHAID (the improved CHAID method by Biggs et al., 1991) with a significance threshold of 5%, the Bonferroni correction, a minimal leaf size of 10 and a minimal parent node size of 30. The descriptors retained are whether the comment explicitly refers to property and financial independence and whether it talks about election.

The tree has 4 terminal nodes, which are called *leaves*. We associate to each of them a rule taking the form *condition* ⇒ *conclusion*. The condition is defined by the path from the root node to the leaf, and the conclusion is, for a classification tree, usually the most frequent class in the leaf.

A similar tree is grown for each type of violation, which results in 6 sets of rules. Some violations (v_1, v_2 and v_5 for instance), are not covered by any comment in the learning sample, and no tree is grown for them. In two cases, we did not rely on the mere statistical criterion and forced the algorithm to split at the first step using the second best variable that seemed theoretically better sounded from our knowledge base.

The classification performance of each tree may be evaluated by means of its classification error, i.e. the percentage of cases which are misclassified

Key concept	Learning	Cross-validation		Test sample (size 21)
(violation)	error rate	error rate	std err	number of errors
v_3	14.10%	n.a.*	n.a.*	3
v_4	5.13%	5.13%	2.50%	0
v_6	12.82%	14.1%	3.94%	4
v_7	15.38%	n.a.*	n.a.*	7
v_8	7.69%	7.69%	3.01%	4
v_9	2.56%	2.56%	1.79%	2

*Cross-validation is not available for v_3 and v_7, because first split is enforced.

Table 4. Error rates, Convention 87

Key	Positives		Negatives		% with key concept			
Concept	true	predicted	true	predicted	reported	predicted	Recall	Precision
v_3	30	32	37	46	50.0%	41.0%	76.9%	93.8%
v_4	29	31	45	47	39.7%	39.7%	93.5%	93.5%
v_6	35	38	33	40	53.8%	48.7%	83.3%	92.1%
v_7	50	59	16	19	67.9%	75.6%	94.3%	84.7%
v_8	29	30	43	48	43.6%	38.5%	85.3%	96.7%
v_9	57	59	19	19	73.1%	75.6%	100.0%	96.6%

Table 5. False Positives, False Negatives, Recall and Precision, Convention 87

by the derived classification rules. Table 4 shows learning error rates (i.e. rates computed on the learning sample) and 10-fold cross-validation error rates with their standard error. It gives in addition the number of errors on a small test sample of 21 comments about the application of Convention 87.

Table 5 exhibits some additional useful indicators. Column 'True positives' gives the number of comments classified as reporting a violation of type k that effectively report it, and column 'Predicted positives' the total number of comments classified as reporting the violation. For key concept v_7, for example, 50 out of 57 comments classified as reporting the violation actually report it. The number of true and predicted negatives is also shown. Table 5 gives the percentage of the 78 comments that report on the relevant key concept and the percentage of comments that are classified as reporting the key concept. For v_7 again, we may check that $59 = 75.6\% \times 78$, are classified as reporting the violation, while there is actually a total $53 = 67.9\% \times 78$ reporting v_7. The 'Recall' is the percentage of this total that is classified as reporting the violation — true positives —, e.g. $94.7\% = 50/53$ for v_7. The 'Precision' is the ratio of the number of true positives on the number of predicted positives, e.g. $84.7\% = 50/59$ for v_7.

These results are quite good when compared with those obtained with other classifiers. For instance, we experimented with support vector machine (SVM) as well as with neighboring graphs. These methods did not produce significantly better results, while producing much less explicit rules. Nevertheless, error rates above 10% as well as recall and precision percentages below 90% may look unsatisfactory. Remember, however, that the learning was done with a sample of only 78 texts. It is also worth mentioning that errors may be more or less important depending on the research objectives. In our case, as stated in the introduction, the text mining has two main purposes: To help the legal expert interested in a given issue in identifying texts reporting this issue (it is not supposed to replace the expert in this task), and to provide material for analysing synthetically the relationship between issues, i.e. types of violations. With such objectives, it is not dramatic to make false predictions for a small number of texts. If the end user wants to find all texts dealing with an issue of interest, false positive cases will generally be easier to identify than false negative ones. Hence we should in that case favor a strategy that limits false negatives even if it is at the cost of more false positives. This can easily be done by lowering for instance the probability threshold used for assigning the outcome class to the rules. For synthetic analyses, on the other hand, we may prefer to retain only the most reliable predictions. We would then primarily limit the number of false positives.

