Strategies in Identifying Issues Addressed in Legal Reports

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

[1]Dept of Econometrics, University of Geneva
http://mephisto.unige.ch
[2]ERIC Laboratory, University of Lyon 2

Compstat 2008, Porto, Portugal, August 24 - 29

Aim of presentation

- Automatic identification of issues reported in texts.
- Experience with reports on application of ILO Conventions.
- Describing the text mining process
  - Quantitative representation of the texts.
  - Learning rules for predicting issues reported by any given text.

Project on Social Dialogue Regimes

- Financially supported by the Geneva International Academic Network (GIAN).
- Joint project between
  - Depts of Econometrics and of Sociology (U. of Geneva),
  - ERIC (U. of Lyon 2)
  - International Institute of Labour Studies (ILO, Geneva).
- Analysis of the determinants and socioeconomic correlates of Social Dialogue Regimes (SDR)
  - Sociopolitical regimes in which workers have freedom to establish organizations of their own choosing, negotiate collectively over working conditions, and participate through their associations in the design and implementation of policies that affect their lives

The Objectives of Text Mining

- Focus on CEACR comments from 1991 to 2002.
  - CEACR = Committee of Experts on the Application of Conventions and Recommendations
- Creation of synthetic indicators of legal rights.
- Use of indicators for aggregate data analysis, aimed at exploring how particular violations are (or are not) linked to others, as well as relationships with socioeconomic indicators.

Why resorting to text mining ?

- Extract useful information from huge number of expert comments (∼1200 reports).
- Two main goals
  - Assist legal experts in the search of relevant information, speed up the process.
  - Produce indicators for synthetic aggregate analysis.

What is text mining ?

- Process of analyzing text to extract information useful for particular purposes.
- More than indexing or search engine.
- Aims at discovering knowledge about :
  - Content
  - Structure
  - Semantic
  - Ontology : typical terminology grouped into concepts and organized into conceptual hierarchies
  - ...

Different text mining usages

- Statistical analysis of used words or sentence structure.
- Relationships between texts
  (Which texts are similar to a given one ?)
- Summarizing automatically articles or documents.
- Unsupervised text categorization (clustering).
- Supervised categorization (detecting spam).
- Technological watch.
- Building ontologies (finding typical terminology and organizing it into conceptual hierarchies).
- Retrieving concepts from texts.
- Information Retrieval.
- ...

Challenges in text mining

- Text are unstructured data
- Polysemy
  - "Mining expert comments"
  - "Mining expert comments"
- Synonymy
  - "Trade union", "Workers' organisation"
- Inflected forms, stop words, ...
- ⇒ Requires pre-processing

The retained approach

- Want a tool that can be used by any non text mining expert.
- ⇒ No pre-processing
  (grammatical tagging, lemmatisation, stemming)
  in the application stage.

Two main steps

1. Representing texts by a set of quantitative variables (whole corpus) :
   - Extracting useful terminology.
   - Grouping terms into reduced number of descriptor concepts.
   - Quantifying descriptor concepts (tf×idf).
2. Learning prediction rules (learning sample) :
   - Classification trees

Retained key concepts, i.e. types of violation

Whose presence is the attribute to predict

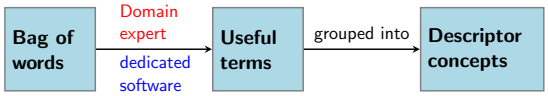
- Original list of 27 key concepts (types of violation)
- Merged into the following 9 key concepts for Convention 87

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

From text to quantitative representation

Defining the descriptor concepts

- Two major approaches
  - *n*-grams (hard to account for semantic)
  - Bag of words



trade	trade union	17 for Conv. 87
union	trade union pluralism	9 for Conv. 98
action	trade union activity	
...	...	

Defining descriptor concepts
With help of legal experts

- Whole corpus (1 file), Grammatically tagged (TreeTagger)
- Process based on both :
  - Statistical characteristics (discriminating terms, co-occurrences)
  - Similarity of meaning (according to domain expert)
- 3 steps :
  - Preliminary set of concepts (by grouping new terms with already extracted most frequent ones, EXIT)
  - Extensional induction with legal expert (preliminary list is adapted according to knowledge base)
  - Amended list is again compared with corpus (find infrequent terms with meaning similar to defined concepts)
- Result : Concepts, each defined by a list of terms.