5 Conclusion

We described in this paper an ad hoc text mining process for identifying issues reported in legal texts. The process described is semi-automatic. The building of the prediction rules relies on an interaction with the domain expert at several points and especially for defining relevant descriptor concepts. This stage of the process could, however, be improved on at least two sides. First, the interest of the descriptor concepts for the targets (each associated to one of the considered violations) is based solely on the opinion of the domain expert. By specifying a global criterion taking simultaneously into account all considered targets, it should be possible to measure the global discriminating power of terms and hence select objectively the most discriminating ones. Likewise, we should be able to measure the similarity in the discriminating capacity of the terms and use these similarities as a guide for grouping them into descriptor concepts. Second, organizing the descriptor concepts into hierarchical ontology would allow for some freedom for choosing between concepts and sub-concepts. It would also produce reusable knowledge material for other applications in similar domains. Beside the systematization of the descriptor concept definition stage, significant improvement may also be expected at the learning level. For instance, taking account of the preference for limiting false positives rather than false negatives (or conversely) during learning and not only during class assignment should most probably generate better suited rules.

References

BIGGS, D., B. DE VILLE, and E. SUEN (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics 18*(1), 49–62.

BOURIGAULT, D. and C. JACQUEMIN (1999). Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pp. 15–22.

BRILL, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics 21*(4), 543–565.

DAMASHEK, M. (1995). Gauging similarity with ngrams: Language-independent categorization of text. *Science 267*, 843–848.

FAN, W., L. WALLACE, S. RICH, and Z. ZHANG (2006). Tapping the power of text mining. *Communications of the ACM 49*(9), 76–82.

FELDMAN, R. and I. DAGAN (1995). Knowledge discovery in textual databases (KDT). In *KDD '95*, pp. 112–117.

FRANTZI, K. T., S. ANANIADOU, and H. MIMA (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries 3*(2), 115–130.

HEITZ, T., M. ROCHE, and Y. KODRATOFF (2005). Extraction de termes centrée autour de l'expert. *Revue des nouvelles technologies de l'information RNTI E-5*, 685–690.

KODRATOFF, Y. (2004). Induction extensionnelle: définition et application l'acquisition de concepts à partir de textes. *Revue des nouvelles technologies de l'information RNTI E-2*, 247–252.

KUMPS, N., P. FRANCQ, and A. DELCHAMBRE (2004). Création d'un espace conceptuel par analyse de donnés contextuelles. In G. Purnelle, C. Fairon, and A. Dister (Eds.), *Le Poids des Mots (JADT 2004)*, Volume 2, pp. 683–691. Presse Universitaire de Louvain.

MAYFIELD, J. and P. MCNAMEE (1998). Indexing using both n-grams and words. In *TREC*, pp. 361–365.

PLISSON, J., N. LAVRAČ, and D. MLADENIĆ (2004). A rule based approach to word lemmatization. In *Proceedings of IS04*.

RITSCHARD, G., D. A. ZIGHED, L. BACCARO, I. GEORGIOU, V. PISETTA, and M. STUDER (2007). Mining expert comments on the application of ILO Conventions on freedom of association and collective bargaining. Working Papers 2007.02, Department of Econometrics of the University of Geneva.

SALTON, G., J. ALLAN, and A. SINGHAL (1996). Automatic text decomposition and structuring. *Information Processing and Management 32*(2), 127–138.

SALTON, G. and C. BUCKLEY (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*(5), 513–523.

SALTON, G., C. BUCKLEY, and J. ALLAN (1992). Automatic structuring of text files. *Electronic Publishing—Origination, Dissemination, and Design 5*(1), 1–17.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing, Manchester.*

SMADJA, F. A. (1993). Retrieving collocations from text: XTRACT. *Computational Linguistics 19*(1), 143–177.

WITTEN, I. H. and E. FRANK (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Amsterdam: Morgan Kaufman (Elsevier).