Retained descriptor concepts for Convention 87

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_{13}	Strike action
c_4	Service	c_{14}	Union establishment limitations
c_5	Pluralism		
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		

Examples of descriptor concept lists

Election (C_6)	Restrictions on trade union activities (C_8)
electoral procedure/	trade union activitie/
representative member/	right to organise/
trade union office/	right of trade unions to organize/
eligibility/	right to publication/
re-election/	right to assembly/
representatives of organization/	right to disseminate information/
representatives of trade union/	freedom of opinion/
vote/	freedom of expression/
elect member/	political opinion/
union leader/	political activity/
elect their representative/	hold meeting/
elect their own representative/	right to organize/
to elect representative/	right of association/
to elect representatives/	right of workers organization/
elect its representatives/	right of workers to organize/
elect their trade union representative/	right to hold trade union/
election of trade union representative/	holding office/
elected representative/	formulate their programmes/
elected workers' representative/	right of organizations to organize/
election/	right of first-level unions to organize/
elect freely their representative/	right of unions to organize/
elected their representative/	right of workers' organizations to organize/
union officer/	right of trade union organizations to organize/
union office/	right of these employees to organize/
representatives of association/	rights of workers' organizations to organize/
representatives of union/	right of workers' trade unions to organize/
representatives of workers' organization/	political activities/
⋮	⋮

Quantifying descriptor concepts

The $tf \times idf$ measure

$tf_{ij} \times idf_j$ weight of concept j in text i

- tf_{ij} frequency of concept j in text i
- idf_j inverse document frequency of concept j
  ($= \ln(d/d_j)$ with d_j #docs with concept j, d total #docs)

Excerpt of data representing comments

CEACR Comment	Descriptor Concepts								
	c1	c2	c3	c4	c5	c6	c7	c8	···
Algeria 1991	0	0	0	0	2.75	0	0.8	0	···
Antigua and Barbuda 1991	3.0	1.53	0	0	0	0	0	0	···
Argentina 1991	0	0	0	0	20.59	2.39	0.8	0	···
Bangladesh 1991	1.0	0.77	2.35	1.24	0	1.59	5.59	0	···
Belgium 1991	0	0	0	0	0	0	0.8	0	···
Bolivia 1991	0	0.77	0	1.24	1.37	4.77	0	0	···
Bulgaria 1991	0	0	0	0	2.75	0	0	0	···
Burkina Faso 1991	0	0	0	0	0	0	0.8	2.19	···
···									

Limits and possibilities of improvement

- Design of descriptor concept is crucial stage of the process.
- Semi-automatize process that strongly relies on the domain expert.
- Time-consuming
- Clever tuning through individual intervention of
  - domain expert
  - text mining expert
- Resulting descriptor concept somewhat subjective (not strictly reproducible)
- **Need for improvement and Systematization**

Improvement and systematization

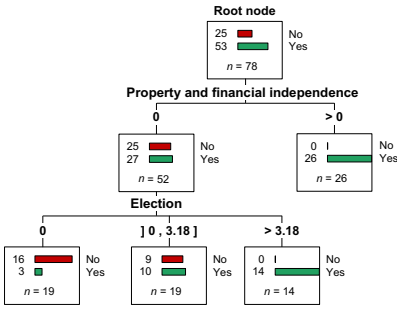
- Through access to lexicon and ontology of the concerned legal domain (which do not yet exist)
- Lexicon : list of words (terms) with synonyms.
- Ontology
  - puts together the characteristic terminology
  - organizes it in terms of concepts and sub-concepts
  - describes interrelation between concepts.

Learning Categorization Rules

- Learning sample
  - 78 texts out of 671 for C87
  - 60 texts out of 509 for C98
- Assign label (violation) to texts in sample
  - Learn a classifier, for each violation k
  - $v_k = f_k(c_1, c_2, \ldots), \quad k = 1, 2, \ldots, 9$
- We used classification trees

Grown tree

for restriction on organization of trade union activities v_7



Error rates, Convention 87

Key concept (violation)	Learning error rate	Cross-validation error rate	Cross-validation std err	Test sample (size 21) 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.

Example of prediction error

False Positives, False Negatives, Recall and Precision

Convention 87

Key Concept	Positives true	Positives predic.	Negatives true	Negatives predic.	% with key concept reported	% with key concept predic.	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%

Degrees of freedom in assignment rules

- Two class problem (violation yes/no)
- Assignment rule : Select "yes" if $p(\text{yes}) > s$
- Assignment controlled through threshold s :
  - large : favors precision (few false positives)
  - small : favors recall (few false negatives)
- Choice depends on what we want to do with predictions
  - Help legal expert identifying issues with given countries : **recall** more important (easier to find out false positives)
  - Synthetic analysis : **precision** more important.

Conclusion 1

- Ad hoc text mining process for identifying issues reported by ILO legal texts.
- Rule construction is long and time-consuming.
  - Semi-automatic building of the prediction system based on strong interaction with domain expert.
- Resulting rules are fully automatic.
  - Can fast and straightforwardly be applied to any new text.

Conclusion 2

- Improvement and systematization
- Characterization of relevant descriptor concepts.
  - Global criterion measuring simultaneously discrimination of all targets (violations).
  - Lexicon and ontology of the domain.
- Improvement at learning level.
  - Account for preference for precision or recall during tree growing (not only for class assignment).

THANK YOU !

References

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.

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

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

Convention 87 : Gabon 1992

(back)

Actual and predicted violations in CEACR 1992 report for Gabon

Violation	3	4	6	7	8	9
Actual	0	**1**	**0**	0	0	**1**
Prediction	9%	**94%**	**92%**	53%	10%	**97%**

- Report contains two references to the "Election" descriptor concept ⇒ 94% probability to mention problems regarding election of representatives (v_6).
- However, first mention of "election" is within a sentence about new rules adopted for insuring representative election, second one in a sentence for asking the Government to indicate the election results.
- Positive reference to "election" confused with a violation.

CEACR 1992 report for Gabon

(back)

The Committee notes, in particular, the Government representative's statement that the recognition of individual liberties in the new Constitution of Gabon, which came into force on 26 March 1991, has a corollary in the overall social plan, which is the abolition of trade union **monopoly**, that is to say the establishment of genuine and complete freedom of association. It notes that a draft new Labour Code which was discussed during a tripartite meeting from January to April 1991, attended both by the unitary employers' and workers' central organisations and by other organisations of workers and employees, has already been examined by the Government and was to be presented before the end of 1991. According to the Government, the amendment envisaged includes the repeal of section 174 of the present Labour Code which obliges all workers' or employers' organisations to affiliate with the Trade Union Confederation of Gabon (COSYGA) or the Employers' Confederation of Gabon (CPG). The Government also states that Act No. 13/80 of 2 June 1980, establishing a trade union solidarity tax deducted for the COSYGA, is no longer applied and that the tax has not been deducted since March 1990. Legislation is to be adopted for its formal repeal.

With regard to the provisions on compulsory arbitration restricting workers' right to **strike** (sections 239, 240, 245 and 249 of the Labour Code), the Government representative stated that a draft law specifically on the right to **strike**, which takes into account the requirements of the Convention, has been prepared and may be incorporated into the revised Labour Code.

The Committee notes the Government's reply in its last report to the effect that : (1) COSYGA, whose members wish the organisation to continue under the same name, has complied with the laws of the Republic of Gabon and adopted new rules under which it is now protected from any influence on the part of political parties and religions ; (2) the new rules of COSYGA settle clearly the problem of the social assets of COSYGA vis-à-vis the new unions ; (3) the sole object of occupational organisations is to examine and defend members' economic, industrial, commercial, agricultural and artisanal interests and there are no longer any restrictions on the establishment of these organisations ; and (4) future **elections** of staff delegates and members of the Economic and Social Cooperation Committees will demonstrate that the various unions in establishments and enterprises are representative.

In the light of this information, the Committee asks the Government to provide a copy of the new COSYGA rules with its next report and to indicate the results of the above-mentioned **elections**.